Generación automática de definiciones mediante explicitación. Una aplicación a los neologismos del dominio médico

Walter Adrián Koza* y Ricardo Martínez-Gamboa**

Resumen: Se describe un método de generación automática de definiciones mediante la explicitación del significado de los morfemas aplicado a neologismos médicos. En primer lugar, se realiza el reconocimiento automático de los morfemas que componen los neologismos y se le asignan a cada uno de ellos los significados correspondientes. A continuación se combinan estos últimos mediante reglas de reescritura, a fin de obtener un sintagma nominal que exprese el significado de la palabra. El procedimiento se probó en una lista de 190 neologismos médicos extraídos del corpus CCM-2009 (Burdiles, 2012), y se obtuvo un 74,34 % de precisión, un 64,21 % de *recall* y un 68,9 % de medida F.
Palabras clave: definiciones, generación automática, morfología, neologismos, terminología médica.

Automatic generation of definitions through explicitation: An application to neologisms in the medical domain

Abstract: This paper describes a method of generating definitions automatically by making the meaning of morphemes explicit, and applies this method to medical neologisms. First, the morphemes that make up the neologisms are automatically recognized and each morpheme is assigned its corresponding meaning. These meanings are then combined through rewrite rules, in order to obtain a noun phrase that expresses the meaning of the neologism. The procedure was tested on a list of 190 medical neologisms extracted from the CCM-2009 corpus (Burdiles, 2012) and obtained 74.34% precision, 64.21% recall, and a 68.9% F-measure.
Key words: automatic generation, definitions, medical terminology, morphology, neologisms.

Panace@ 2016; 17 (44): 133-142 **Recibido: 13.VII.2016. Aceptado: 12.IX.2016.**

1. Introducción

El gran desarrollo de las tecnologías de la comunicación ha permitido producir, acceder e intercambiar un enorme flujo de información y conocimiento científico a usuarios de todo el mundo. No obstante, para acceder a esa gran cantidad de datos, se hace necesario disponer de herramientas que puedan procesarlos y que cuenten con sistemas de almacenamiento y de recuperación de la información (López-Huertas *et al.*, 2004). Una de las actividades principales en el desarrollo de dichos sistemas es la detección automática de términos de dominios específicos. Un término es una unidad léxica que designa un concepto en un campo temático particular (Cabré, 2002; Marincovich, 2008). La extracción de términos representativos de un área suele constituir el punto de partida para realizar tareas más complejas, como la elaboración de listas de entradas para diccionarios especializados, la creación de bases de datos, ontologías, taxonomías, etcétera.

Ahora bien, una de las áreas de conocimiento fundamentales es la de la medicina, no solo por la función social que cumple —conservar la integridad física y psíquica de los seres humanos— sino también por la creciente producción y circulación de textos que se producen en este dominio —artículos de investigación, casos clínicos, reportes técnicos, etcétera—. En este campo, ya en los trabajos de Krauthamer y Nenadić (2004) se menciona que, entre las barreras para una extracción de términos exitosa, se incluyen fenómenos como las variaciones léxicas, la sinonimia o la homonimia. Por otro lado, el mantenimiento de los recursos terminológicos, además, se dificulta ante el constante cambio de la terminología, puesto que se producen neologismos prácticamente a diario (Krauthamer y Nenadić, 2004). No obstante, se puede observar que una cantidad importante de ellos se forma a partir de la combinación de morfemas —raíces y afijos— propios del área médica (Herrero-Zorita *et al.*, 2015). Precisamente, en relación con esto último, el presente trabajo describe el desarrollo de un método de generación automática de definiciones aplicado en neologismos del dominio médico a partir del procesamiento de información morfosintáctica. En este sentido, se considera definición a la evidencia del significado de una expresión a partir de la explicitación ordenada de los elementos morfológicos que la componen. En este caso, con explicitación nos referimos al establecimiento del significado de una expresión a partir de la transformación de los elementos morfológicos que la componen en expresiones claras, precisas y definidas. Dicha perspectiva se encuentra estrechamente relacionada con la noción de explicitación propuesta desde el

* Pontificia Universidad Católica de Valparaíso (Chile). Dirección para correspondencia: walter.koza@pucv.cl.
** Universidad Diego Portales, Santiago de Chile (Chile). Dirección para correspondencia: ricardomartinez@gmail.com.

ámbito de la traducción (Herrezuelo, 2008; Alcántara, 2013; Soto, 2013). Al respecto, Vinay y Darbelnet (1995) definen este término como la técnica de traducción que se pone en práctica cuando se quiere poner de manifiesto, en una lengua meta, la información que se sobreentiende en una lengua origen. En el presente trabajo, se considera lengua especializada, es decir, propia del ámbito médico, y lengua meta a la definición generada por explicitación conformada mediante expresiones del lenguaje general. A modo de ejemplo, un término como *hemograma* se puede definir mediante una explicitación como 'representación gráfica de la sangre'. Para ello, es necesario reconocer los morfemas que conforman la palabra (*hemo* y *grama*), asignarles el significado correspondiente ('sangre' y 'representación gráfica', respectivamente), para, finalmente, ordenarlos y adicionarle los elementos funcionales pertinentes —determinantes, preposiciones, etcétera—.

A tales efectos, el objetivo del presente trabajo es establecer una formalización de dicho procedimiento de explicitación, a fin desarrollar una implantación computacional que permita generar definiciones para neologismos de forma —creados mediante la combinación de elementos morfológicos— del dominio médico. Para este objetivo, en primer lugar, se tomó la lista de neologismos obtenida a partir de una metodología de extracción basada en el procesamiento de información lingüística planteada en el proyecto Fondecyt 11130469 (Koza, 2015; Koza *et al.*, 2015). En segundo lugar, se realizó la segmentación de los candidatos a partir del reconocimiento automático de los morfemas propios del área médica, es decir, los denominados formantes cultos (Vivaldi, 2003), para, posteriormente, asignarle a cada lexema su significado y, finalmente, en tercer lugar, mediante reglas de reescritura, reordenar estos a fin de conformar un sintagma nominal que diera cuenta del significado del neologismo. La implantación en máquina fue realizada en Excel.

El artículo se organiza de la siguiente manera. En la segunda sección, se presenta el marco teórico que sustenta el presente trabajo. En la tercera se describe la metodología. En la cuarta sección se analizan los resultados obtenidos y finalmente, en la quinta, se presentan las conclusiones derivadas de la investigación.

2. Marco teórico

Si bien, hasta el momento, no se han encontrado trabajos similares, la presente investigación entabla relación con los trabajos llevados a cabo sobre extracción automática de candidatos a término y el análisis de la neología en comunicación especializada. A la vez, también se hace pertinente revisar las distintas concepciones sobre la definición y el concepto de explicitación. A continuación se especifican los lineamientos teóricos de dichas disciplinas que sustentan nuestra propuesta.

2.1. Extracción de candidatos a término en la medicina mediante técnicas lingüísticas

Las tareas de extracción automática de candidatos a término tienen ya un amplio desarrollo y una larga tradición (Kageura y Umino, 1996; Frantzi y Ananiadou, 1996; Park *et al.*, 2002). Esta disciplina se inicia a finales de los años 80, a partir de la necesidad de extracción automática de las unidades terminológicas de textos especializados en diversos campos (Marciniak y Mykowiecka, 2015). La extracción tanto de unidades simples como compuestas es fundamental, tanto para el desarrollo de recursos en el ámbito del procesamiento del lenguaje natural —glosarios, tesauros, ontologías, etcétera— como para tareas computacionales más complejas, como el reconocimiento automático de la información, la clasificación de textos, la elaboración de resúmenes o la traducción automática (Periñán-Pascual, 2015). Las metodologías en este ámbito se basan en tres tipos de enfoques: lingüísticos, estadísticos e híbridos. Los lingüísticos parten del planteo de que los términos de diversas áreas de especialidad siguen ciertos patrones morfosintácticos (Barrón, 2007). Los estadísticos, a su vez, se dividen en dos: los que se basan en medidas de asociación léxica (Pecina, 2010) y los que lo hacen exclusivamente en información estadística (Pazienza *et al.*, 2005). Por último, los enfoques mixtos o híbridos combinan tanto información lingüística como técnicas estadísticas.

En relación con los métodos basados en el procesamiento de información lingüística, Periñán-Pascual (2015) menciona que estos implican tres tareas consecutivas: (i) reconocimiento de palabras a partir de etiquetadores, (ii) reconocimiento de candidatos mediante patrones morfosintácticos y (iii) filtrado de candidatos mediante una lista de exclusión de palabras funcionales y genéricas. De acuerdo con esta perspectiva, la detección de candidatos a término estaría supeditada a la detección de los patrones sintácticos. Entre las críticas a este enfoque, el autor señala lo restrictivo del método al basarse exclusivamente en patrones lingüísticos, y añade que, en estos casos, las tareas de extracción son dependientes de las lenguas particulares en las que se aplican, lo que traería aparejado un alto costo de tiempo y trabajo manual. No obstante, el primero de los problemas puede reducirse significativamente si a la información morfosintáctica se le adiciona información léxica propuesta por diccionarios. Asimismo, el hecho de focalizarse en lenguas particulares no necesariamente ha de implicar un costo mayor de procesamiento si se compara con los tiempos que suponen algunas de las tareas propias de los métodos estadísticos, como, por ejemplo, el entrenamiento del corpus (Silberztein, 2016).

En lo que atañe al área de la medicina, desde hace tiempo se han venido desarrollando diversos sistemas de reconocimiento de términos para muchas clases de entidades de este dominio. Estos se basan tanto en características internas de clases específicas como en pistas externas que pueden ayudar al reconocimiento de secuencias de palabras que representan conceptos del dominio. Para ello, apelan a diferentes tipos de información, como ortografía —mayúsculas, dígitos, caracteres griegos—, elementos morfológicos —afijos específicos y formantes cultos— o a los resultados derivados del análisis sintáctico (Krauthamer y Nenadić, 2004). Asimismo, también se encuentran metodologías basadas en aprendizaje automático, como, por ejemplo, la propuesta por Vivaldi y Rodríguez (2015), quienes desarrollan un anotador semántico de entidades médicas basado en distancia de aprendizaje. Para ello, recurren a la taxonomía de SNOMED CT y a clasificadores

binarios, cuyos resultados se combinan luego con un metaclasificador.

En lo que atañe a los métodos basados exclusivamente en el procesamiento de información lingüística, estos pueden dividirse en dos enfoques: los basados en diccionarios y los basados en reglas morfosintácticas. Por un lado, los métodos constituidos a partir de diccionarios utilizan recursos terminológicos existentes con el propósito de localizar las ocurrencias de términos en los textos (Krauthamer y Nenadić, 2004). La limitación obvia que presentan es que las ocurrencias aún no registradas no pueden ser reconocidas. Por otro lado, también pueden influir negativamente factores como la homonimia y las variaciones en la escritura de los términos, como las variaciones en la puntuación (bmp-4/bmp4), el uso de diferentes numerales (syt4/syt iv), diferencias en la transcripción de letras del alfabeto griego (igα/ig alpha) o variaciones en el orden (integrin alpha 4/integrin4 alpha) (Tuason *et al.*, 2004).

Los enfoques basados en reglas morfosintácticas, por su parte, intentan recuperar términos por el restablecimiento asociado a los patrones de formación que han sido utilizados para construir los términos en cuestión. Desde esta perspectiva, se puede mencionar el trabajo de Segura *et al.* (2008), focalizado en la detección automática de fármacos genéricos mediante la utilización del metatesauro *ULMS* y las reglas de nomenclatura para la formación de fármacos genéricos propuestas por el consejo United States Adoptated Names (USAN), el cual permite clasificar los fármacos en familias farmacológicas. Con esta técnica, los autores han detectado fármacos no incluidos en el *UMLS*: han logrado un 100 % de cobertura y un 97 % de precisión utilizando el *UMLS*, y un 99,3 % de precisión y un 99,8 % de cobertura recurriendo a una combinación de información lexicográfica propuesta por el *UMLS* y las reglas de formación de nombres de fármacos propuestas por el USAN. Posteriormente, Gálvez (2012) elabora un trabajo similar, aunque basado solamente en reglas morfológicas propuestas por el USAN, al igual que Segura *et al.* (2008), y recurriendo a herramientas de estados finitos. De esta manera, la autora logra un 99,8 % de precisión y un 92 % de cobertura.

Finalmente, en el proyecto FONDECyT (Koza, 2015; Koza *et al.*, 2015) se plantea un método de detección automática de candidatos a término del dominio médico para el español a partir del procesamiento de información lingüística, esto es, léxica —aportada por diccionarios de la especialidad—, sintáctica y morfológica. En relación con el trabajo computacional, se asume que, en un corpus compuesto por textos del dominio médico, es factible acceder a su terminología mediante un proceso de detección automática basado en información lingüística porque los términos presentan, al menos, una de las siguientes características: (i) su lema es una entrada léxica de un diccionario electrónico del dominio médico; (ii) incluye un neologismo que posee una estructura morfológica propia del dominio médico que se puede formalizar y ser implantada en máquina, y (iii) incluye un neologismo que no posee una estructura morfológica, pero su categoría gramatical puede ser deducida automáticamente a través del contexto sintáctico. En la sección 3.1 se describe la metodología aplicada en este proyecto.

2.2. Generación y reconocimiento automático de neologismos

La creación de neologismos obedece a la necesidad constante de contar con las palabras adecuadas para dar cuenta de los cambios y descubrimientos de la sociedad. En contextos de producción de conocimiento, se presentan nuevos términos cuando (i) se descubre o inventa una nueva entidad; (ii) en un ámbito de traducción, cuando se requieren equivalentes para expresiones que, hasta el momento, solo eran mencionadas en su lengua de origen; y/o (iii) en contextos de planificación lingüística, para adoptar o adaptar préstamos de otras lenguas (Cabré *et al.*, 2012).

De acuerdo con Varo (2013), puede analizarse el reconocimiento de estas nuevas unidades en la medida en que se las puede considerar como una clase especial de pseudopalabras, «cuyos efectos en el proceso de acceso difieren bastante de los constatados en unidades consideradas palabras reales» (Varo, 2013:132). En cuanto a sus posibilidades de clasificación, Cabré (2006) analiza la propuesta del Observatori de Neología (OBNEO), que contempla los siguientes casos:

1. Neologismos de forma:
   a. Por afijación
   b. Por composición
   c. Por lexicalización
   d. Por conversión sintáctica
   e. Por sintagmación
   f. Por acortamiento (siglación, acronimia y abreviación)
   g. Por variación (variante formal ortográfica)
2. Neologismos sintácticos
3. Neologismos semánticos
4. Préstamos
5. Otros

A la vez, todos estos casos pueden englobarse en dos tipos: por un lado, los neologismos formales, como por ejemplo *vitritis* ('inflamación del humor vítreo'), que son los que se expresan mediante secuencias fonológicas que no existen como tal en el lexicón. Por otro, los casos de neología semántica, unidades que, si bien están fonológicamente presentes en el lexicón, se vinculan con significados que no se corresponden con el conjunto de contenidos asociados a cada entrada en cuestión (Varo, 2013); como, por ejemplo, el vocablo *paraguas* en el caso de *revisión paraguas*, que alude a la recopilación de información sobre una afección o enfermedad mediante varias revisiones sistemáticas.

En cuanto a las metodologías para el reconocimiento automático, desde hace tiempo se vienen desarrollando diversas herramientas para ello. De acuerdo con Janssen (2009) existen tres métodos: (i) utilización de una lista independiente de palabras conocidas —lista de exclusión—, (ii) recurrencia a patrones lingüísticos que caracterizan los neologismos, y (iii) contabilización de las ocurrencias de las palabras comparando un texto reciente con un corpus de textos de fechas anteriores.

La primera consiste en contrastar una lista de palabras conocidas, a la que se denomina lista de exclusión. Aquí, los candidatos a neologismo son aquellas expresiones que se encuentran dentro del corpus de estudio y no forman parte de la lista de exclusión. La segunda recurre a la utilización de patrones léxico-sintácticos que permiten que los neologismos sean reconocidos a partir de la observación de las palabras en su contexto. Por último, la metodología estadística consiste en cuantificar las ocurrencias de las palabras en un corpus de estudio, por lo general, en comparación con el corpus de referencia. Al respecto, Janssen (2009) señala, para esta última, cuatro aproximaciones. La primera no utiliza corpus de referencia, sino que considera neologismos a aquellas palabras que ocurren una sola vez en el texto. La segunda propone que los neologismos son aquellas palabras que tienen frecuencia cero en el corpus de referencia. La tercera consiste en comparar la frecuencia de todas las palabras en los dos corpus y, cuanto mayor sea la frecuencia de la palabra en el corpus de estudio en comparación con su frecuencia en el corpus de referencia, más probabilidades tendrá de ser un neologismo.

En el presente trabajo, se consideró neologismo a aquella palabra que no se encontraba en el diccionario de especialidad y el proceso de generación de definiciones está orientado a los neologismos de forma, productos de la derivación y la composición. Si bien esto puede considerarse una limitación, vale señalar que el método de generación de definiciones que se presenta fue probado en este tipo de expresiones con el propósito de mostrar su efectividad; sin embargo, la metodología podría ser aplicada a cualquier palabra médica que presente una combinación de formantes cultos.

Precisamente, en relación con los estudios llevados a cabo sobre neología en este dominio, para el caso del español, se pueden mencionar los trabajos descriptivos de Díaz (2001[a], 2001[b] y 2001[c]), quien se dedica a analizar diversos aspectos de la formación de términos médicos en relación con particularidades fonéticas, morfológicas, semánticas y etimológicas. En relación con los neologismos de forma, el autor se refiere a fenómenos como la acentuación en palabras terminadas en *-ía*, las variaciones del prefijo *des-* y los sufijos *-oides*, *-oide*, *-oideo*, *-oidal* y *-oídico*. En esta ocasión, se han tenido en cuenta los aportes de Díaz para las variaciones que pueden presentar los morfemas médicos al momento de elaborar la base de datos.

2.3. Sobre la noción de definición

En lo que atañe al concepto de definición, en el presente trabajo se toma en consideración el trabajo de Sierra (2009) sobre la extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos. Allí se expone una metodología para la detección automática de aquellas partes en el texto que definen los términos presentados por los autores. En la sección dedicada al análisis lingüístico de las definiciones, presenta una tipología sustentada en los modelos analíticos de Aguilar y Sierra (2008 y 2009) y Sierra *et al.* (2003), la cual se estructura de la siguiente manera:

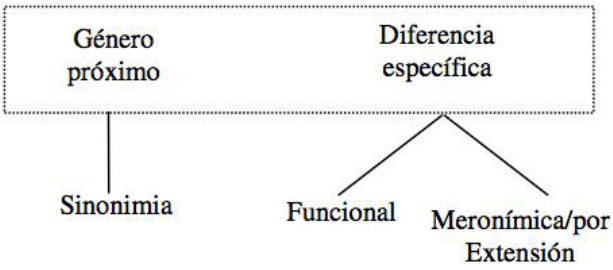


Figura 1. Estructura de la tipología de la definición, tomado de Sierra (2009:21)

Precisamente a partir de la relación observada entre la presencia o ausencia del género próximo y la diferencia específica, así como el tipo de predicación que introduce y asocia a la definición con un término, los autores observan cuatro tipos de definiciones básicas con los siguientes rasgos:

- Definición analítica o aristotélica: en este caso, la predicación verbal introduce explícitamente el género próximo y la diferencia específica; por ejemplo: «La dismutación es *una reacción química del tipo A+A→ A'+A' en la que dos especies químicas idénticas*, como dos moléculas de la misma saturación, *se transforman en dos especies distintas entre sí y de la especie original*». Adaptado del *Diccionario de Términos Médicos* (RANM, 2013).
- Definición sinonímica: la predicación hace explícito el género próximo y no considera ninguna diferencia específica: «La tos ferina es la *tos convulsa*».
- Definición funcional: se da cuando se reconoce solo la diferencia específica, la cual se describe como rasgo distintivo de un objeto, su función en un contexto dado: «Un bisturí se usa *para diseccionar*».
- Definición extensional: se presenta cuando la predicación introduce una definición en donde se explicita la diferencia específica sin mencionar el género próximo: «La sonda de Foley se compone de un catéter que dispone en su extremo distal de un globo hinchable con aire o con líquido para que pueda mantenerse en posición dentro de la vejiga» (RANM, 2013).

Las definiciones generadas por el algoritmo propuesto en el presente trabajo se correspondieron en su mayoría con la definición analítica o aristotélica, aunque también se observaron casos de definiciones sinonímicas.

No obstante, es necesario dejar claro lo que aquí se entenderá por definición. Al respecto, se propone una definición *ad hoc* del término que dé cuenta de sus principios fundamentales, pero que, a la vez, sea acorde con el trabajo computacional. Específicamente, en el presente trabajo se considera definición a la evidencia del significado de una expresión a partir de la explicitación de los elementos que componen su estructura morfológica. Como se mencionó en la introducción, la noción de explicitación está relacionada con el recurso de traducción —que consiste en aportar información descriptiva en una lengua meta—, que permite com-

prender de mejor manera un concepto de una lengua origen (Herrezuelo, 2008). No obstante, se diferencia del procedimiento de la traducción en que no se trata de dos idiomas diferentes, sino que se explicita una lengua especializada, en este caso el lenguaje de la medicina, a partir de expresiones propia de la lengua general. Dicho de manera simple, la explicitación consiste en transformar una palabra en una expresión perifrástica equivalente que exprese lo que está implícito en dicho término.

A tales efectos, definir un término perteneciente al lenguaje especializado de la medicina es, en esencia, traducirlo mediante el procedimiento de la explicitación, de modo que se logra una expresión compuesta por expresiones más generales. A modo de ejemplo, el término *electrocorticograma* se define como 'representación gráfica electrónica de la corteza cerebral'. Ahora bien, el procedimiento para lograr dicha definición consiste en: (i) reconocer los morfemas que componen la palabra —*electro*, *cortico* y *grama*—; (ii) explicitar dichos morfemas —*electro*: 'electrónico/a'; *cortico*: 'corteza cerebral', y *grama*: 'representación gráfica'—; (iii) reordenar las explicitaciones; (iv) añadir elementos funcionales —determinantes, preposiciones, etcétera— y adecuar los rasgos de flexión para las exigencias de concordancia.

De este modo, el objetivo del presente trabajo fue establecer una formalización de los pasos señalados, a fin de desarrollar una implantación computacional que permitiera generar definiciones para los neologismos de forma. Para ello, se tomó la lista de neologismos médicos obtenida a partir de una metodología de extracción basada en el procesamiento de información lingüística planteada en el proyecto Fondecyt 11130469. La implantación realizada se describe en la sección siguiente.

3. Metodología

La metodología se compone de los siguientes pasos: (i) detección automática de neologismos de forma en un corpus; (ii) elaboración de una base de datos de lexemas, sufijos y prefijos del dominio médico —formantes cultos—, junto con sus significados; (iii) segmentación de los neologismos de acuerdo con los formantes que los componían y asignación de significados; y (iv) generación de reglas de reescritura. A continuación, se detallan dichas etapas de trabajo.

3.1. Detección automática de neologismos de forma

En primer lugar se realizó un trabajo de extracción de candidatos a término que implicó desarrollar reglas de reconocimiento en el nivel léxico, morfológico y sintáctico. Para el nivel léxico, la detección fue realizada mediante la aplicación de diccionarios médicos (Mosby, 2005; Navarro, 2005; Cárdenas, 2012; RANM, 2013), junto con un listado de medicamentos extraído del sitio http://www.vademecum.es (consulta: 31.V.2016). Las expresiones no incluidas en los diccionarios fueron consideradas neologismos y sometidas a un proceso de deducción de la categoría gramatical a partir del análisis morfológico y sintáctico. Para el trabajo computacional, se utilizaron los softwares Smorph (Aït Mokthar, 1998) y Módulo Post Smorph (MPS) (Abacci, 1999), que trabajan en bloque, y Xfst (Beesley y Karttunen, 2003). Smorph realiza el análisis morfológico y MPS se aplica sobre gramáticas locales. Xfst, por su parte, es una herramienta de estados finitos que opera sobre cadenas de caracteres a las que asigna categorías previamente declaradas. Las etapas de trabajo fueron las siguientes:

- **Etapa I:** análisis morfosintáctico y reconocimiento de los signos de puntuación por medio del etiquetador Smorph. A las palabras desconocidas se les asignó la etiqueta PD, y a las palabras incluidas en los diccionarios de especialidad, MED.
- **Etapa II:** reconocimiento de candidatos a términos a partir de estructuras morfológicas mediante XFST. Si la expresión contenía, al menos, un formante culto con un mínimo de tres caracteres, se lo etiquetaba como MORF.
- **Etapa III:** creación y aplicación de reglas sintácticas para deducir la categoría de las PD no compuestas por morfemas propios del área.
- **Etapa IV:** extracción en calidad de candidatos a término de aquellos SSNN que incluían al menos una expresión PD, MED o MORF y se correspondieran con una de las siguientes estructuras: (i) unigramas: nombre (*cáncer*); (ii) bigramas: nombre + adjetivo (*cáncer mamario*), y (iii) trigramas: nombre + preposición *de* + nombre (*cáncer de mama*).

Esta metodología fue probada en el Corpus de Casos Clínicos Médicos (CCM-2009), compilado por Burdiles (2012), que contiene casos clínicos aparecidos entre los años 1999 y 2008 y que suma un total de 969 textos y 801 224 palabras. Los textos reunidos pertenecen a las especialidades de parasitología, neuropsiquiatría, enfermedades respiratorias, otorrinonaringología y cirugía, infectología, pediatría, obstetricia, cirugía y ciencias biomédicas y, finalmente, medicina interna y especialidades derivadas.

El trabajo de generación de definiciones fue realizado con los unigramas que contenían la etiqueta MORF —en trabajos futuros se abordarán los casos correspondientes con bigramas y trigramas—. Dichas expresiones sumaban un total de 190 neologismos.

3.2. Elaboración de bases de datos

El proceso de generación automática de definiciones del dominio médico fue desarrollado en Excel y comenzó con la elaboración de una base de datos de morfemas del ámbito —raíces, sufijos y prefijos—. Estas partículas léxicas y morfológicas se extrajeron del *Diccionario médico-biológico, histórico y etimológico* (*Dicciomed*), publicado en versión electrónica por la Universidad de Salamanca (versión 2011). En dicha colección se almacenan 7052 palabras, que incluyen entradas como *abductor*, *dactilomegalia* o *lecitoblasto* y, a la vez, se incluyen 3233 partículas correspondientes a lexemas y sufijos clasificados por conceptos —como «acciones, generar, nacer», «biología, taxonomía» o «fisiología, respiración»—. A modo de ejemplo, la entrada del lexema *ito* presenta la siguiente fisonomía:

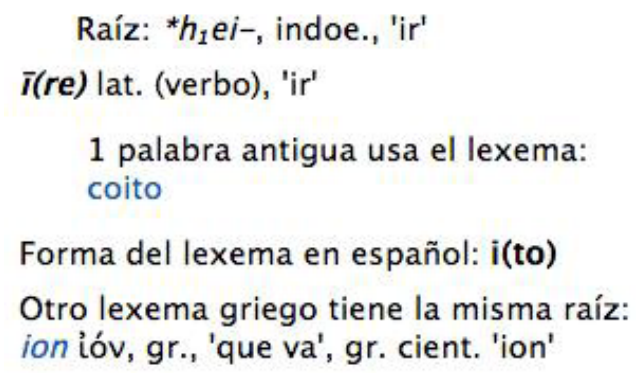


Figura 2. Ejemplo de entrada en *Dicciomed*

En la imagen se puede observar la exposición de la raíz —en este caso proveniente del latín—, la traducción de la misma —'ir'—, la presencia de la raíz en las palabras del diccionario —coito—, las variaciones del lexema en español —i(to)— y la clasificación en conceptos. En la siguiente tabla, se presenta un fragmento de la base de datos elaborada:

Tabla 1. Fragmento de la base de datos elaborada a partir de la información extraída de *Dicciomed*

Forma en español	Significado	Concepto	Prefijo	Sufijo	Raíz
bostezar	boca	Acciones, abrir			Raíz
apto	atar	Acciones, agarrar, fijar, unir			Raíz
grama	rayar	Acciones, escribir/leer		Sufijo	
grafía	rayar	Acciones, escribir/leer		Sufijo	
re	repetición	Acciones, repetir	Prefijo		
frac	romper	Acciones, romper	Prefijo		

A los efectos de la presente investigación se ha extraído el total de entradas del diccionario en conjunto con sus formas en español, así como sus definiciones o traducciones. El resultado de dicho proceso ha sido la generación de una base de datos de tres columnas, correspondientes a la entrada en español, la traducción y el concepto asociado. Posteriormente, se realizó el mismo procedimiento a nivel palabras, con el fin de evitar definiciones etimológicas, como el caso de *colágeno*, que en un primer momento fue definido como 'el generador de cola', lo que ocasionó inconvenientes en la definición de términos de mayor complejidad, como, por ejemplo, *colagenopatía*, que primeramente se definió como 'la enfermedad de el generador de cola' —por el momento, las definiciones no contienen contracciones, lo que implica, en algunos casos, la aparición de la expresión «de el»—.

En un tercer momento se desagregó cada forma en español de acuerdo con las opciones que entrega el diccionario mediante el uso de paréntesis, comas o guiones. Por ejemplo, en el caso indicado anteriormente, la partícula *i(to)* permite dos posibilidades: 1) como *i*, o 2) como *ito*. El máximo de posibilidades de combinaciones de segmentos de morfema para una entrada lexemática es siete.

Llevada a cabo dicha desagregación, el total de entradas es de 8398 alomorfos a partir de las 3233 partículas originales, cada uno asociado a una traducción y a un concepto. El objetivo de esta expansión de las formas es el de explicitar todos los modos en los que un morfema puede presentarse en el interior de una palabra: como raíz, como prefijo o como sufijo.

3.3. Segmentación y análisis de los neologismos

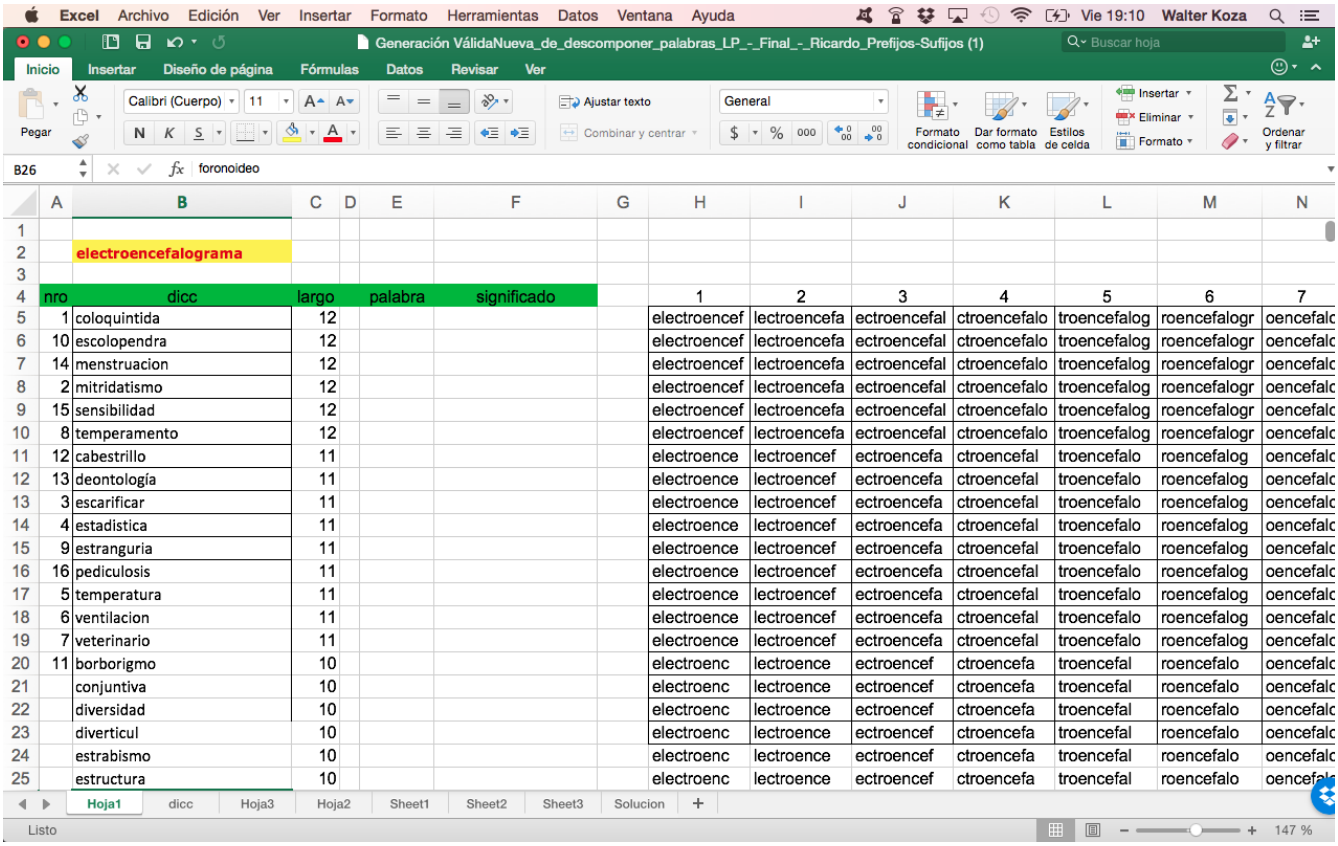
A continuación se ha desarrolló un procedimiento para segmentar las palabras del diccionario a fin de emparejar los segmentos con las partículas lexemáticas y de sufijos. Por ejemplo, en el caso de la palabra *electroencefalograma* se ha segmentado en tres unidades, a saber: *electro*, *encéfalo* y *grama*. Para cada una de ellas se ha extraído la definición de la base de datos —'imagen electrónica', 'cerebro' y 'representación gráfica', respectivamente—. No obstante, estas definiciones fueron adaptadas, esto es, modificadas en su conformación sintáctica a fin de que pudieran combinarse unas con otras para formar una cláusula coherentemente estructurada. De este modo, las definiciones del ejemplo fueron modificadas de la siguiente manera:

- 'imagen electrónica' → 'electrónico/a'
- 'cerebro' → 'el cerebro'
- 'representación gráfica' → 'la representación gráfica de'

De este modo, mediante reglas de reordenamiento, se logra la siguiente definición para *electroencefalograma*: 'la representación gráfica de el cerebro electrónico/a'. Como se puede observar, una de las dificultades del presente trabajo radicó en la concordancia.

El procedimiento ha eliminado los acentos gráficos o tildes a fin de evitar posibles disimilitudes entre la lista de partículas léxicas y sufijales y el modo en que estas se presentan en los términos. Por otro lado, en este mismo entendido, el sistema de segmentación ha operado desde las segmentaciones más extensas hacia las más breves con el objetivo de extraer en primer lugar los tramos más desarrollados de cada palabra que se pueden emparejar con la lista de la base de datos, siguiendo el procedimiento planteado por Rehman *et al.* (2013), que proponen como uno de los modos más eficientes para la segmentación morfemática el que ordena los morfemas desde los segmentos más extensos hasta los más breves —*longest matching technique*—. Esta regla permite elegir las segmentaciones más adecuadas para cada palabra.

De esta manera, en *encefalograma* se evitan posibles reconocimientos erróneos de segmentos, ya que *grama*, al ser más extenso, tiene prevalencia sobre los lexemas *gram* y *am*. Para objetivos prácticos, el orden de búsqueda de los 8398 alomorfos se ha ordenado desde los alomorfos más extensos en número de caracteres hasta los más breves. Este procedimiento fue realizado con las funciones Extrae y Buscar de Excel. Cada palabra se va separando de a un carácter y, a medida que los fragmentos coinciden con la lista de morfemas, se les van asignando a estos los significados correspondientes (v. figura 3).



nro	dicc	largo	palabra	significado	1	2	3	4	5	6	7
1	coloquintida	12			electroencef	lectroencefa	ectroencefal	ctroencefalo	troencefalog	roencefalogr	oencefalc
10	escolopendra	12			electroencef	lectroencefa	ectroencefal	ctroencefalo	troencefalog	roencefalogr	oencefalc
14	menstruacion	12			electroencef	lectroencefa	ectroencefal	ctroencefalo	troencefalog	roencefalogr	oencefalc
2	mitridatismo	12			electroencef	lectroencefa	ectroencefal	ctroencefalo	troencefalog	roencefalogr	oencefalc
15	sensibilidad	12			electroencef	lectroencefa	ectroencefal	ctroencefalo	troencefalog	roencefalogr	oencefalc
8	temperamento	12			electroencef	lectroencefa	ectroencefal	ctroencefalo	troencefalog	roencefalogr	oencefalc
12	cabestrillo	11			electroence	lectroencef	ectroencefa	ctroencefal	troencefalo	roencefalog	oencefalc
13	deontología	11			electroence	lectroencef	ectroencefa	ctroencefal	troencefalo	roencefalog	oencefalc
3	escarificar	11			electroence	lectroencef	ectroencefa	ctroencefal	troencefalo	roencefalog	oencefalc
4	estadistica	11			electroence	lectroencef	ectroencefa	ctroencefal	troencefalo	roencefalog	oencefalc
9	estranguria	11			electroence	lectroencef	ectroencefa	ctroencefal	troencefalo	roencefalog	oencefalc
16	pediculosis	11			electroence	lectroencef	ectroencefa	ctroencefal	troencefalo	roencefalog	oencefalc
5	temperatura	11			electroence	lectroencef	ectroencefa	ctroencefal	troencefalo	roencefalog	oencefalc
6	ventilacion	11			electroence	lectroencef	ectroencefa	ctroencefal	troencefalo	roencefalog	oencefalc
7	veterinario	11			electroence	lectroencef	ectroencefa	ctroencefal	troencefalo	roencefalog	oencefalc
11	borborigmo	10			electroenc	lectroence	ectroencef	ctroencefa	troencefal	roencefalo	oencefalc
	conjuntiva	10			electroenc	lectroence	ectroencef	ctroencefa	troencefal	roencefalo	oencefalc
	diversidad	10			electroenc	lectroence	ectroencef	ctroencefa	troencefal	roencefalo	oencefalc
	diverticul	10			electroenc	lectroence	ectroencef	ctroencefa	troencefal	roencefalo	oencefalc
	estrabismo	10			electroenc	lectroence	ectroencef	ctroencefa	troencefal	roencefalo	oencefalc
	estructura	10			electroenc	lectroence	ectroencef	ctroencefa	troencefal	roencefalo	oencefalc

Figura 3. Cálculo de segmentaciones posibles para una palabra

3.4. Elaboración de reglas de reescritura

Una vez que se dispone de todas las coincidencias entre segmentos de la palabra por traducir y la lista de partículas léxicas, prefijales y sufijales, corresponde realizar un ordenamiento y reescritura de dichas partículas. Para ello, se ha operado con once reglas:

1. pref - raíz (1) - raíz (2) → pref - raíz (1) - raíz (2)
2. pref - raíz - suf → suf - pref - raíz
3. raíz (1) - pref - raíz (2) → raíz (1) - pref - raíz (2)
4. pref (1) - pref (2) - raíz → pref (1) - pref (2) - raíz
5. pref (1) - pref (2) → raíz - suf suf - pref (1) - pref (2) - raíz
6. raíz (1) - raíz (2) - raíz (3) → raíz (3) - raíz (2) - raíz (1)
7. raíz (1) - raíz (2) - suf → suf - raíz (1) - raíz (2)
8. pref - raíz → pref - raíz
9. raíz (1) – raíz (2) → raíz (2) - raíz (1)
10. raíz - suf → suf - raíz
11. raíz → raíz

Estas reglas seleccionan once tipos frecuentes de neologismos del dominio médico y permiten reescribir y, por ende, generar las definiciones. Para ello, el sistema automático asigna a cada partícula un valor que puede corresponder a tres categorías diferentes: prefijo, sufijo y raíz, que son extraídos de la tabla de tres columnas que contiene el valor en caracteres del alomorfo, su definición y su categoría morfológica. Del mismo modo, el sistema puede asignar a cada ocurrencia de una misma categoría un valor numérico ordenado de izquierda a derecha. Por ejemplo, si el sistema encuentra dos raíces en la palabra, a la que primero aparece de izquierda a derecha le asigna el valor Raíz (1) y a la segunda le asigna el valor Raíz (2). La tabla de reescrituras reordenará estos elementos en las circunstancias en que resulte necesario.

El paso final consiste en extraer de la tabla de traducciones aquellas traducciones correspondientes a cada uno de los elementos reordenados en ese mismo último orden para dar con una definición del término. Mediante la elaboración de una macro, se pasaron los 190 neologismos de forma detectados y se generó, en cada uno de ellos, una explicitación correspondiente. A continuación se presentan ejemplos de las definiciones generadas:

coronariografia: 'la representación gráfica de la arteria coronaria'
antibradicardia: 'lo contrario a el ritmo lento de el corazón' *histerosonografia*: 'la representación gráfica de el útero por sonido'
vitreitis: 'la inflamación de el humor vítreo'

En la sección siguiente se analizan los resultados obtenidos.

4. Resultados

El método presentado, como se mencionó más arriba, fue probado en una lista de 190 neologismos. De dicho total, se generaron 164 definiciones, de las cuales 122 fueron correctas, lo que implicó un 74,34% de precisión, un 64,21% de cobertura (*recall*) y un 68,9% de medida F.

En relación con las estructuras observadas, se pudo apreciar que la más frecuente fue la estructura Raíz (1) Raíz (2), seguida de Prefijo Raíz. En la tabla 2, se presenta la frecuencia de cada una de ellas.

Tabla 2. Frecuencia de las estructuras de los neologismos

Estructura	Número de casos	%	% acumulado
Raíz Raíz	60	31,58	31,91
Prefijo Raíz	31	16,32	48,40
Prefijo Raíz Raíz	17	8,95	57,45
Raíz	16	8,42	65,96
Raíz Raíz Raíz	13	6,84	72,87
Raíz Sufijo	12	6,32	79,26
Raíz Raíz Sufijo	6	3,16	82,45
Raíz Prefijo Raíz	5	2,63	85,11
Prefijo Raíz Sufijo	4	2,11	87,23
Prefijo Prefijo	2	1,05	88,30
Prefijo Prefijo Raíz Raíz	2	1,05	89,36
Raíz Prefijo	2	1,05	90,43
Raíz Sufijo Raíz Raíz	2	1,05	91,49
Prefijo	2	1,05	92,02
Prefijo Prefijo Prefijo	2	1,05	92,55
Prefijo Prefijo Prefijo Raíz	1	0,53	93,09
Prefijo Prefijo Raíz Prefijo	1	0,53	93,62
Prefijo Raíz Prefijo	1	0,53	94,15
Prefijo Raíz Raíz Sufijo	1	0,53	94,68
Prefijo Raíz Sufijo Prefijo	1	0,53	95,21
Prefijo Sufijo Prefijo Raíz	1	0,53	95,74
Prefijo Sufijo Raíz	1	0,53	96,28
Prefijo Sufijo Sufijo	1	0,53	96,81
Raíz Prefijo Prefijo Raíz	1	0,53	97,34
Raíz Prefijo Raíz Raíz	1	0,53	97,87
Raíz Raíz Prefijo Raíz	1	0,53	98,40
Raíz Raíz Prefijo Sufijo	1	0,53	98,94
Raíz Raíz Raíz Raíz	1	0,53	99,47
Raíz Sufijo Prefijo Raíz	1	0,53	100,00
	190	100,00	

Como se puede apreciar, se hallaron estructuras no consideradas en el momento de establecer las reglas de reescritura, como por ejemplo en casos como el de *taquicardiomiopatía*, cuya estructura es: Prefijo (*taqui*) - Raíz (*cardio*) - Raíz (*mio*) - Raíz (*patía*). Dado que la secuencia de los cuatro lexemas no se encontraba contemplada, el algoritmo no generó ninguna definición.

No obstante, en algunos casos, las estructuras propuestas no se correspondían con la organización morfológica de la palabra, y esto es debido a que, al no hallarse ningún lexema en la base de datos, se realizaba una segmentación errónea. Tal fue el caso de términos como *neurodesarrollo*. Dado que *desarrollo* no se encontraba en la base de datos, la segmentación fue la siguiente:

Tabla 3. Segmentación de *neurodesarrollo*

Lexema 1	arroll	envolver en forma de rollo
Lexema 2	neuro	lo neuronal
Lexema 3	des	la inversión de una acción

A partir de esto se generó la siguiente definición:

> **neurodesarrollo*: 'lo neuronal la inversión de una acción envolver en forma de rollo'

Otro de los problemas que se observó fue la coocurrencia de lexemas pertenecientes al mismo dominio que deberían aparecer coordinados, como por ejemplo *transcardiopulmonar*, cuya estructura es: Prefijo (*trans*) - Raíz (*cardio*) - Raíz (*pulmón*) - Sufijo (*ar*). Debido a que *cardio* y *pulmón* pertecen al subdominio de la anatomía, y que en la base de datos se encuentra la expresión *pulmonar* clasificada como Raíz, se generó la siguiente construcción:

> **transcardiopulmonar*: 'a través de el corazón lo relativo al pulmón'.

No obstante, a pesar de los inconvenientes observados, las reglas de reescritura tuvieron resultados altos en las estructuras más frecuentes. En la tabla 4 se presentan los números de aciertos y errores en cada definición.

Tabla 4. Resultados obtenidos por estructura

Estructura	Casos	Aciertos	Errores	Precisión
Raíz Raíz	60	68	12	80%
Prefijo Raíz	31	26	5	83,87%
Prefijo Raíz Raíz	17	11	6	64%
Raíz	16	10	6	62,5%

Estructura	Casos	Aciertos	Errores	Precisión
Raíz Raíz Raíz	13	6	7	46,15 %
Raíz Sufijo	12	10	2	83,33 %
Raíz Raíz Sufijo	6	5	1	83,33 %

En las definiciones logradas para los casos de los neologismos conformados por Raíz Raíz, se obtuvo un 80 % de efectividad, al lograr 68 definiciones correctas en 80 expresiones que presentaban dicha estructura. Los resultados más bajos se dieron en las expresiones compuestas por tres raíces —como el caso de *hepatocoledoco*, *poliquimioterapia*, *ectoidectomia*, *glucoglicinuria*, entre otros—, en donde solo en 6 de los 7 casos se lograron definiciones adecuadas. Esto se debió, principalmente, a casos de composición en que dos raíces léxicas pertenecientes a la misma subárea debían unirse mediante coordinación, por ejemplo en el caso de *coloproctologo*, cuya definición generada fue 'el estudioso de el ano el colon'. En el trabajo futuro se pretende incluir en las reglas de reescritura información acerca del área médica a la que pertenece cada lexema —anatomía, química, etcétera— a los efectos de poder establecer con mayor precisión los elementos de coordinación en cada significado.

5. Consideraciones finales

Se presentó un método automático de escritura de definiciones para términos médicos compuestos por formantes cultos aplicados a neologismos de forma basado en el procesamiento de información morfológica. A fin de lograr dicho objetivo, se reunió la información morfológica alojada en *Dicciomed*, adaptándola a los requerimientos propuestos. Al respecto, vale aclarar que, si bien se agregaron elementos que actuaran a modo de enlace, no se modificó la información semántica contenida en las definiciones de los morfemas. Específicamente, se agregaron determinantes y preposiciones —«representación gráfica» por «la representación gráfica de»—. Los resultados obtenidos demuestran que el método propuesto es válido para este tipo de tareas, aunque es pertinente ampliar la base de datos.

Dentro de las limitaciones, se pueden observar algunos problemas en cuanto a la ilación de los elementos y sus posiciones, y, en el caso de adjetivos, variaciones de género y número. Para ello, una solución establecida fue la inclusión de elementos optativos, como, por ejemplo, comas y variaciones de género (el/la). De este modo, un morfema como *-emia* fue cargado junto con las posibles preposiciones con la que podía aparecer: *emia*: '(en/de) la sangre(,)'.

Otro problema radicó en la obtención de definiciones etimológicamente correctas, pero que requirieron ser actualizadas. Tal es el caso mencionado más arriba de *colágeno* (*cola*, *geno*), cuya primera definición fue 'generador de cola'.

En relación con los aportes, por un lado el presente trabajo contribuye a los estudios de morfología, al presentar un segmentador y analizador de términos médicos. Conviene aclarar que, si bien la metodología se probó en los neologismos, dicha herramienta es pertinente para segmentar toda expresión formada a partir de morfemas del área médica. Como ejemplo, un término como *otorrinolaringólogo* es segmentado de la siguiente manera:

Tabla 5. Segmentación de *otorrinolaringólogo*	
oto	el oído
rino	la nariz
laringo	la laringe
logo	dedicado a

De este modo, el método proporciona una herramienta de ayuda para lexicógrafos, traductores, profesionales de la salud y público en general. La segmentación de mayor a menor, esto es, del morfema con más letras a los de menor extensión, permite lograr el reconocimiento de los componentes de los neologismos sin inconvenientes. De hecho, los casos erróneos, como se mencionó en la sección anterior, se debieron a limitaciones de la base de datos, pero no a las reglas de segmentación.

En relación con la asignación de significados a cada morfema, se pudieron observar algunos inconvenientes derivados de la polisemia, como en el caso de *híper*, que puede significar 'grande', 'en exceso', 'más que', etcétera; en este caso se optó por tomar el primer significado que presenta *Dicciomed*. Asimismo, también se evidenció un caso de ambigüedad en *test*, que puede significar tanto 'examen' como 'testículo'. Dada la frecuencia de aparición del primer caso, se optó por no incluir el segundo. No obstante, es pertinente tener en cuenta esta particularidad en próximas operaciones de reescritura.

En cuanto a las definiciones logradas, estas se correspondieron en su mayoría con las definiciones aristotélicas. Esto se justifica en la medida en que uno de los lexemas se vuelve núcleo del sintagma nominal definitorio Y constituye el género próximo, mientras que los demás aportan las diferencias específicas. A modo de ejemplo, se puede observar el caso de *coronariografía* y *vitreitis*:

- *coronariografía*: 'la representación gráfica [género próximo] de la arteria coronaria [diferencia específica]'
- *vitreitis*: 'inflamación [género próximo] de el humor vítreo [diferencia específica]'

Asimismo, se encontraron también definiciones sinónimas, en las que solo se explicitaba el género próximo. Entre estos casos se puede mencionar *hipomagnesemia* —'bajo/a magnesio (en/de) la sangre'— y *comitógeno* —'algo que actúa en conjunto con la sustancia que induce la mitosis'—.

El trabajo a futuro se organiza en torno a los siguientes ejes: en primer lugar, se ampliará la base de datos de formantes cultos del dominio médico, incluyendo aquellos pertenecientes a las subáreas de la química y la farmacología. En

segundo lugar, se revisarán las reglas de reescritura, a la vez que se propondrán otras que no fueron consideradas, como por ejemplo Prefijo Prefijo Raíz. Asimismo, se planteará una combinatoria entre la información morfosintáctica del lexema y el área a la que pertenece a fin de establecer con mayor definición las partículas funcionales, como preposiciones y conjunciones.

Referencias bibliográficas

Abacci, Faiza (1999): *Développement du module post-smorph*. Clermont-Ferrand: Universidad Blaise-Pascal.

Aguilar, César y Gerardo Sierra (2008): «Hacia una tipología de definiciones basada en el modelo analítico», en *Memorias del XV Congreso Internacional ALFAL*. Montevideo: ALFAL. Disponible en <http://alfal.easyplanners.info/programa/buscar.php#> [consulta: 27.XI.2016].

Aguilar, César y Gerardo Sierra (2009): «Reconocimiento de definiciones asociadas a frases predicativas en contextos definitorios», *Procesamiento de Lenguaje Natural*, 43: 151-158.

Aït Mokthar, Salah (1998): *SMORPH: guide d'utilisation: rapport technique*. Clermont-Ferrand: Universidad Blaise Pascal.

Alcántara, Berenice (2013): «Evangelización y traducción. La Vida de san Francisco de san Buenaventura vuelta al náhuatl por fray Alonso de Molina», *Estudios de cultura Náhuatl*, 46: 89-158.

Barrón, Luis (2007): *Extracción automática de contextos definitorios*. Tesis de maestría. México D. F.: Universidad Nacional Autónoma de México.

Beesley, Kenneth y Lauri Karttunen (2003): *Finite state morphology*. Stanford: CSLI Stanford University.

Burdiles, Gina (2012): *Descripción de la organización retórica del género caso clínico de la medicina a partir del corpus CCM-2009*. Tesis doctoral. Valparaíso: Pontificia Universidad Católica de Valparaíso.

Cabré, M. Teresa (2002): «Textos especializados y unidades de conocimiento: metodología y tipologización». En Joaquín García Palacios y María Teresa Fuentes Morán (eds.): *Texto, terminología y traducción*. Salamanca: Ediciones Almar, pp. 15-36.

Cabré, M. Teresa (2006): «La clasificación de neologismos: una tarea compleja», *Alfa: revista de lingüística*, 50 (2): 229-250.

Cabré, M. Teresa; Rosa Estopà y Chelo Vargas (2012): «Neology in specialized communication», *Terminology*, 18 (1): 1-8.

Cárdenas, E. (2012): *Terminología médica*. México D. F.: Mac Craw Hill.

Díaz, José (2001ª): «Nociones de neología. La formación de derivados y compuestos a partir de nombres propios de personas», *Panace@*, 2 (5): 25-30.

Díaz, José (2001ᵇ): «Nociones de neología: el prefijo –des», *Panace@*, 2 (6): 83-84.

Díaz, José (2001ᶜ): «Nociones de neología: los sufijos -oides, -oide, -oideo, -oidal y -oídico en terminogía médica», *Panace@*, 2 (3): Panace@, 2 (3): 67-70.

Frantzi, Katerina y Sophia Ananiadou (1996): «Extracting nested collocations», en *16th Conference on Computational Linguistics*. Copenhague: Association for Computational Linguistics, pp. 41-46.

Gálvez, Carmen (2012): «Reconocimiento y anotación de nombres de fármacos genéricos en la literatura biomédica», *Acimed*, 4: 326-345.

Herrero-Zorita, Carlos; Clara Molina y Antonio Moreno-Sandoval (2015): «Medical term formation in English and Japanese», *Review of Cognitive Linguistics*, 13 (1): 81-105.

Herrezuelo, María Inmaculada (2008): «Estudio de la explicitación en dos publicaciones periódicas gratuitas bilingües (*Ronda Iberia* y *Sur in English*). Análisis de casos», *Trans*, 12: 169-188.

Janssen, Maarten (2009): «Detección de neologismos: una perspectiva computacional», *Debate terminológico*, 5: 68-75.

Kageura, Kio y Bin Umino (1996): « Methods of automatic term recognition. A review», *Terminology*, 3 (2): 259-289.

Koza, Walter (2015): «Proposal for automatic extraction of medical term candidates with linguistic information processing description and evaluation of results», *Alfa. Revista de Linguística*, 59 (1): 113-125.

Koza, Walter; Mirian Muñoz y María José Mánquez (2015): «Desarrollo de un diccionario electrónico para la detección automática de candidatos a término del dominio médico. Una aplicación con Smorph y MPS», en Sociedad Argentina de Informática: *CAIS 2015*, Congreso Argentino de Informática y Salud. Rosario: Sociedad Argentina de Informática, pp. 1-11.

Krauthamer, Michael y Goran Nenadić (2004): «Term identification in the biomedical literature», *Journal of Biomedical Informatics*, 37: 512-526.

López Huertas, M.; I. Torres y M. Barité (2004): «Terminological representation for specialized areas in conceptual structures: the case of gender studies», en M. López Huertas, M. Barité e I. Torres (eds.): *Proceedings of the 8th International ISKO Conference*. Londres: Ia C. McIlwaine, pp. 263-268.

Marciniak, Małgorzata y Agnieszka Mykowiecka (2015): «Nested term recognition driven by word connection strength», *Terminology*, 21 (2): 180-204.

Marincovich, J. (2008): «Palabra y término. ¿Diferenciación o complementación?», *Revista Signos. Estudios de Lingüística*, 41 (67): 119-126.

Mosby (2005): *Diccionario Mosby*. Versión electrónica. Madrid: Harcourt.

Navarro, Fernando A. (2005): *Diccionario crítico de dudas inglés-español de medicina* (2.ªed.). Madrid: McGraw-Hill Interamericana.

Park, Youngja; Roy Byrd y Branimir Boguraev (2002): «Automatic glossary extraction: beyond terminology identification», en *Proceedings of the 19th International Conference on Computational Linguistics*. Taipei: Association for Computational Linguistics, pp. 1-7.

Pazienza, María Teresa; Marco Pannacchiotti y Fabio Zanzotto (2005): «Textual entailment as syntactic graph distance: a rule-based adn svm based approach», en *Proceedings of the PASCAL Challenges Workshop on recognizing textual entailment*. Southampton: PASCAL, pp. 25-28.

Pecina, Pavel (2010): «Lexical association measures and collocation extraction», *Language resources and evaluation*, 44 (1): 137-158.

Periñán-Pascual, Carlos (2015): «The underpinnings of a composite measure for automatic term extraction. The case of SRC», *Terminology*, 21 (2): 151-179.

Real Academia Nacional de Medicina (2012). *Diccionario de términos médicos*. Buenos Aires: Editorial Médica Panamericana.

Rehman, Zobia; Waqas Anwar, Usama Ijaz Bajwa, Wang Xuan y Zhou Chaoying (2013): «Morpheme Matching Based Text Tokenization for a Scarce Resourced Language», *PLoS One*, 8 (8). <http://

journals.plos.org/plosone/article?id=10.1371/journal.pone.0068178> [consulta: 27.XI.2016].

Segura, Isabel; Paloma Martínez y Doaa Samy (2008): «Detección de fármacos genéricos en textos biomédicos», *Procesamiento de Lenguaje Natural*, 40: 27-34.

Sierra, Gerardo (2009): «Extracción de contextos definitorios en textos de especialidad a partir del reconocimiento de patrones lingüísticos», *LinguaMática*, 2: 13-38.

Sierra, Gerardo; A. Medina, R. Alarcón y César Aguilar (2003): «Towards the extraction of conceptual information from corpora», en D. Archer, P. Rayson, A. Wilson y T. McEnery (eds.): *Proceedings of the Corpus Linguistics 2003 conference*. Lancaster: Lancaster University, pp. 691-697.

Silberztein, Max (2016): *Formalizing Natural Languages. The NooJ approach*. Londres: ISTE.

Soto, Jorge (2013): «La traducción de términos culturales en el contexto turístico español-inglés: recepción real en usuarios anglófonos», *Quaderns. Revista de Traducció*, 20: 235-250.

Tuason, O.; L. Chen, H. Liu, J. Blake y C. Friedman (2004): «Biological nomenclature: a source of lexical knowledge and ambiguity», en *Pacific Symposium of Biocomputing*. Oak Ridge: PSB, pp. 238-249.

Varo, Carmen (2013): «Aproximación teórico-práctica al procesamiento lingüístico de neologismos léxicos», *Revista Signos. Estudios de Lingüística*, 46 (81): 132-152. <http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342013000100006> [consulta: 27.XI.2016].

Vinay, Jean-Paul y Jean Darbelnet (1995): *Comparative Stylistics of French and English: A Methodology for Translation*. Ámsterdam: John Benjamins. Traducción al inglés de J. Sager.

Vivaldi, Jorge (2003): *Sistema de extracción de candidatos a término YATE. Manual de utilización*. Barcelona: IULA, Universidad Pompeu Fabra.

Vivaldi, Jorge y Horacio Rodríguez (2015). «Medical entities tagging using distant learning», *Computational Linguistics and Intelligent text processing*, 9042: 631-642.