Metodología para la extracción e identificación de candidatos a términos en el ámbito de la bioquímica

Coral López Mateo
Universitat Politècnica de València
clopezm@idm.upv.es

Françoise Olmo Cazevieille
Universitat Politècnica de València
folmo@idm.upv.es

Françoise Olmo Cazevieille és doctora en Filologia Francesa i professora titular del Departament de Lingüística Aplicada de la Universitat Politècnica de València (UPV). Actualment és membre responsable del grup d'investigació GALE (Grup d'Anàlisi de Llengües d'Especialitat), professora de terminologia en el màster Llengües i Tecnologia i de llengua francesa en l'Escola Superior d'Enginyeria Agronòmica i del Medi Natural. Les seves línies d'investigació són la terminologia cientificotècnica i la didàctica de llengües amb fins específics, en particular, la metodologia i les tecnologies digitals.

Coral López Mateo és llicenciada en Filologia Alemanya per la Universitat de València i és professora del Departament de Lingüística Aplicada de la Universitat Politècnica de València (UPV), on imparteix docència en assignatures de llengua alemanya. Actualment es troba realitzant una tesi doctoral a la UPV centrada en la descripció terminològica de la llengua d'especialitat alemanya, en concret, en l'àmbit de la bioquímica.

Resum

Metodologia per a la extracció i identificació de candidats a termes en l'àmbit de la bioquímica

En aquest article descrivim el procés d'extracció de candidats a termes de textos alemanys de l'àmbit de la bioquímica. A més, resolem de manera efectiva el silenci generat en l'extracció a través de la recerca de «termes mare» (Ahmad i Rogers, 2001) i de morfemes específics del domini (Heid, 1998). La metodologia aplicada és extrapolable a altres camps científics relacionats.

PARAULES CLAU: *extracció terminològica; sufixos de termes especialitzats; criteris d'identificació; bioquímica.*

Resumen

En este artículo describimos el proceso de extracción de candidatos a términos de textos alemanes del ámbito de la bioquímica a partir de una herramienta de análisis textual. Resolvemos de manera efectiva el silencio generado en la extracción a través de la búsqueda de «términos madre» (Ahmad et al, 2001) y de morfemas específicos del dominio (Heid, 1998). La metodología empleada es extrapolable a otros campos científicos relacionados.

PALABRAS CLAVE: *extracción terminológica; sufijos de términos especializados; criterios de identificación; bioquímica.*

Abstract

Methodology for term extraction and identification in the domain of Biochemistry

In this article, we describe the candidate term extraction process from German texts in the field of Biochemistry using a textual analysis tool. Moreover, we effectively resolve silence generated in the extraction by searching for "mother terms" (Ahmad et al., 2001) and specific domain morphemes (Heid, 1998). The methodology used is transposable to other related scientific fields.

KEYWORDS: *terminological extractions; suffixes of specialized terms; identification criteria; biochemistry.*

TERMINÀLIA 16 (2017): 18-28 · DOI: 10.2436/20.2503.01.108
Data de recepció: 27/2/2017. Data d'acceptació: 27/3/2017
ISSN: 2013-6692 (impresa); 2013-6706 (electrònica) · *http://terminalia.iec.cat*

1 Introducción

El desarrollo de programas informáticos para la extracción automática de términos agiliza de forma espectacular la tarea de vaciado de términos. Hoy en día, los sistemas de extracción automática de candidatos a términos se pueden dividir siguiendo a Lossio-Ventura *et al.* (2014) en cuatro grandes categorías: las lingüísticas (Krauthammer y Nenadic, 2004), las estadísticas (Van Eck *et al.*, 2010), las híbridas y las que se basan en un aprendizaje automático (Newman *et al.*, 2012). Estas últimas diseñan y desarrollan algoritmos para que la herramienta mejore el comportamiento de forma automática a través de la experiencia y sea capaz de aprender por sí sola por medio de datos empíricos. Concretamente, un algoritmo es capaz de extraer características y patrones comunes de un conjunto de datos y aplicarlos a nuevos conjuntos de datos. La metodología estadística se basa, principalmente, en recontar unidades léxicas, realizar estadísticas de frecuencias y comparar frecuencias entre corpus especializados y de referencia. La metodología lingüística no se basa en contabilizar términos, sino que busca patrones sintácticos y morfológicos etiquetando el corpus para su posterior estudio lingüístico como el lexicográfico, por ejemplo. Ambos sistemas tienen sus limitaciones (Cabré *et al.*, 2001), de ahí que la tendencia actual se decante por un sistema híbrido que los integra a los dos. De este modo, se consigue ajustar la extracción a objetos de estudio particulares (Benavent y Parrilla, 2006).

No obstante, la gran mayoría de estos extractores automáticos están confeccionados para el estudio de la lengua inglesa, por lo tanto, no resultan tan efectivos en lenguas con características morfológicas diferentes como es el caso de la lengua alemana (Ahmad y Rogers, 2001; Heid, 1998), objeto de este estudio. En este trabajo, proponemos una metodología para extraer e identificar términos del campo de la bioquímica para posteriormente poder describir el comportamiento de los mismos desde el punto de vista lingüístico. Para ello, hemos estructurado este artículo en dos grandes partes. En la primera, explicamos la elaboración del corpus textual, establecemos los criterios seguidos para la detección de las unidades léxicas especializadas y justificamos la selección de la herramienta WordSmith Tools para llevar a cabo la extracción de los candidatos a términos. En la segunda parte, describimos el proceso de extracción poniendo de manifiesto las limitaciones surgidas y proponemos soluciones a las mismas.

2 Preliminares

2.1 Diseño y recopilación del corpus

Antes de pasar a la explicación de la metodología utilizada en la extracción de candidatos a términos, presentaremos el diseño y la recopilación de nuestro corpus textual. Nuestro objeto de estudio se centra en la lengua de especialidad del ámbito de la bioquímica, con un nivel de especialización alto y en alemán. Para el diseño del corpus nos basamos, naturalmente, en una serie de criterios. Al igual que Sinclair (1996), distinguimos dos tipos de criterios lingüísticos: externos e internos. Los criterios externos contemplan básicamente el contexto sociocultural y la función comunicativa del texto, mientras que los criterios internos hacen referencia a aspectos puramente lingüísticos como: la distribución de palabras, aspectos gramaticales, léxico, etc. La tendencia actual es usar ambos criterios y tal es nuestro caso. Otros criterios externos específicos contemplados y que hemos considerado son: la *cantidad* (criterio muy debatido y todavía sin esclarecer completamente), la *calidad* (los textos han de ser recientes y con una autoridad explícita y contenido fiable), la *lengua* (corpus monolingüe o multilingüe), el *nivel de la lengua* (dependerá de los interlocutores), la *temática* y la *forma* (textos escritos-orales, coloquial, culto, lenguaje especializado).

La fuente de nuestros textos es la revista *Angewandte Chemie*, que pertenece a la Sociedad Alemana de Química (Gesellschaft Deutscher Chemiker, GDCh) y es publicada por la editorial Wiley-VCH. Es una revista científica de gran reconocimiento, revisada por pares, que publica investigaciones originales que abarcan todo el campo de la química. Se edita semanalmente y el número de revistas anuales asciende a un total de 52. Se publican dos ediciones, una edición alemana, *Angewandte Chemie*, cuyo primer número data de 1887, y la internacional, totalmente en inglés, la *Angewandte Chemie International Edition*, que se estrenó 125 años después. Las ediciones son idénticas en cuanto al contenido y ambas se publican en papel y en línea. No obstante, es relevante resaltar que, en la edición alemana, predominan los artículos en lengua inglesa y además también contiene traducciones al alemán, que se han descartado para esta investigación por no tratarse de versiones originales. Recopilamos un total de 528 textos íntegros, de longitud variable, de diferentes especialistas y publicados en el periodo de tiempo de 2010 a 2014 (ambos incluidos).

Como hemos mencionado anteriormente, el diseño de nuestro corpus textual contempla además criterios internos, concretamente, el criterio del léxico. Este nos permitió delimitar nuestro campo de estudio. En efecto, la gran extensión del ámbito de la bioquímica nos llevó a acotar el árbol de campo (véase la figura 1) y limitarlo a la bioquímica humana, excluyendo los apartados señalizados en gris claro.

1	**Estructuras moleculares del ser vivo**
1.1	Biomoléculas
1.1.1	Inorgánicas
1.1.1.1	Agua
1.1.1.2	Sales minerales
1.1.2	Orgánicas
1.1.2.1	Glúcidos (hidratos de carbono)
1.1.2.2	Lípidos
1.1.2.3	Prótidos (compuestos nitrogenados)
1.1.2.4	Ácidos nucleicos
1.2	La célula
1.2.1	Animal
1.2.2	Bacteriana
1.2.3	Vegetal
2	**Reacciones metabólicas**
2.1	Enzimas
2.1.1	Coenzimas
2.2	Metabolismos
2.2.1	Metabolismos de los glúcidos
2.2.2	Metabolismo de los lípidos
2.2.3	Metabolismo de los prótidos
2.2.4	Metabolismo de los ácidos nucleicos
2.2.5	Fotosíntesis
3	**Métodos y técnicas instrumentales**
3.1	Cromatografía
3.2	Electroforesis
3.3	Técnicas de diálisis y ultracentrifugación
3.4	Espectroscopia
3.5	Isótopos radioactivos
3.6	Autorradiografía
3.7	Espectrometría de masas
3.8	Microscopia electrónica
3.9	Radioinmunoanálisis
3.10	Cristalografía de rayos X
3.11	Fluorometría
3.12	Inmunoprecipitación
4	**Aplicaciones**
4.1	Medicina y terapias químicas
4.2	Inmunología
4.3	Ingeniería genética y clonación
4.4	Nutrición
4.5	Química clínica
4.6	Farmacología
4.7	Toxicología
4.8	Nanotecnología
4.9	Ecología
4.10	Agricultura

FIGURA 1. Estructura conceptual de la bioquímica

Para seleccionar los textos pertenecientes a esta área, adaptamos el criterio de identificación de las unidades léxicas especializadas (ULE) de L'Homme (2004). Como veremos más adelante, esta autora identifica las ULE de un ámbito basándose en los términos que lo rodean. Y considera, entre otros criterios, que, si estos son del campo estudiado, la ULE también lo es. Para llevar a cabo la selección de los textos del corpus, aplicando este criterio, analizamos las ULE del título, del resumen y de las palabras clave. Así, si una ULE de estos tres apartados se combinaba con otras pertenecientes al dominio de la bioquímica humana, el texto entraba a formar parte del corpus. De esta forma, constituimos el corpus textual que usamos para la extracción de candidatos a términos.

2.2 Identificación de los términos

Antes de pasar a describir los criterios de selección, conviene aclarar la denominación de *término* que aplicaremos en este artículo. Ante la amplia variedad de denominaciones existentes para designar al término con todas sus diferentes matizaciones diferenciadoras, hemos decidido utilizar la denominación alemana de *término* porque se ajusta más a nuestro objeto de estudio, a saber, el estudio de términos alemanes especializados. El Instituto Alemán de Normalización (DIN) define el término simple o *Einwortbenennung* de la siguiente manera:

> Einwortbenennung: Eine aus einem Wort bestehende Benennung. ANMERKUNG: Zu den Einwortbenennungen zählen auch die zusammengesetzten einschließlich der mit Bindestrich durchgekoppelten Benennungen [...]. (DIN 2330, 1993, 2.4)

Es decir, un término simple es un término que consta de una palabra (*Molekül*) que puede estar compuesta por otras, bien unidas por guion (*Spacer-Molekül*) o sin él (*Molekülthermophorese*).

Y el término compuesto o agrupación de palabras o *Mehrwortbennenung oder Wortgruppe* lo define como un término que consta de al menos dos palabras separadas entre sí por espacios:

> Mehrwortbenennung: Eine Benennung, die aus mindestens zwei durch Leerstellen getrennten Wörtern besteht. (DIN 2330, 1993, 2.5)

El corpus analizado en este trabajo, como ya se ha mencionado anteriormente, se compone de textos especializados en lengua alemana, por lo que consideramos apropiado tener en cuenta las características morfosintácticas típicas de esta lengua especializada.

La composición es el recurso prototípico y, por tanto, más productivo y frecuente en la formación de términos especializados, seguido de la derivación, que también es muy empleada. Otros recursos característicos y abundantes en el dominio de la química son las

abreviaciones y las siglas, la conversión y los préstamos léxicos o extranjerismos.

Teniendo en cuenta las definiciones de *término simple* y *compuesto* que acabamos de indicar, nos centramos en la elección de términos simples considerando términos de una única palabra o *Simplizia*, composiciones, derivaciones, conversiones, abreviaciones, siglas y préstamos. Los adverbios no se contemplaron en esta ocasión y fueron descartados. No obstante, sí que se recogieron los verbos con valor terminológico. En la detección de términos compuestos nos fijamos en las construcciones nominales propuestas por Fluck (1997): (Adj. + N - Particip. + N - N + Prp + N - Nombres propios flexionados + N - N + NGen).

Antes de efectuar el vaciado terminológico, establecimos y aplicamos de forma sistemática unos criterios claros y sencillos para asegurar la calidad de la selección de candidatos (Estopà, 2001) que desarrollamos en los siguientes apartados.

2.2.1 Criterio de la frecuencia

El criterio de selección basado en la frecuencia de aparición es, naturalmente, clave. Es el primer acercamiento en el proceso de extracción para obtener una lista de candidatos a términos y estudiar con qué frecuencia aparecen. Si un término con valor terminológico muestra una elevada frecuencia, significa que es un término base del dominio o término madre o *mother term*, según Ahmad y Rogers (2001, p. 742). Estos términos fueron de gran ayuda en la búsqueda de otros términos, compuestos o simples —formados por composición o derivación—, no detectados por la herramienta informática. A pesar de la innegable importancia de la frecuencia, no contemplamos este criterio como de exclusión.

A través de la estrategia empleada para ampliar la lista de candidatos, se reunieron términos de frecuencia más baja de tres, que era nuestro umbral. Compartimos con Edo (2012) que el hecho de que un término aparezca en un corpus con una frecuencia igual a uno no significa que ese término no sea significativo ni representativo del dominio. Puede tratarse del fenómeno conocido como *hápax legómenon*, es decir, un término que aparece una única vez en un corpus. Esta autora recomienda, por tanto, no descartar *a priori* estos hápax sino más bien considerarlos para un posible análisis terminológico. Hemos decidido considerar e incluir en nuestra lista preliminar de candidatos a términos los hápax. En la fase de validación, junto a especialistas del campo, se descartarán o en su caso se conservarán estas unidades léxicas para su posterior estudio lingüístico. El papel de los especialistas será de gran importancia y determinante porque otorgará validez y fiabilidad al trabajo y, consecuentemente, al producto final que se pueda derivar del mismo.

2.2.2 Criterios semánticos

Mathews *et al.* (2006) definen la bioquímica como la ciencia que estudia los seres vivos a nivel molecular mediante técnicas y métodos físicos, químicos y biológicos. Es, por tanto, una ciencia interdisciplinar, experimental y de investigación que interactúa con otras disciplinas como son la química orgánica, biofísica, medicina, nutrición, microbiología, fisiología celular, genética, etc. Según estos autores, es, además, una disciplina diferenciada con identidad propia que se distingue por su énfasis en las estructuras y reacciones de las biomoléculas, por la explicación de rutas metabólicas y su control y por el principio de que los procesos vitales pueden comprenderse mediante leyes de la química. Esta definición nos inspiró en la elaboración de nuestro árbol de campo (véase la figura 1). Así, con ayuda de textos del dominio y de los especialistas en el ámbito confeccionamos la estructura conceptual de la bioquímica. Es, por tanto, requisito imprescindible que la unidad léxica extraída pertenezca a este árbol de campo para ser considerada unidad léxica de la especialidad.

2.2.3 Criterios léxico-semánticos

Para fijar los criterios léxico-semánticos de selección de términos nos basamos, al igual que hicimos para la elección de los textos del corpus, en L'Homme (2004) (véase el apartado 2.1), que propone cuatro criterios léxico-semánticos en la selección de posibles candidatos. Un término ha de estar relacionado con el dominio especializado para poder ser candidato, por ejemplo, sería el caso de: *Enzym*, *Protein*, *Synthase*, *Ligand*, *Reaktion*, *spektroskopisch*, *katalytisch*, etc. No obstante, existen términos que en ocasiones no resultan tan fáciles de identificar como pertenecientes a un dominio. En ese caso, L'Homme (2005) propone analizar los actantes semánticos que acompañan al candidato en el fragmento de texto en el que aparece. Si los actantes resultan ser términos especializados del dominio, fácilmente también lo será el candidato. A modo de ejemplo, el término *Immunität* se seleccionaría como posible candidato del ámbito de la bioquímica porque, como podemos observar más abajo, está acompañado por los actantes semánticos *T-Helferzellen*, *Gedächtnis-B-Zellen*, *Antigene*, que sí que pertenecen al dominio de la bioquímica (*Auch sind T-Helferzellen an der Bildung von Gedächtnis-B-Zellen beteiligt, die langandauernde* ***Immunität*** *gegen Antigene gewähren*). Además, considera como términos especializados los candidatos que muestran un mismo paradigma morfológico que los términos seleccionados como tal previamente. Así, los términos *Immunadsorbent-Untersuchungen* (ELISA), *Immunantwort*, *immunisiert*, *immunstimmulierendem* de los ejemplos siguientes, se seleccionaron como candidatos a términos por compartir el mismo lexema.

- *Die Seren wurden durch Enzymgebundene* ***Immunadsorbent-Untersuchungen (ELISA)*** *auf Vakzin-induzierte Serumantikörper des IgG-Typs geprüft.*
- *eine T-Zell-vermittelte* ***Immunantwort*** *begünstigt*
- *Dabei waren die Titer in den Mäusen signifikant höher, die mit der Vier-Komponenten-Vakzine 8* ***immunisiert*** *wurden (Abbildung 1A).*
- *Das B-Zellepitop kann z.B. mit zwei T-Zellepitopen und einem* ***immunstimmulierendem*** *Lipopetid wie Pam3Cys kombiniert werden, das als internes Adjuvans wirkt.*

Como es sabido, el uso de léxico de la lengua general para designar conceptos especializados es frecuente. En estos casos, es necesario comprobar si viene acompañado o no de actantes específicos del ámbito para poder seleccionarlos o descartarlos como candidatos. Por ejemplo, el término *Zweig* es un término de la lengua general y significa «rama». Pero, además, como se puede apreciar más abajo, aparece acompañado por actantes específicos del dominio de la bioquímica (palabras subrayadas), lo que nos indica su uso especializado y, por tanto, lo seleccionamos como candidato por su significado terminológico:

a) *Die Proteinsequenzen bisher untersuchter PS-Domänen zeigen hohe Homologien zu denen von DH-Domänen, erstere bilden jedoch einen eigenen phylogenetischen* ***Zweig.***

b) *Dagegen ist ein untereinheitenspezifischer Bindemodus eher für die Abschwächung des immunologischen* ***Zweigs*** *der proteasomalen Signalkaskade geeignet.*

Una vez fijados los criterios de identificación de los candidatos a términos, pasamos a explicar la elección de la herramienta informática elegida para la extracción.

2.3 Programa de concordancias WordSmith Tools

Para la extracción de las ULE del presente trabajo hemos elegido un programa de concordancias y no un extractor automático de términos por diferentes motivos que desglosamos a continuación:

a) Los conceptos semánticamente complejos en lengua alemana se expresan con mayor frecuencia a través de la composición de términos simples en lugar de términos compuestos (Heid, 1998). No compensa, por lo tanto, extraer términos con métodos estadísticos, es decir, aplicando criterios de búsqueda con *n-grams* de palabras consecutivas porque nos interesan sobre todo los términos simples, lo que equivaldría a los *unigrams*, es decir, a todas las palabras del texto.

b) Según Ahmad *et al.* (2001), en los textos existen palabras especialmente productivas que se combinan con otras formando términos compuestos o simples o bien forman palabras derivadas, incluso en dominios diferentes. Estas se pueden detectar con un simple análisis de frecuencias, creando una lista de términos simples potenciales y, además, posibles términos madre o *mother terms*. Frecuentemente, estas unidades léxicas van unidas a otra formando candidatos a términos compuestos o bien pueden formar parte de una composición de candidatos simples, como ocurre con gran frecuencia en la lengua alemana. Estos autores indican que para detectar términos madre es suficiente una herramienta informática de análisis de texto y no es necesario, por tanto, un extractor automático si no es objeto de estudio un análisis estadístico más complejo.

c) Como indican estos mismos autores (*ibidem*) la polisemia y los diferentes tipos de variaciones utilizados en los textos tampoco son detectables por extractores automáticos ya que estos no son capaces de distinguir otros significados de una misma palabra, ni de reconocer construcciones sintácticas diferentes con el mismo significado, ni omisiones empleadas por la economía del lenguaje, como tampoco las variaciones formales (ortográficas, abreviaciones, el uso de números, separación de palabras, etc.).

d) Gran número de extractores automáticos son creados para proyectos particulares y ajustados a unos objetivos específicos. Este hecho dificulta, por un lado, el acceso a los mismos, por no estar disponibles a investigadores no pertenecientes al proyecto, y, por otro, su implementación por las características propias del extractor, creado para proyectos con unos objetivos muy específicos y una lengua en particular. Otros, de libre acceso, como el LexTerm (Oliver *et al.*, 2007) resultan difíciles de instalar y ejecutar.

En la actualidad hay gran variedad de programas de concordancias tanto comerciales (WordSmith Tools, ConcGram, Collocate, etc.) como de libre acceso (Antconc, Monoconc, KWIC Finder, TextSTAT, SCP, Simple Concordance Program, etc...). Hemos optado por el programa WordSmith Tools por las siguientes razones:

1) Como indica Edo (2011), se trata de una herramienta que, por la gran variedad de opciones y prestaciones que ofrece, es más indicada para terminólogos e investigadores profesionales de este ámbito que otros programas de libre acceso como AntConc y MonoConc Pro.

2) Es, además, una herramienta que, a pesar de contar ya con veintiún años, no para de actualizarse ampliando y mejorando sus prestaciones. Su última versión, la 7.0, es de 2017. No obstante, para este trabajo empleamos la versión 5.0 porque es de la que disponemos y cumple perfectamente para conseguir nuestro objetivo de búsqueda de morfemas específicos y términos madre.

3 Metodología de extracción y resultados previos

Como acabamos de mencionar, para la extracción de los candidatos a términos de nuestro corpus utiliza-

mos el programa de concordancias comercial WordSmith Tools, versión 5 (Scott, 2011). En primer lugar, se extrajo una lista de términos con la herramienta WordList con el propósito de analizar la frecuencia de aparición de los mismos. En una primera extracción y con un umbral de frecuencias igual a 5, se obtuvieron únicamente 15.698 *types* (palabras diferentes) o candidatos a términos de un corpus que consta de 1.186.484 *tokens*, extraídos de los 528 artículos. Tras la supresión de los *tokens* correspondientes a números resultó un corpus con un total de 1.111.885. Como suele ocurrir en este tipo de extracciones, se generó, por un lado, demasiado silencio, es decir, no se detectaron los términos de frecuencia baja que podrían ser términos potenciales. Y, por otro, también se generó gran cantidad de ruido, es decir, palabras funcionales sin valor terminológico que convenía eliminar. Así, en la lista resultante de la primera extracción, el primer candidato a término *Zelle* (célula) ocupaba la posición 69 en el *ranking* de frecuencias.

3.1 Tratamiento del ruido y del silencio generado en la extracción

Para solucionar el problema del ruido, se creó un archivo de texto con una lista de exclusión que contenía este tipo de palabras gramaticales sin relevancia terminológica (artículos, conjunciones, preposiciones, verbos auxiliares, léxico no especializado, etc.), llamada *stopword list* (Ahmad y Rogers, 2001, p. 741). Para la elaboración de la lista de exclusión se tomó como punto de partida la lista que proporciona la Universidad de Neuchâtel[1] y se amplió con otras palabras funcionales sin relevancia terminológica extraídas de nuestro corpus. Una vez cargada esta *stopword list* en la aplicación, se repitió el proceso de extracción y se logró excluir 582.742 *tokens* y 885 *types*.

Con la aplicación de esta *stopword list* se consiguió acercar los términos especializados a los primeros puestos de la lista de frecuencia. Por ejemplo, *Zelle* pasó de ocupar el puesto 69 a ocupar el primer lugar en la lista.

No obstante, continuaba existiendo ruido que generaba una visión general poco depurada de la lista de frecuencias. La lengua alemana es una lengua muy flexiva, lo que significa que un mismo adjetivo, por ejemplo, puede aparecer de forma repetida cada vez con una terminación diferente según su función sintáctica. Además, aparecen sustantivos en singular y plural, verbos conjugados, etc., lo que genera gran cantidad de formas diferentes de una misma palabra que distorsiona el recuento final de la frecuencia de los candidatos a términos. Para resolver este problema se decidió lematizar el corpus. Una posibilidad es realizarlo de forma manual y otra, elaborar una lista y cargarla en forma de un archivo de texto en la herramienta. Se creó una lista con 1.124 palabras y sus formas lematizadas. Lematizar un corpus de este tamaño, teniendo en cuenta las características morfosintácticas de la lengua alemana, por un lado, y la escasez de recursos humanos, por otro, resulta muy fatigoso y requiere mucho tiempo, por lo que se optó por continuar lematizando de forma manual y directamente sobre la lista ordenada alfabéticamente arrastrando las formas lematizadas a un único lema.

En la siguiente tabla presentamos la resolución de la cantidad de ruido generada por las duplicaciones de palabras. Se observan cambios en el *ranking* de frecuencias. Por ejemplo, *Protein* ha pasado del puesto 30 a ocupar el segundo, al reunir todas sus formas posibles de aparición en un único lema. De forma detallada vienen indicadas a la derecha del candidato a término las formas lematizadas y la cantidad de formas que reúne. Este proceso permite detectar y descartar términos no especializados de forma más fácil, al aparecer aglutinadas todas sus posibles formas en un único lema. En este segundo acercamiento, el verbo *zeigen* (mostrar) se ha situado en primer lugar. Aparentemente, este verbo no es un candidato a término ya que carece de contenido terminológico. Lo mismo sucede con los demás términos sombreados en la tabla 1.

Una vez resuelta en parte la generación de ruido, intentamos mitigar, en la medida de lo posible, el silencio generado. Para ello, se repitió el proceso de extracción modificando el umbral de frecuencia. En el segundo intento, se modificó la configuración de la herramienta WordSmith Tools con el fin de que detectase todos los términos especializados, es decir, situamos el nivel de frecuencia en 1. Los resultados obtenidos tampoco fueron demasiado satisfactorios a pesar de que aumentase el número de *types* considerablemente. El número de candidatos pasó de 15.698 a 93.054, es decir, a simple vista se logró reducir considerablemente el silencio. No obstante, tras analizar la estadística de frecuencias se observó que a partir del término 34.035 la frecuencia de aparición era igual a uno, es decir, aproximadamente un 63 % de términos aparecía una única vez. Ante estos resultados, se optó por realizar una tercera extracción ajustando el umbral de frecuencia a 3. En esta ocasión, se amplió la lista en 9.447 términos logrando así un total de 25.141 posibles candidatos a términos.

Con el propósito de minimizar todavía más el silencio existente, realizamos búsquedas en el programa de concordancias Concord. Para ello, seguimos la propuesta de Heid (1998), que sugiere llevar a cabo búsquedas de morfemas específicos del dominio en compuestos nominales y derivaciones con el fin de detectar términos especializados simples que detallaremos en el siguiente apartado.

1	ZEIGEN	4199	zeigen[1149] gezeigt[906] zeigt[1016] zeigte[578] zeigten[550]
2	PROTEIN	2595	protein[682] proteine[971] proteinen[652] proteins[290]
3	FÜHREN	2132	führen[471] führt[797] führte[540] führten[244] geführt[80]
4	VERBINDUNG	1600	verbindung[670] verbindungen[930]
5	STRUKTUR	1589	struktur[969] strukturen[620]
6	CHEMISCH	1553	chemisch[198] chemische[513] chemischem[6] chemischen[640] chemischer[178] chemisches[18]
7	ZELLE	1518	zelle[254] zellen[1264]
8	VERWENDEN	1392	verwenden[126] verwendet[968] verwendeten[298]
9	METHODE	1387	methode[731] methoden[656]
10	PEPTID	1345	peptid[385] peptide[541] peptiden[253] peptids[166]
11	ENZYM	1337	enzym[375] enzyme[508] enzymen[249] enzyms[205]
12	BEOBACHTEN	1321	beobachten[174] beobachtet[905] beobachteten[242]
13	STARK	1309	stark[567] starke[282] starkem[8] starken[136] starker[39] starkes[20] stärker[109] stärkere[57] stärkeren[29] stärkerer[9] stärkeres[4] stärksten[49]
14	REAKTION	1262	reaktion[815] reaktionen[447]
15	SYNTHESE	1223	synthese[1139] synthesen[84]
16	BINDEN	1213	binden[427] band[17] banden[161] bindet[304] gebunden[304]
17	KOMPLEX	1196	komplex[389] komplexe[273] komplexen[262] komplexer[82] komplexes[190]
18	ERMÖGLICHEN	1169	ermöglichen[442] ermöglicht[673] ermöglichten[54]
19	BILDUNG	1163	
20	AKTIV	1157	aktiv[112] aktive[281] aktivem[10] aktiven[604] aktiver[90] aktives[36] aktivste[9] aktivstem[1] aktivsten[14]
21	AKTIVITÄT	1156	aktivität[1036] aktivitäten[120]
22	MOLEKÜL	1151	molekül[182] moleküle[564] molekülen[298] moleküls[107]

TABLA 1. Listado de frecuencia tras la lematización del corpus

3.2 Detección de términos especializados

3.2.1 Detección de adjetivos especializados por su composición

Revisamos la literatura existente en relación con estudios lingüísticos precedentes en lengua alemana y en el dominio de la química con el objetivo de recabar morfemas específicos. Nos acogimos a la lista de morfemas característicos y productivos del dominio propuestos por Banionyté (2008) en su trabajo. A continuación, enumeramos los sufijos alemanes más frecuentes acompañados de ejemplos extraídos de nuestro corpus:

- *-isch: chemisch, proteolytisch, termisch, NMR- spektroskopisch, genetisch, katalytisch, photometrisch, biochemisch, metabolisch, zytotoxisch*, etc.
- *-bar: nichthydrolysierbar, synthetisierbar, codierbar, photoschaltbar, oxidierbar, hyperpolarysierbar, kuppelbar, photo-depolarisierbar, bioverfügbar, bioaktivierbar*, etc.

Los sufijos *-lich* (*entzündlich, nichtnatürlich, löslich*), *-ig* (*hochgradig, zweistellig, zellgängig, kurzlebig, zellgängig, mittig, großskalig, großflächig*) y *-los* (*präzedenzlos*) apenas aparecen en el corpus y los sufijos *-haft*, *-weise*, *-mäßig-* prácticamente no aparecen y, si se da el caso, sin valor terminológico. Los sufijos de origen griego y latino sí muestran una elevada frecuencia en comparación con los de origen alemán. Son frecuentes los adjetivos acabados en:

- *-al: lysosomal, isothermal, ribosomal, bioorthogonal, konfokal, intraperitoneal, chromosomal, unidirektional, spektral*, etc.
- *-iv: selektiv, hochreaktiv, hochpromiskuitiv, putativ, präparativ, seronegativ, redoxaktiv, konformativ, regioselektiv, radioaktiv*, etc.
- *-ent: kovalent, latent, monovalent, nukleaseresistent, transient, biopersistent*, etc.
- *-ar: intramolekular, homonuklear, linear, planar, laminar*, etc.

Sufijoides productivos en química serían, sobre todo:

- *-reich: elektronenreich, Prolin-reich, G-reich, AEG-reich, FG-reich, Aspartat-reich, glykotoxinreich, glucosereich, lipidreich*, etc.

- *-förmig*: *rhombusförmig*, *haarnadelförmig*, *Donut-förmig*, *X-förmig*, *T-förmig*, *kreuzförmig*, *trichterförmig*, *faserförmig*, *α-Helix-förmig*, *Gaußförmig*, etc.
- *-haltig*: *olefinhaltig*, *schwefelhaltig*, *Fe3-haltig*, *DH-haltig*, *PS-haltig*, *Pyran-haltig*, *Adenosin-haltig*, *Glutamirid-haltig*, *Häm-haltig*, *serumhaltig*, etc.
- *-ähnlich*: *virusähnlich*, *wirkstoffähnlich*, *GliT-ähnlich*, *Rh-ähnlich*, *Ps-ähnlich*, *Partikel-ähnlich*, *inhibitorähnlich*, *Nukleinsäure-ähnlich*, *Antagomir-ähnlich*, *lipidähnlich*, etc.
- *-geschützt*: *Propargylgeschützt*, *Cbz-geschützt*, *Boc-geschützt*, *MOM-geschützt*, *basengeschützt*, *Fmoc-geschützt*, etc.
- *-artig*: *Lectin-artig*, *Cluster-artig*, *chymotripsinartig*, *faserartig*, *Claisen-artig*, *haarnadelartig*, *reißverschlussartig*, *Michaelis-Mentenartig*, *inhibitorenartig*, *substratartig*, etc.
- *-frei*: *nukleotidfrei*, *kupferfrei*, *CO2-frei*, *metallfrei*, *enzymfrei*, *zellfrei*, *waschfrei*, *Löscher-frei*, *donorfrei*, *DNA-frei*, etc.
- *-los*: *nahtlos*, *präzedenzlos*, *geruchslos*, *kontaktlos*, *funktionslos*, *strahlungslos*, *konturlos*, *rückstandslos*, *strukturlos*, *wasserlos*, etc.
- El resultado de la búsqueda del sufijo *-ig* (**ig**) indica que este sufijo es más productivo en adjetivos no especializados que en los de contenido terminológico. En el análisis de los candidatos se observó que si se repetían con frecuencia los adjetivos acabados en *-abhängig* y *-fähig* en combinación con otros adjetivos o sustantivos. Los hemos considerado sufijoides por ser muy productivos y porque semánticamente conserva más valor semántico la base (sustantivo o adjetivo) que el sufijoide (Fleischer, 1969). Ampliamos así la lista de partida propuesta por Banionyté (2008). Algunos ejemplos son:
- *-abhängig*: *Ligase-unabhängig*, *Mevalonat-unabhängig*, *FAD-abhängig*, *Häm-abhängig*, *substratnabhängig*, etc.
- *-fähig*: *wirkungsfähig*, *reaktionsfähig*, *paarungsfähig*, *expansionsfähig*, *wachstumsfähig*, etc.

Por otro lado, los sufijoides *-arm* (*elektronenarm*, *asparaginarm*, *ermüdungsarm*), *-fest*, *-dicht* (*gasdicht*, *hochdicht*), *-verträglich* y *-leer* mostraron escasa o ninguna presencia en nuestro corpus.

3.2.2 Detección de sustantivos especializados por su composición

3.2.2.1 Sustantivos simples

Se siguió la misma metodología en la búsqueda de morfemas especializados en sustantivos. Para ello, se escogió la lista propuesta por Lippert (1978), que señala en su trabajo una serie de sufijos típicos en el ámbito de la bioquímica y de la medicina,[2] además de otros prefijos representativos del campo. Por ejemplo, hicimos búsquedas con los sufijos[3] siguientes:

- *-an* (*Triisopropylsilan*, *Pyran*, *Glycan*, *Phodphoglycan*, *Tryptophan*, *Cyclohexan*, *Dimethyldioxiran*, etc.)
- *-ol* (*Diaryltetrazol*, *Ethanol*, *Thiol*, *Alkohol*, *Phenol*, *Methanol*, *Triol*, etc.)
- *-on* (*Analogon*, *Ketoclomazon*, *Keton*, etc.)
- *-ose* (*Toxoplasmose*, *Meiose*, *Endozytose*, *Diagnose*, *Apoptose*, *Agarose*, *Glukose*, *Aminoaldose*, *Ribose*, etc.)
- *-ase* (*Rab-GTPase*, *Carboxypeptidase*, *Proteinase*, *Leberesterase*, *Oxidoreduktase*, *ß-Sekretase*, etc.)
- *-itis* (*Thyroiditis*, *hepatitis*, *Arthritis*, *Kolitis*, etc.)
- *-om* (*Mantelzellymphom*, *Proteasom*, *Genom*, *Myelom*, *Thermosom*, *Tn-Syndrom*, *Metabolom*, *Epigelon*, *Mammakarziom*, etc.)
- *-pathie* (*Neuropathie*, *Myopathie*, *Nefropathie*, *Enzefalopathie*, etc.)

Este sistema es útil para detectar y estudiar el término en su contexto, pero también genera ruido puesto que muestra todas las palabras que incluyen el morfema de búsqueda. Se podría eliminar el asterisco a la derecha del morfema, pero en ese caso no se detectarían otros términos relevantes. Por ejemplo, en la búsqueda con el morfema *-*ase**, se obtuvo un total 6.271 *tokens*, que incluían repeticiones de un mismo término y además otros que no hacían referencia a enzimas, como: *Base*, *Phase*, *Laser* y sus formas compuestas: *Schiff-Base*, *Festphase*, *Wachstumphase*, *Laserlicht*, *Laserbeschuss*, etc. Se identificaron incluso términos con un morfema específico *-stase* no considerado en un principio (*Homoöstase*, *Epistase*, *Metastase*, *Hämostase*). En la misma búsqueda detectamos además un préstamo léxico (extranjerismo) *chase*. A título ilustrativo, presentamos la tabla 2 con las primeras 22 entradas.[4]

Se renunció a buscar por el sufijo *-en*, propuesto por Lippert (1978), a pesar de tratarse de un morfema específico para designar *hidrocarburos insaturados* porque no incluimos en nuestro corpus artículos referidos a hidrocarburos saturados (alcanos) ni a insaturados (alquenos y alquinos).

Una vez detectados términos simples mediante esta técnica, se emplearon estos mismos en realizar sucesivas búsquedas con el propósito de ampliar los candidatos a términos y así minimizar en lo posible el silencio generado. A continuación, mostraremos algunos ejemplos de la metodología empleada. Con el término *Laser*, recuperado en la búsqueda con el morfema específico *-ase*, se realizaron dos tipos de búsqueda:

a) Añadiendo asteriscos a ambos lados del término (**laser**). De esta manera aparece como palabra base (o *Grundwort*)[5] en el caso de una composición nominal. De la búsqueda se obtuvieron entre otros: *Quantenkaskadenlaser*, *Abreicherungslaser*, *Freie-Elektronen-Laser*, *Punktlaser*, *Diodenlaser*, etc.

b) Añadiendo asterisco únicamente a la derecha (*laser**), el término *Laser* funciona como un determinante o complemento de la palabra base (o *Bestimmungswort*) en caso de una composición nominal. De la búsqueda se extrajeron los siguientes términos: *Laserbeschuss* (al otro grupo), *Laserfokalvolumen*, *Laserleistungsdichte*, *Laserfarbstoff*, *Laseranregungsleistung*, *Laserlichtquelle*.

1	Die RCM wurde an der festen **Phase** unter Verwendung des Grubbs-Katalysators
2	StRIP3 der erste Binder einer **Rab-GTPase** in ihrer aktiviert
3	opsin (Rh) ist die protonierte **Schiff-Base** des 11-cis-Retinals (PSB11)
4	Proteasestabilität wurde mit **Carboxypeptidase** Y, Chymotrypsin
5	ypeptidase Y, Chymotrypsin und **Proteinase K** getestet. In fast
6	Die einzige Ausnahme war, dass **Carboxypeptidase** Y langsam die letzten
7	orthogonale UAS ersetzt wird **(Chase).** Solch eine Methode
8	Doppelmutante (Y306A, Y384F) der **Pyrrolysyl-tRNA-Synthetase** (PylRSAF)
9	Pullse), bevor anschließend ein **4 h-Chase** mit der zweiten UAS(BCN)
10	durch Zugabe von anorganischer **Diphosphatase** erhöht, indem
11	Lipoxygenasen, nichtspezifische **Leberesterase** und Myosin-ATPa
12	Flaninadenindinukleotid(FAD)-abhängige **Oxidoreduktase**, 4 zum
13	Cysteinreste wurde bereits für die **Dihydrolipoamid-Dehydrogenase**
14	Substrat kovalent mit Cys145 der **Oxidoreduktase** verknüpft, und
15	Voraussetzung für die zelluläre **Homöostase**. Der komplexe Aufbau
16	Proteine, L7Ae, Nop5 und die **Methyltransferase** Fibrillarin
17	dazu beitragen, die **Methyltransferase** relativ zur Substrat RNA
18	Struktur-Aktivitäts-Beziehungen von **ß-Sekretase(BACE1)-Hemmern**
19	um Wirkstoffe zur Hemmung von **ß-Sekretase** herzustellen. Dieses Enzym
20	HGVTSAPDTRPAPGSTAPPA, an der **Festphase** wie beschrieben synthetisiert
21	die Mortalitätsrate durch den **Lactatdehydrogenase(LDH)-Test** gemessen
22	in seiner exponentiellen **Wachstumsphase** bei einer optischen Dichte
	...
31	erfolgt 30 Minuten lang mit **Laserlicht** von 568 nm bei 77
32	vermessen, um bei jedem **Laserbeschuss** die vollständig Ersetzung

Tabla 2. Resultados de búsqueda con el morfema *-*ase**

También se optó por identificar nuevos candidatos mediante búsquedas con términos simples básicos de elevada frecuencia (términos madre) de nuestra lista de frecuencias. A modo de ejemplo, con el término simple *Molekül*, se procedió de la misma manera que en la búsqueda anterior añadiendo asteriscos y se obtuvieron:

a) con (**molekül*) entre otros: *Spacer-Molekül*, *FAD-Molekül*, *Disciformycinmolekül*, *Transmitternmolekül*, *Strukturwassermolekül*, *Vorläufermolekül*, etc.
b) Y con (*molekül**): *Molekülpopulation*, *Molekülfalle*, *Molekül-Wasser-Cluster*, *Molekül-Wasser-Wechselwirkung*, *Moleküldynamik-Simulation*, *«Shuttle»-Moleküle*, etc.

3.2.2.2 Sustantivos compuestos

En la detección de términos especializados compuestos o agrupaciones de palabras nos basamos de nuevo en Heid (1998). Él identificó en su investigación básicamente tres tipos de construcción nominal que aplicamos a nuestro corpus:

a) dos sustantivos, el segundo en genitivo (N + N Gen)
Diagnose akuter Infektionen
Diagnose verschiedener Stadien
Diagnose epigenetisch bedingter Krankheiten
Homöostase des Wirts
b) dos sustantivos unidos por una preposición (N + Prp + N)
Laserfleck mit Licht
Laserphotolyse mit sichtbarem Licht
Base zur Thioldeprotonierung
Basenstapelung für die DNA-Schäden
c) adjetivo y sustantivo (Adj + N)
ribosomale Bindetasche
ribosomale Inhibitoren
ribosomale Peptidbiosynthese
(nicht)ribosomale Peptide

Fluck (1997) señala, además de las indicadas por Heid, otras dos posibles combinaciones en las lenguas especializadas en el ámbito de las ciencias y la tecnología, a saber, la agrupación de un participio y un sustantivo (Particip. + N) y la de un nombre propio

flexionado y un sustantivo (Nombres propios flexionados + N). Las numerosas búsquedas de morfemas específicos realizadas en el programa de concordancias nos permitieron constatar que estos dos patrones eran muy recurrentes en el corpus, lo que nos indujo a realizar búsquedas basándonos en ellos. En el caso del patrón participio + sustantivo, buscamos participios de verbos acabados en *-ieren* porque gran número de estos verbos tiene como base léxica un término especializado, como por ejemplo: *Acetyl* (sustantivo) *acetylieren* (verbo) *acetyliert-*[6] (participio pasado - adjetivo). A continuación, presentamos ejemplos extraídos de nuestro corpus:

a) (Particip. + N)
säulenbasierter Affinitätstest
zellbasierte Aktivität
Chip-basierter Enzymhemmtest
tumorhemmende Aktivität

b) (Nombres propios flexionados + N)
Van der Waals-Abstandes
Michelis-Menten-Kinetik

3.2.3 Observaciones sobre las variaciones terminológicas

La búsqueda con los términos madre y morfemas base del dominio nos permitió además detectar variaciones denominativas que cabe mencionar. Basándonos en la tipología propuesta por Freixa (2002), distinguimos las siguientes y ofrecemos un ejemplo de cada una:

— variación gráfica: *Grün fluoreszierendes Protein / GFP*.
— variación morfosintáctica: *hydrophobe Seitenkettenverbrückung / hydrophob verbrückter Seitenketten*.
— variación morfológica: *das Modell der Wechselwirkung / das Wechselwirkungsmodell*.
— variación léxica: *schwingungsspektroskopische Technik / schwingungsspektroskopischer Ansatz*.
— variación por reducción: *eine Diels-Alder-Reaktion / eine Alder-Reaktion*.

Un caso de variación que nos llamó la atención fue un término que recoge casi todas las variaciones mencionadas: *Förster-Resonanzenergietransfer / FRET / Förster-Resonanz-Energietransfer* (v. gráficas); *resonantem Föster-Energietransfer / resonantem Energietransfer nach Föster* (v. morfosintáctica); *Förster-Resonanzenergietransfer / Fluoreszenz-Resonanzenergietransfer* (v. léxica); Förster-*Resonanzenergietransfer / Förster-Energietransfer* (v. por reducción).

4 Conclusión

Para fijar los criterios de identificación con el propósito de realizar un trabajo sistemático y, por tanto, más fiable, aplicamos los criterios léxico-semánticos de L'Homme (2004) tanto en la selección del corpus textual como en la de los candidatos a términos. Para la extracción de las ULE empleamos el programa de concordancia WordSmith Tools. Extrajimos una lista de candidatos a términos, que como era de esperar en un principio contenía demasiado ruido y mucho silencio. Eliminar el ruido no resultó complicado y para mitigar el silencio modificamos el umbral de frecuencia en la herramienta informática y realizamos búsquedas con morfemas típicos del campo de la bioquímica basándonos en la estrategia de Heid. Además, ampliamos la búsqueda de candidatos a partir de los términos simples (términos madre) de frecuencia elevada obtenidos en el listado de frecuencias. Este método de trabajo nos ha permitido identificar, por un lado, dos sufijoides productivos en el campo de la bioquímica no incluidos ni en Banionyté ni en Lippert (*-abhängig* y *-fähig*) y, por otro, variaciones denominativas gráficas, morfosintácticas, morfológicas, léxicas y por reducción. Por consiguiente, en este trabajo hemos reunido y adaptado los sufijos extraídos en trabajos realizados en química y medicina al campo de la bioquímica abriendo a investigadores de otras disciplinas científicas posibilidades de adaptación de esta metodología a sus ámbitos de estudio.

Bibliografía

AHMAD, Khurshid; ROGERS, Margaret (2001). «Corpus linguistics and terminology extraction». En: WRIGHT, Sue Ellen; BUDIN, Gerhard. *Handbook of terminology management*. Vol. 2: *Application-oriented terminology management*. Amsterdam: John Benjamins, p. 725-760.

BANIONYTÉ, Vita (2008). «Zur Terminologie und zum Wortschatz der deutschen Fachsprache der Chemie». *Santalka: Filologija, Edukologija*, 16(4), p. 4-11.

BENAVENT, Paloma; PARRILLA, Sara (2006). *Análisis de la extracción automática de términos con el programa informático ExtraTerm* [en línea]. <http://repositori.uji.es/xmlui/bitstream/handle/10234/78647/forum_2006_25.pdf?sequence=1> [Consulta: 30 marzo 2016].

DIN 2342 Teil 1 (1992). *Begriffe der Terminologielehre - Grundbegriffe*. Berlin: Beuth.

Cabré, M. Teresa; Estopà, Rosa; Vivaldi, Jordi (2001). «Automatic term detection: A review of current systems». En: Bourigault, D.; Jacquemin, C.; L'Homme, M-C. *Recent advances in computational terminology*, p. 53-88.
Edo Marzá, Nuria (2011). «Terminology management systems for the development of (specialised) dictionaries: a focus on WordSmith Tools and Termstar XV». *Language Value*, vol. 3, p. 162-173.
Edo Marzá, Nuria (2012). «Lexicografía especializada y lenguajes de especialidad: fundamentos teóricos y metodológicos para la elaboración de diccionarios especializados». *Lingüística*, vol. 27, n.º 1, p. 98-135.
Estopà, Rosa (2009). «El diseño de aplicaciones terminológicas: los extractores de terminología». *Boletín de los Traductores Españoles de las Instituciones de la Unión Europea* [en línea], n.º 115-S, p. 15-21. <http://ec.europa.eu/translation/spanish/magazine/documents/pyc_115_supl_es.pdf> [Consulta: 30 marzo 2016].
Estopà, Rosa (2001). «Extracción de terminología: elementos para la construcción de un extractor». *TradTerm*, vol. 7, p. 225-250.
Fleischer, Wolfgang (1969). *Wortbildung der deutschen Gegenwartssprache.* Leipzig: VEB.
Fluck, Hans Rüdiger (1997). «Fachdeutsch in Naturwissenschaft und Technik: Einführung in die Fachsprachen und die Didaktik». *Methodik des Fachorientierten Fremdsprachenunterrichts (Deutsch als Fremdsprache)*, vol. 2.
Freixa, Judit (2002). *La variació terminològica: Anàlisi de la variació denominativa en textos de diferent grau d'especialització de l'àrea de medi ambient.* Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. (Sèrie Tesis; 3)
Heid, Ulrich (1998). «A linguistic bootstrapping approach to the extraction of term candidates from German text». *Terminology*, vol. 5:2, p. 161-181.
Krauthammer, Michael; Nenadic, Goran (2004). «Term identification in the biomedical literature». *Journal of Biomedical Informatics*, vol. 37, n.º 6, p. 512-526.
Lippert, Herbert (1978). «Fachsprache Medizin». *Interdisziplinäres Deutsches Wörterbuch in der Diskussion. Düsseldorf*, p. 86-101.
L'Homme, Marie Claude (2004). *La terminologie: principes et techniques.* Pum.
L'Homme, Marie Claude (2005). «Sur la notion de «terme»». *Meta: Journal des Traducteurs Meta: / Translators' Journal*, vol. 50, n.º 4, p. 1112-1132.
López, Coral; Olmo, Françoise (2015). «Compiling texts for a specialized corpus in the biochemistry domain: Theoretical and methodological aspects». *Procedia-Social and Behavioral Sciences*, 198, p. 300-308.
Lossio-Ventura, Juan Antonio; Jonquet, Clement; Roche, Mathieu; Teisseire, Maguelonne (2014). «Biomedical terminology extraction: a new combination of statistical and web mining approaches». En: *Proceedings of Journées Internationales d'Analyse Statistique des Données Textuelles (JADT2014)*. París: France.
Mathews, Christopher; Van Holde, Kensal; Ahern, Kevin (2002). *Bioquímica.* 3.ª ed. Madrid: Pearson Education.
Newman, David; Koilada, Nagendra; Lau, Jey Han; Baldwin, Timothy (2012). «Bayesian text segmentation for index term identification and keyphrase extraction». *Proceedings of 24th International Conference on Computational Linguistics (COLING)*, p. 2077-2092.
Oliver, Antoni.; Vázquez, Mercè; Moré, Joaquim (2007). «Linguoc Lexterm: una eina d'extracció automàtica de terminologia gratuïta». *Translation Journal* [en línea], vol. 11, n.º 4. <http://translationjournal.net/journal/42linguoc.htm> [Consulta: 27 febrero 2017].
Scott, Mike (2011). *WordSmith Tools version 5*. Oxford: Oxford University Press.
Sinclair, John (1996). «Preliminary recommendations on corpus typology». *EAGLES Document TCWG-CTYP/P* [en línea]. <http://www.ilc.pi.cnr. it/EAGLES/corpustyp/corpustyp.html> [Consulta: 30 marzo 2016].
Van Eck, Nees Jan; Waltman, Ludo; Noyons, Ed; Buter, Reindert (2010). «Automatic term identification for bibliometric mapping». *Scientometrics*, vol. 82, n.º 3, p. 581-596.
Wright, Sue Ellen; Budin, Gerhard (ed.) (2001). *Handbook of terminology management*. Vol. 2: *Application-oriented terminology management*. Amsterdam. John Benjamins Publishing.

Notes

1. *http://members.unine.ch/jacques.savoy/clef/index.html*
2. Se incluyen estos sufijos porque nuestro árbol de campo comprende aplicaciones médicas.
3. En las búsquedas añadimos delante y detrás del sufijo el operador truncamiento *(asterisco) con el fin de abarcar desinencias y la raíz del adjetivo y/o sustantivo.
4. Hemos añadido dos filas más para mostrar un par de ejemplos con el acrónimo laser.
5. La palabra base o *Grundwort* es la palabra situada al final de la composición y es la que confiere a la nueva formación los rasgos morfológicos y además lleva el peso semántico del compuesto.
6. El guion detrás del participio ocupa el lugar de las posibles flexiones que puedan tomar del adjetivo en función predicativa.