COMBINACIÓN DE ESTRATEGIAS LÉXICAS Y ESTADÍSTICAS PARA EL RECONOCIMIENTO AUTOMÁTICO DE TÉRMINOS: SU APLICACIÓN A UN CORPUS DE MEDICINA[1]

ANTONIO MORENO SANDOVAL
Universidad Autónoma de Madrid
LEONARDO CAMPILLOS LLANOS
LIMSI-CNRS
Universidad Autónoma de Madrid

RESUMEN

El reconocimiento automático de términos a partir de textos técnicos tiene básicamente dos aplicaciones: la ayuda al lexicógrafo y documentalista, por un lado, y por otro, la identificación de conceptos clave en sistemas de traducción automática o recuperación de información. El reconocimiento de términos depende del dominio temático y de los patrones de formación léxica de la lengua en cuestión, en nuestro caso, el español. En este artículo, presentamos la evaluación de tres estrategias diferentes de selección de precandidatos a término, que en una segunda fase se combinan con un filtrado basado en afijos médicos para dar una propuesta de término. El artículo analiza los problemas de cobertura y precisión de cada estrategia y expone cómo se ha empleado estas técnicas para compilar un lexicón de términos médicos y un sistema de reconocimiento automático para el español.

PALABRAS CLAVE: términos, lingüística computacional, corpus de especialidad.

COMBINED STRATEGIES FOR AUTOMATIC TERM RECOGNITION AND ITS APPLICATION TO A SPANISH CORPUS OF MEDICINE

ABSTRACT

Automatic Term Recognition from technical texts has mainly two applications: on one side, assistance to lexicographers and documentalists; and on the other side, identification of key concepts for machine translation and information retrieval. Term recognition depends on the subject domain and the lexical patterns of a given language, in our case, Spanish. In this article, we present an evaluation of three different strategies for selecting candidates to term. In a second step, each

[1] Trabajo financiado por el MICINN (TIN2010 20644 C03 03) y la Comunidad de Madrid (programa MA2VICMR).

list is filtered by a set of medical affixes to provide a final proposal of terms. This paper also discusses the problems of recall and precisión of each strategy and shows how these techniques have been used to compile a lexicon of medica terms and an ATR system for Spanish.

KEYWORDS: terms, computational linguistics, domain corpus.

RECEPCIÓN: 24/III/2014 ACEPTACIÓN: 3/IV/2014

1. INTRODUCCIÓN

La terminología es una rama de la lingüística aplicada que tiene por objeto, entre otros, la elaboración de diccionarios de lenguaje especializado. Los dominios temáticos componen una sublengua específica y precisa, adaptada a la designación de los conceptos de cada materia o área de conocimiento. En esta sublengua coexisten términos exclusivos y otros que adquieren significaciones propias diferentes de las de la lengua general. La elaboración de un diccionario terminológico es una tarea multidisciplinar especializada, pues necesita tanto de lexicógrafos como de especialistas en la disciplina, para decidir lo que es un término y cómo definirlo de la manera más rigurosa. Los terminólogos, además de contar con la ayuda de expertos, trabajan con diferentes fuentes: desde la consulta a obras de referencia hasta tesauros y bases de datos especializadas (p. ej., IATE o TERMIUM). Sin embargo, algunos campos, que se caracterizan por una rápida evolución de la disciplina y la consiguiente incorporación vertiginosa de nuevos conceptos, requieren un constante trabajo de detección y normalización. La terminología médica es uno de esos campos donde el número de términos excede, considerando lemas simples y formas compuestas, el número habitual de vocablos especializados en otras disciplinas. Por dar una cifra, el *Diccionario de términos médicos*[2] contiene casi 52 000 términos.

La dinámica de generación de nuevos términos justifica la necesidad de herramientas informáticas como los reconocedores automáticos. Estas aplicaciones analizan textos digitalizados e identifican candidatos que pueden ser términos en un dominio dado, para su posterior validación por un experto.

El objetivo del artículo es exponer cómo se recopiló una lista de términos a partir de un corpus de textos médicos (MultiMedica). Dicho listado se emplea en un extractor terminológico para el español creado en el

[2] REAL ACADEMIA NACIONAL DE MEDICINA, *Diccionario de términos médicos*, Editorial Médica Panamericana, Madrid, 2011.

Laboratorio de Lingüística Informática de la Universidad Autónoma de Madrid. Las secciones siguientes explican la finalidad de estos programas, sus dificultades y las técnicas empleadas (§2). Posteriormente se describe nuestro método (§3) y se evalúan los resultados (§4). Tras la discusión de los datos (§5) incluimos unas conclusiones finales (§6).

2. El reconocimiento automático de términos en textos digitalizados

2.1. *Objetivos*

El reconocimiento automático de términos (RAT) consiste en identificar candidatos en textos o listas de palabras[3]. Se trata de una aplicación informática extensamente utilizada en lingüística computacional[4] y especialmente en el procesamiento de textos médicos[5]. Su interés original no fue la aplicación a la creación de recursos terminológicos sino la extracción de palabras o locuciones para identificar los temas de un documento. En este sentido, el RAT es una técnica de extracción de información y de minería de textos[6]. En efecto, para identificar el contenido semántico relevante de un texto es prioritario detectar sus conceptos representativos.

Naturalmente, para poder detectar nuevos términos se requieren textos recientes y representativos. La Lingüística de Corpus, con un fuerte desarrollo en los últimos veinte años, tiene como principal objetivo la compilación de textos de una variante lingüística o género textual. Los documentos han de estar en un formato digitalizado para permitir búsquedas y el tratamiento computacional (por ejemplo, la anotación morfosintáctica y el análisis estadístico). Una vez creado el corpus de textos médicos, el programa se encarga de extraer automáticamente candidatos. Esta aproximación empírica tiene la ventaja de la búsqueda exhaustiva sobre datos textuales, que está exenta del cansancio o descuido del ojo humano. Si bien los reconocedores automáticos carecen del conocimiento experto humano, su labor agiliza y facilita el trabajo del terminólogo o documentalista. Otras aplicaciones no

---

[3] Kyo Kageura y Bin Umino, "Methods of automatic term recognition: A review", *Terminology,* 3(2) (Ámsterdam, 1996), págs. 259-289; Michael Krauthammer y Goran Nenadic, "Term identification in the biomedical literature", *Journal of Biomedical Informatics,* 37 (Ámsterdam, 2004), págs. 512-526.

[4] Nitin Indurkhya y Fred J. Damerau (eds.), *Handbook of natural language processing,* 2ª ed., Chapman and Hall, Boca Raton, 2010.

[5] Sophia Ananiadou y John McNaught (eds.), *Text Mining for Biology and Biomedicine*, Artech House, Boston, MA, 2006.

[6] Kevin B. Cohen, "Biomedical Text Mining", en Nitin Indurkhya y Fred J. Damerau (eds.), *op. cit.*, 2010, págs. 605-625.

menos interesantes de los reconocedores automáticos son utilizar su lexicón para buscar ejemplos, recuperar y clasificar documentos o traducir términos a otra lengua.

En el ámbito terminológico, como explica Vivaldi[7], existen tradiciones metodológicas asentadas para enriquecer los recursos lexicográficos y construir bancos de datos con procedimientos estandarizados. No obstante, el ritmo de creación de neologismos en ciertos dominios hace muy costosa esta tarea. Es en este punto donde los sistemas de extracción terminológica resultan de gran ayuda, entendiendo siempre que las propuestas de términos tienen que ser aprobadas por un experto para su utilización final.

2.2. *Ámbito y dificultades*

La tradición terminológica define *término* o *unidad terminológica* como la realización lingüística de un concepto en un dominio especializado[8]. Desde la perspectiva del RAT, la tarea consiste en identificar los dos rasgos definitorios de un término[9]:

- *Unicidad* (*unithood*): el grado de cohesión o estabilidad de las palabras de una locución.
- *Termicidad* (*termhood*): el grado de especificidad del término con respecto a una disciplina. Por ejemplo, *hepático* está relacionado con el dominio médico, no con el aeronáutico.

Con respecto a la *unicidad*, las principales dificultades se concentran en el reconocimiento de las estructuras sintagmáticas y las fronteras entre palabras en los términos compuestos (*multiword terms*). Por ejemplo, el reconocedor debería detectar como candidato tanto *infarto*, *infarto de miocardio* e *infarto agudo de miocardio*, pero no *posible infarto*. Hay diferentes combinaciones sintagmáticas (N + ADJ, ADJ + N, N + PREP + N, N + ADJ + PREP + N…) pero la mayoría de las combinaciones que aparecen en un texto no son términos. Por tanto, hay que desarrollar estrategias concretas que separen unas combinaciones de otras.

Desde el punto de vista de la *termicidad*, es frecuente encontrarse con términos polisémicos que pertenecen a diferentes disciplinas. Por ejemplo, *nuclear* es tanto un término de la física como de la genética y biología. Por

[7] Jorge Vivaldi, *Extracción de candidatos a término mediante la combinación de estrategias heterogéneas*, Tesis doctoral, Universidad Politécnica de Cataluña, 2001, pág. 2.

[8] M.ª Teresa Cabré, *Terminology: Theory, methods and applications*, John Benjamins, Ámsterdam, 1999.

[9] Kyo Kageura y Bin Umino, *op. cit.*

ello, la utilización de recursos terminológicos de otra especialidad para realizar el contraste puede producir resultados erróneos.

Por otra parte, hay dos fenómenos que complican el reconocimiento de términos biomédicos: la *variación* y la *homonimia*. En el primer caso, el problema ocurre cuando el dominio contiene numerosas variantes formales del mismo término. Esto afecta tanto a los términos simples (como *aterosclerosis ~ateroesclerosis*) como a términos compuestos (*carcinoma microcítico de pulmón ~ carcinoma microcítico pulmonar*). Ananiadou y Nenadic[10] distinguen cinco tipos de variación terminológica, que básicamente son alternancias formales, y que adaptamos al español aquí:

- Ortográfica: *alfa-amilasa ~amilasa alfa ~ -amilasa*
- Morfológica: *obsesiva-compulsiva ~ obsesivo-compulsivas*
- Léxica: *infarto de corazón ~ infarto cardíaco*
- Estructural: *virus del papiloma humano ~ papilomavirus humano*
- Acrónimos y abreviaturas: *SST ~ ST*, ambos para referirse a *somatostatina*.

Además de la creación constante y continua de neologismos en el discurso biomédico, la influencia extranjera es fuente de nuevas variantes. Así, se pueden citar los múltiples calcos y préstamos adaptados de manera poco certera. Por ejemplo, en los textos españoles suelen convivir *bypass*, con *by pass* y *baipás*. Para terminar la clasificación de causas de la variación terminológica, hay que añadir la frecuente inserción de modificadores a términos ya formados: *deficiencia de hexosaminidasa A ~ deficiencia total de hexosaminidasa A*. En consecuencia, una tarea esencial tanto de los terminógrafos humanos como de los extractores automáticos es normalizar las variantes formales que representan el mismo concepto. Para ello se cuenta con la ayuda esencial de las ontologías y metatesauros multilingües, como los integrados en el *UMLS* (*Unified Medical Language System*)[11]. Este recurso incluye diversos tesauros y terminologías; entre ellos, *Medical Subject Headings* (*MeSH*), la *Systematized Nomenclature of Medicine – Clinical Terms* (*SNOMED-CT*) o la versión 10 de la Clasificación Internacional de Enfermedades (*ICD-10*, según las siglas en inglés). *UMLS* recoge códigos identificadores unicos de concepto (en inglés, *CUI*) asociados a cada variante terminológica en los distintos recursos. Por ejemplo, el código C0817096 designa al término *pecho* o *caja torácica* en *MeSH,* y también al término *torácico* o *tórax* en *SNOMED-CT*.

Por último, la homonimia de términos, especialmente acrónimos, supone otro desafío para los reconocedores automáticos. Por ejemplo, *IM* puede

[10] Sophia Ananiadou y Gorin Nenadic, "Automatic terminology management in biomedicine", en Sophia Ananiadou y John McNaught (eds.), *op. cit,* 2006, págs. 67-98.

[11] Olivier Bodenreider, "The Unified Medical Language System (UMLS): integrating biomedical terminology", *Nucleic Acids Research*, 32(Database issue) (Oxford, 2004), págs. 267-270.

referirse tanto a *insuficiencia mitral* como a *infarto de miocardio*. Sin contar con el conocimiento contextual y del dominio que tienen los terminólogos, no es fácil decidir a cuál de los dos conceptos se refiere la abreviatura. Algunos sistemas intentan solucionarlo con la restricción del lexicón a un campo concreto[12], pero en muchos casos esto es problemático porque los límites entre disciplinas biomédicas son difusos.

2.3. *Aproximaciones*

Aunque varios autores distinguen básicamente entre técnicas lingüísticas y técnicas estadísticas[13], en el reconocimiento de términos se suelen combinar métodos heterogéneos para conseguir los mejores resultados, como mostraremos en nuestro caso. De manera convencional, las diferentes aproximaciones a RAT se clasifican en cuatro tipos: a) basadas en diccionario; b) basada en reglas; c) basadas en estadística y aprendizaje automático; d) híbridas[14].

Los enfoques basados en diccionarios usan recursos en formato electrónico, como listas de palabras gramaticales y sin contenido (conocidas como *stop lists*), así como ontologías, glosarios y tesauros del dominio. Estas listas permiten filtrar el texto: con las primeras se eliminan palabras que no interesan y con las segundas se reconocen términos. Esta aproximación es la más simple y eficiente, pero suele ser muy incompleta y sobre todo no está disponible en todos los dominios ni a todos los investigadores. Un ejemplo es el experimento explicado en Segura-Bedmar *et al.*[15], donde utilizaron el metatesauro *UMLS* y otras listas de nombres de medicamentos genéricos, con el fin de reconocer y clasificar nombres farmacológicos en textos de biomedicina.

Los métodos basados en reglas se basan en el análisis de patrones de formación de términos (por ejemplo, compuestos por composición, uso de guiones, patrones sintagmáticos) y en conocimiento gramatical (análisis morfológico de los términos, listas de raíces y afijos). Este enfoque es bastante usado desde los 90. Por ejemplo, se ha empleado la descripción morfológica de raíces y afijos para detectar términos médicos[16]. Otros investiga-

---

[12] Almudena Ballester, Ángel Martín Municio, Fernando Pardos, Jordi Porta, Rafael J. Ruiz y Fernando Sánchez, "Combining statistics on *n*-grams for automatic term recognition", en *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*. Universidad de Las Palmas de Gran Canaria, 29-31 de mayo de 2002, págs. 1209-1213.

[13] Kyo Kageura y Bin Umino, *op. cit.*

[14] Michael Krauthammer y Goran Nenadic, 2004, *op. cit.*; Sophia Ananiadou y Goran Nenadic, *op. cit.*

[15] Isabel Segura-Bedmar, Paloma Martínez Fernández y Doaa Samy, "Detección de fármacos genéricos en textos biomédicos", *Procesamiento del Lenguaje Natural*, 40 (Jaén, 2008), págs. 27-34.

[16] Sophia Ananiadou, "A methodology for Automatic Term Recognition", *COLING'94 – Proc. of the 15th Int. Conf. on Computational Linguistics*, págs. 1034-1038, 1994.

dores utilizaron algoritmos basados en patrones de categorías concatenadas[17]. Para el español, se ha abordado el reconocimiento de sintagmas nominales para extraer términos médicos[18]. En general, se puede estimar una estrategia efectiva si procesa una lengua que recurre a bases grecolatinas para formar términos. Sin embargo este método no está extendido igualmente en todos los dominios e idiomas[19].

Con respecto a las técnicas estadísticas, su fundamento se basa en medir el grado de distintividad[20] de una palabra o lema en un contexto especializado en contraste con su frecuencia en un corpus general. Las dos más empleadas son el *test de la razón de verosimilitud* (*log-likelihood ratio test*[21]) y la métrica *logDice* empleada en *The Sketch Engine*[22]. La idea central de estas técnicas es conocer qué palabras o términos son sobreutilizados o infrautilizados en nuestro corpus de análisis en comparación con la frecuencia de las mismas palabras en un corpus de referencia. En nuestro caso, tomaremos un corpus de textos médicos (MultiMedica) y lo compararemos con el Corpus de Referencia del Español Actual (CREA), que contiene un conjunto equilibrado de textos de diferentes géneros y registros. Varios autores[23] han seguido esta aproximación. Sin embargo, hay más técnicas estadísticas, como la métrica de *información mutua* (*Mutual Information, MI*)[24] o el uso de semántica distribucional y coaparición léxica[25]. Para el español, se ha llevado a cabo un

---

[17] IDO DAGAN Y KEN CHURCH, "TERMIGHT: Identifying and Translating Technical Terminology", en *4th Conference on Applied Natural Language Processing*, págs. 34-40, 1994; JOHN S. JUSTESON Y SLAVA M. KATZ, "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering*, 1(1) (Cambridge, 1995), págs. 9-27.

[18] WALTER KOZA, ZULEMA SOLANA, MERLEY DA S. CONRADO, SOLANGE O.REZENDE, THIAGO A. PARDO, JOSUKA DÍAZ-LABRADOR Y JOSEBA ABAITUA, "Extracción terminological en el dominio médico a partir del reconocimiento de sintagmas nominales", *INFOSUR*, 5 (Rosario, Argentina, 2011), págs. 27-40.

[19] Para una comparación entre el uso de sufijos grecolatinos en inglés y japonés, dos lenguas de tradiciones muy distintas, consúltese CARLOS HERRERO ZORITA, CLARA MOLINA Y ANTONIO MORENO SANDOVAL, "Medical term formation in English and Japanese: A study of the suffixes -gram, -graph and -graphy", *Review of Cognitive Linguistics*, 13(1) (Ámsterdam, 2015), págs. 81-101.

[20] ANTONIO MORENO SANDOVAL Y JOSÉ Mª. GUIRAO, "Frecuencia y distintividad en el uso lingüístico: casos tomados de la lematización verbal de corpus de distintos registros", en *Actas del I Congreso Intl. de Lingüística de Corpus*, Universidad de Murcia, Murcia, 2009.

[21] TED DUNNING, "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics*, 19(1) (Cambridge, MA, 1993), págs. 61-74.

[22] ADAM KILGARRIFF, PAVEL RYCHLY, PAVEL SMRZ Y DAVID TUGWELL, "The Sketch Engine", en *Proceedings of EURALEX 2004*, Lorient, France, págs. 105-116.

[23] CHUNYU KIT Y XIAOYUE LIU, "Measuring mono-word termhood by rank difference via corpus comparison", *Terminology*, 14(2) (Ámsterdam, 2008), págs. 204-229.

[24] Un ejemplo de aplicación de la técnica de la IM se describe en HIROSHI NAKAGAWA Y TATSUNORI MORI, "Automatic term recognition based on statistics of compound nouns and their components", *Terminology*, 9(2) (Ámsterdam, 2003), págs. 201-219.

[25] ROGELIO NAZAR, JORGE VIVALDI Y LEO WANNER, "Automatic taxonomy extraction for spe-

experimento para detección de términos en un corpus de textos científicos empleando n-gramas y su probabilidad y distribución en un corpus[26]. También se ha empleado un algoritmo que analiza rasgos léxicos, morfológicos y sintácticos y los compara con un corpus de referencia[27].

Las aproximaciones basadas en aprendizaje automático son un tipo de estrategia estadística que consiste en entrenar los programas con los datos de un corpus previamente anotado con términos por especialistas. Los algoritmos de aprendizaje (entre otros, modelos ocultos de Markov, *HMM* en inglés; máquinas de soporte vectorial, *SVM*; y árboles de decisión) identifican rasgos característicos de los términos anotados y los aplican en un conjunto nuevo de datos. Son los llamados *clasificadores*, que dividen las palabras de un texto en términos y no términos.

Las técnicas híbridas combinan dos o más métodos de los mencionados. Normalmente, eligen una aproximación lingüística (o bien diccionarios o reglas de formación) y una métrica estadística. Existe ya un algoritmo desarrollado para la lengua española[28].

Hacer un repaso completo de la bibliografía en RAT queda fuera de los objetivos de este artículo. Remitimos a los interesados a la consulta de los trabajos que revisan el tema[29].

3. Explicación del experimento

Nos planteamos la siguiente cuestión de partida: ¿cuál es la mejor estrategia para obtener una lista de términos a partir de un corpus, sin contar con ninguna lista previa de ese dominio? Entendíamos que, si se partía de un cor-

---

cialized domains using distributional semantics", *Terminology*, 18(1) (Ámsterdam, 2012), págs. 188-225.

[26] Almudena Ballester et al., *op. cit.*, 2002.

[27] Rogelio Nazar y Mª. Teresa Cabré, "Un experimento de extracción de terminología utilizando algoritmos estadísticos supervisados", *Debate Terminológico,* 7 (2010), págs. 36-55.

[28] Alberto Barrón Cedeño, Gerardo Sierra, Patrick Drouin y Sophia Ananiadou, "An Improved Automatic Term Recognition Method for Spanish", en Alexander Gelbukh (ed.), *CICLing2009 LNCS 5449.* Springer, Berlín, 2009, págs. 125-136.

[29] Existen dos excelentes libros: Didier Bourigault, Christian Jacquemin y Marie-Claude L'Homme (eds.), *Recent Advances in Computational Terminology*, J. Benjamins, Ámsterdam, 2001; y Sophia Ananiadou y John McNaught, *op. cit.* Otros trabajos relevantes son los de Kyo Kageura y Bin Umino, *op. cit.*; Adam Kilgarriff, "Which words are particularly characteristic of a text? A survey of statistical approaches", en *Proc. AISB Workshop on Language Engineering for Document Analysis and Recognition,* Sussex University, abril de 1996, págs. 33-40; Mª. Teresa Cabré, Rosa Estopá y Jorge Vivaldi, "Automatic term detection: A review of current systems", en Didier Bourigault, Christian Jacquemin y Marie-Claude L'Homme (eds.), *op. cit.*, págs. 53-87; Jorge Vivaldi, *op. cit.*; y Michael Krauthammer y Goran Nenadic, *op. cit.*

pus suficientemente amplio y representativo, se podían extraer un buen número de candidatos que validar posteriormente con fuentes fiables. Sin embargo, no teníamos ninguna hipótesis previa de cuál de las tres aproximaciones (léxica, basada en reglas, o estadística) daría los mejores resultados para nuestro caso concreto.

3.1. *El corpus*

El experimento se ha realizado empleando los datos de la parte española del corpus MultiMedica[30]. El subcorpus está formado por 4200 documentos con un total de más de 4 millones de palabras. La tipología textual va desde artículos divulgativos escritos por médicos para usuarios no especialistas (generalmente editados por periodistas) hasta textos científicos dirigidos a profesionales de la salud. Claramente, predominan los textos especializados y técnicos (más del 80%), donde la mayor parte de las especialidades médicas están representadas de manera equilibrada. Por ello, consideramos que era una fuente fiable y suficiente para la obtención de términos, de modo que se pudiera realizar el experimento con resultados válidos.

Aunque no se empleó exhaustivamente en el experimento, el corpus fue etiquetado morfosintácticamente con categoría y lema, con el fin de permitir búsquedas y concordancias[31].

3.2. *La metodología y etapas del proceso*

En el experimento solo consideramos el reconocimiento de términos simples formados por una única palabra (como *aspirina* o *ADN*) o por palabras que forman parte de un compuesto (p. ej., *ascórbico* en *ácido ascórbico*, o *Down* en *síndrome de Down*). El objetivo fue evaluar cuál de las tres estrategias de selección proporciona mejores resultados. El proceso siguió tres pasos (Figura 1):

1) Preselección de los candidatos mediante cada uno de tres métodos.
2) Filtrado de candidatos a término mediante una lista de afijos y raíces biomédicos.

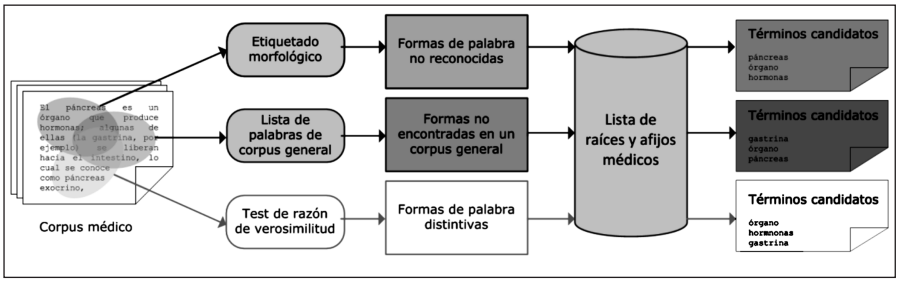
---

[30] Antonio Moreno Sandoval y Leonardo Campillos Llanos, "Design and annotation of MultiMedica - a multilingual text corpus of the biomedical domain", en C. Vargas-Sierra (ed.), *Procedia*, 95, Elsevier, Berlín, 2013, págs. 33-39.

[31] Los textos y el extractor de términos se encuentran en la dirección: <www.lllf.uam.es/ESP/Multimed.html> [11/09/2015]

3) Comprobación manual de cada candidato consultando referencias bibliográficas.

FIGURA 1. *Esquema de los pasos seguidos en el experimento.*



3.2.1. Preselección de los términos por cada método

Cada método de extracción de candidatos se basa en una estrategia distinta, y por tanto la lista obtenida en cada uno tiene un tamaño diferente aunque se aplique al mismo conjunto de datos. Sin embargo, obtener más candidatos no significa que se acierte más. Veamos cada caso.

El primer método consiste en utilizar un etiquetador morfológico. Es un ejemplo del tipo *basado en reglas*: el analizador contiene un conjunto de reglas de reconocimiento y análisis de palabras en español. Para este experimento solo nos interesan aquellas palabras que etiqueta como *desconocidas*, porque asumimos que los términos médicos tienen un tipo de estructura morfológica que no está recogida en un analizador como el que utilizamos: GRAMPAL[32], que cubre un lexicón de más de 50 000 lemas de uso general y analiza más de 500 000 formas flexionadas. Obviamente, GRAMPAL contiene entre sus lemas bastantes palabras médicas que se han incorporado al léxico común, como recoge cualquier diccionario monolingüe de referencia (por ejemplo, el *DRAE* o el María Moliner). Pero al igual que un diccionario general, la mayoría de palabras específicas del dominio no están presentes en él (p. ej., *ADN* o *distal*). Después de procesar con el analizador el corpus de 4 millones de palabras se recogió un total de 22 413 casos desconocidos, con los que formamos una lista de candidatos.

El segundo método es del tipo *basado en corpus*: contrastamos todas las palabras del corpus MultiMedica con la lista de palabras de un corpus gene-

[32] ANTONIO MORENO SANDOVAL Y JOSÉ Mª. GUIRAO MIRAS, "Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation", en YUJI KAWAGUCHI, SUSUMU ZAIMA Y TOSIHIRO TAKAGAKI (eds.), *Spoken Language Corpus and Linguistic Informatics*, John Benjamins, Ámsterdam, 2006, págs. 199-218.

ral del español, el *CREA*[33]. Por tratarse de un corpus grande y equilibrado en géneros y registros, se puede considerar como una referencia fiable sobre el uso general del español. Así, el *CREA* contiene más 150 millones de palabras y unas 700 000 formas diferentes. Sin embargo, esta lista presenta casi un 50% de palabras *ruidosas* para el experimento: extranjerismos, errores orto-tipográficos y nombres propios[34]. Una limpieza exhaustiva redujo la lista a unas 350 000 formas. De nuevo, muchos términos médicos de uso general aparecen en esta lista y, por otra parte, eliminamos nombres propios como Down y Alzheimer que forman parte de términos. Sin embargo, al contrastar el número de nombres propios no relevantes para la medicina, preferimos eliminarlos todos a pesar de la utilidad de algunos de ellos. Con este método se seleccionó un total de 23 239 candidatos, que son palabras que no están en la lista depurada del corpus CREA[35].

El tercer método es puramente estadístico: aplicamos el test de la razón de verosimilitud (*Log-Likelihood* o *LLH*[36]) al reconocimiento de palabras distintivas en el corpus médico. Este test es de uso generalizado en programas de concordancias (por ejemplo, Wordsmith o AntConc) para extraer palabras clave de un texto. Consiste en contrastar las frecuencias de aparición de las distintas palabras de nuestro corpus con otro de referencia. Para el contraste con el corpus MultiMedica usamos la lista del *CREA* depurada por nosotros. Afortu-nadamente, esta lista contiene las frecuencias absolutas y normalizadas de cada palabra y, por tanto, es la mejor lista disponible de manera abierta para hacer comparaciones con un léxico general del español. Para obtener un 99.9% de tasa de confianza, aplicamos el corte de significancia en 10.83. En consecuencia, solo incluimos en la lista de candidatos aquellas que tenían un valor del test por encima de 10. El resultado fue una lista de 8667 candidatos[37] .

3.2.2. Filtrado con la lista de afijos y raíces

Un vistazo rápido a los candidatos proporciona la evidencia de que hay que añadir algún filtro a la lista pues contiene palabras que simplemente no

---

[33] La lista completa del corpus CREA se puede obtener en la dirección: <http://corpus.rae.es/lfrecuencias.html> [11/09/2015]

[34] Esto se debe principalmente a que contiene textos académicos y literarios, donde abundan los extranjerismos y los nombres propios. Los errores orto-tipográficos se deben a que los textos se escanearon directamente de originales que los contenían.

[35] Queremos llamar la atención sobre el hecho de que un lexicón de 50000 lemas como el de GRAMPAL genera unas 150000 formas diferentes más que las contenidas en un corpus de 150 millones de palabras.

[36] TED DUNNING, *op. cit.,* 1993.

[37] Este valor de corte se ha eligió en función de nuestra experiencia previa con otros corpus

estaban recogidas ni en el lexicón del analizador morfológico ni en la lista del CREA, pero que son palabras del dominio general (p. ej., *tabúes o vinculador*). Para mejorar la precisión de los términos seleccionados aplicamos un programa de reconocimiento de afijos y raíces de términos médicos. El programa recoge 2128 ítems (incluyendo variantes ortográficas como ADEN- y ADENO-)[38], a saber:

- Afijos grecolatinos del dominio médico (CARDIO-, -ITIS) y raíces médicas frecuentes (PANCREA-), recopiladas a partir de distintas fuentes de terminología médica[39]. Para evitar falsos positivos, se excluyeron de esta lista los afijos de uso muy frecuente y general que no están restringidos al dominio biomédico (como PRE- o -ABLE).
- Raíces y afijos para el reconocimiento de compuestos farmacológicos (-CAVIR) y sustancias bioquímicas (BUT- o -STEROL). Todos ellos fueron compilados de las listas propuestas por la Organización Mundial de la Salud (OMS)[40] así como por los listados aprobados por la American Medical Association (AMA) para la nomenclatura de compuestos clínicos[41]. Como la mayoría de los afijos en inglés tienen una correspondencia unívoca con los términos españoles, la adaptación fue directa, especialmente para los acabados en vocal como -INE > -INA (p. ej., *creatine* > *creatina*).

Para obtener la lista final se han generado todas las posibles variantes de cada afijo o raíz. Por una parte, las variantes gráficas debidas a la tilde: así

---

de menor tamaño. Esta decisión explica el reducido número de casos que se obtienen, en comparación con los otros métodos. Después de analizar los datos, nos percatamos que este valor de corte es probablemente bastante estricto para un corpus del tamaño del usado en este experimento. Este hecho podría explicar los pobres valores obtenidos en cobertura y los excelentes en precisión, como discutiremos más adelante.

[38] También se usaron formas grecolatinas en: ROSA ESTOPÀ, JORGE VIVALDI Y Mª. TERESA CABRÉ, "Use of Greek and Latin forms for term detection", en *Proc. of the 2nd Intl, Conf. on Language Resources and Evaluation (LREC 2000),* Atenas, Grecia, 31 de mayo-2 de junio del 2000; JORGE VIVALDI, *op. cit.*, 2001; y OLATZ PÉREZ-DE-VIÑASPRE, MAITE ORONOZ, MANEX AGIRREZABAL Y MIKEL LERSUNDI, "A Finite-State Approach to Translate SNOMED CT Terms into Basque Using Medical Prefixes and Suffixes", en *Proc. of the 11th Intl. Conf. on Finite State Methods and Natural Language Processing,* St Andrews, 15-17 de julio de 2013, págs. 99-103.

[39] JOSÉ Mª. LÓPEZ PIÑERO Y Mª. LUZ TERRADA FERRANDIS, *Introducción a la terminología médica*, Masson, Barcelona, 2005; Mª. ELENA JIMÉNEZ, "Afijos grecolatinos y de otra procedencia en terminos medicos", *MEDISAN,* 16(6) (Santiago de Cuba, 2012), págs. 1005-1021; MIGUEL A. SÁNCHEZ GONZÁLEZ, *Historia de la medicina y humanidades médicas,* 2ª ed., Elsevier/Masson, Barcelona, 2012.

[40] OMS, "The use of stems in the selection of International Nonproprietary Names (INN) for pharmaceutical substances", 2011a, <http://apps.who.int/medicinedocs/documents/s19117en/s19117en.pdf> [11/09/2015]; OMS, "International Nonproprietary Names (INN) for biological and biotechnological substances (a review)", 2011b, <http://apps.who.int/medicinedocs/documents/s19119en/s19119en.pdf> [11/09/2015].

[41] <www.ama-assn.org/resources/doc/usan/stem-list-cumulative.pdf>, <www.ama-assn.org/

PRÓST- (como en *próstata*) y PROST- (como en *prostático*). Por otra parte, la alternancia de variantes debida a una vocal epentética: ESCOLI- y SCOLI-. Por último, las variantes debidas a la flexión de género y número: así, el sufijo -GÉNICO se representa por cuatro formas, -GÉNICO, -GÉNICA, -GÉNICOS, -GÉNICAS.

El programa que contrasta los afijos con los candidatos funciona de la siguiente manera. Primero, se contrasta cada candidato con todos los afijos de dos listas diferentes (prefijos y sufijos). La búsqueda es recursiva, de manera que cada cadena de caracteres del candidato es comparada hasta no registrar más coincidencias[42]. Cuando un candidato contiene un afijo o raíz biomédico, se le considera un término potencial. La Figura 1 resume todo el proceso. Véase también la Tabla 2 para los datos de términos aceptados y rechazados por el filtrado de afijos.

3.2.3. Verificación manual de cada término propuesto

Finalmente, todos los términos seleccionados en el proceso automático fueron revisados manualmente. Uno de los autores (LCL) confirmó o rechazó cada término propuesto, marcándolo como *Aceptado* o *Rechazado*. El resultado final fue un *gold standard* o conjunto de términos de referencia con todas las formas aceptadas. ¿Qué criterios se han seguido para considerar que un término es aceptado o rechazado? La idea central es que un término válido debe aparecer en una fuente médica reconocida. De hecho, para evitar la subjetividad o discrecionalidad, hemos basado la decisión en las siguientes obras de referencia y por este orden de autoridad:

- *Diccionario de términos médicos*[43]: recoge casi 52 000 términos con una perspectiva normativa: muestra preferencia por ciertas variantes sobre otras y advierte contra el uso de extranjerismos y calcos.
- *Diccionario médico enciclopédico Dorland*[44]: contiene más de 112 000 términos.
- *Diccionario Espasa Medicina*[45]: es un diccionario con 18 000 entradas, recopiladas por un equipo de médicos de la Universidad de Navarra.

---

resources/doc/usan/new-stem-list.pdf> [11/09/2015]. También se consultó la lista de afijos de Michael Quinion: <hwww.affixes.org> [11/09/2015]

[42] Los autores agradecen la ayuda de la Dra. Alicia González Martínez con el procesamiento de los afijos.

[43] REAL ACADEMIA NACIONAL DE MEDICINA, *op. cit.*

[44] DORLAND, *Diccionario enciclopédico ilustrado de medicina Dorland*, Elsevier, Madrid, 30ª edición, 2005; y 32ª edición en línea, 2012: <https://dorlandsonline.com/> [11/09/2015].

[45] LUIS Mª. GONZALO SANZ (coord.), *Diccionario Espasa Medicina,* Espasa S.L., Madrid, 1999.

- *Dicciomed*[46]: es un recurso de enfoque etimológico y de referencia histórica en la aparición de términos. Solo está disponible en línea y recoge unos 7000 términos.

También hemos aceptado como términos válidos aquellos que hemos podido localizar en revistas y libros de investigación biomédica. Para ello hemos empleado el corpus de Google Books en línea[47]. La Tabla 1 resume los criterios de clasificación que hemos empleado para aceptar o rechazar un término, así como algún ejemplo característico.

TABLA 1. *Las cuatro clases de términos*

	Clasificación de los términos	*Ejemplos*
Aceptados	■ Lista 1: términos que tienen una entrada en algún diccionario médico de referencia. ■ Lista 2: términos sin entrada en diccionario de referencia, pero registrados en libros o artículos científicos.	*páncreas, ADN* … *RAS, cisteínico* …
Rechazados	■ Lista 3: términos rechazados por especialistas debido a errores ortotipográficos o una mala adaptación al español. ■ Lista 4: términos no biomédicos.	**perirenal, *croup*… *Aragón, Pfizer* …

La biomedicina es un área extensa de investigación en la que es problemático reducir los límites del dominio. Los términos del *gold standard* proceden de campos como la Anatomía (*hígado, nefrona*), la Microbiología (*cilio, "Escherichia"*), la Genética (*transcripción, ARN*), la Oncología (*oncogén, leucemia*), la Bioquímica (*fosforilación, amina*), la Farmacología (*aspirina, prozac*), la Historia de la Medicina (*frenología, miasma*), o la cirugía y procedimientos o técnicas médicas (*tomografía, maniobra*), entre otros. También aceptamos en nuestro listado, en menor medida, términos de otras disciplinas no estrictamente relacionados con la biomedicina, pero muy comunes en textos del dominio. Por ejemplo, conceptos referidos a medidas estadísticas (*variable, significancia*), agentes implicados en un trastorno, como animales venenosos o condiciones ambientales (*Anopheles, vipéridos, contaminación*) o plantas productoras de sustancias farmacológicas (*Vinca, cornezuelo*). En total la lista final

[46] FRANCISCO CORTÉS GABAUDÁN (coord.), *Dicciomed,* 2007-2013. <http://dicciomed.eusal.es> [11/09/2015].

[47] GOOGLE BOOKS, <https://books.google.com/> [11/09/2015].

de términos aceptados contiene 24 639, algo por encima de los más de 20 000 candidatos extraídos mediante el analizador y la lista del CREA; pero claramente muchos términos aceptados se repetían en las tres listas. Lo analizamos en la siguiente sección.

4. Resultados de la evaluación de los tres métodos

Hemos empleado la lista de términos aceptados (*gold standard*) para evaluar el grado de acierto y error de cada método. La Tabla 2 muestra los resultados de los candidatos filtrados por la lista de afijos y raíces, descrito en el paso segundo del procedimiento. Se aprecia que el método con el etiquetador GRAMPAL ha obtenido un 65% de aceptados y el método estadístico se ha quedado en el 44%. La selección basada en los ítems no incluidos en el corpus CREA se ha quedado en posición intermedia con un 53% de acierto. En términos absolutos, también el método basado en el analizador proporciona muchos más casos fiables, especialmente en comparación con el método estadístico. Por tanto, estos serían los resultados que el programa propone como términos. El siguiente paso es comprobar cuántos de estos términos propuestos son verdaderamente términos correctos, según nuestro *gold standard*, que contiene 24 639 términos y que son la intersección, sin repeticiones, de las tres listas.

Tabla 2. *Candidatos aceptados y rechazados por la lista de afijos y raíces*

	Aceptados		*Rechazados*		*Total*
	Formas	%	*Formas*	%	*Formas*
No en GRAMPAL	14551	64.92%	7862	35.08%	22413
No en CREA	12307	52.96%	10932	47.04%	23239
Log-Likelihood (*LLH*)	3832	44.21%	4835	55.79%	8667

En Lingüística Computacional y concretamente en el campo de la recuperación de información se emplean de manera generalizada tres medidas de evaluación: precisión, cobertura (*recall*) y la medida F.

La *precisión* es el porcentaje de términos correctos con respecto al total de términos propuestos. Así, por ejemplo, en el caso del método basado en GRAMPAL, el programa ha seleccionado 14 551 candidatos, pero solo 12 718 son correctos (aparecen en el *gold standard*). Entonces, la precisión se calcula mediante 12 718 / 14 551 = 87.40%.

La *cobertura* es el porcentaje de términos correctos con respecto al total de términos correctos que se han encontrado, es decir, el *gold standard*. Por tanto, la cobertura del método basado en GRAMPAL se calcula mediante 12 718 / 24 639 = 51.62%.

Finalmente, la *medida F* es la media harmónica de la precisión y la cobertura. Se calcula mediante la siguiente fórmula[48]:

$$F = \frac{P\ R}{+P + R,}$$

Los cálculos completos para los tres métodos se muestran en las Tablas 3 y 4 (en negrita se muestran los valores más altos para cada métrica)

TABLA 3. *Recuentos absolutos*

	Propuestos	*Correctos*	*Erróneos*
No en GRAMPAL	14 551	12 718	1833
No en CREA	12 307	10 213	2094
Log-Likelihood (*LLH*)	3832	3525	307

TABLA 4. *Medidas de evaluación estándares para los tres métodos*

	Propuestos	*Correctos*	*Medida F*
No en GRAMPAL	87.40%	**51.62%**	**64.90%**
No en CREA	82.99%	41.45%	55.29%
Log-Likelihood (*LLH*)	**91.99%**	14.31%	24.76%

5. DISCUSIÓN

5.1. *Análisis de la evaluación*

Las tres medidas estándares aportan información sobre el funcionamiento de cada método. Una mayor precisión implica que el programa se equivoca menos y por tanto da una medida de fiabilidad en cuanto a los candidatos que ofrece. Por su parte, una mayor cobertura significa que el programa tiende a ser más exhaustivo en la búsqueda de candidatos (independientemente de su corrección). Finalmente, la medida F se suele tomar como la más apropiada para evaluar la actuación global del programa. En conjunto, cada una ofrece de manera cuantificada un aspecto diferente sobre el rendimiento práctico del sistema evaluado. Distintos usuarios del programa preferirán un aspecto a otro en función de sus intereses. Por ejemplo, un traductor acaso prefiera contar con más cobertura que precisión: la búsqueda deta-

[48] DAN JURAFSKY y JAMES MARTIN, *Speech and Language Processing*, 2ª ed., Prentice Hall, New Jersey, 2009.

llada de términos en un texto extenso supone más esfuerzo que descartar los no términos en una lista ordenada.

Los resultados mostrados en la Tabla 4 permiten avanzar algunas conclusiones:

- El empleo de un analizador morfológico, apoyado con el filtrado de afijos y raíces, proporciona la mejor medida F y cobertura de los tres métodos, consiguiendo al mismo tiempo una excelente precisión. Estos resultados están en línea con los obtenidos por los sistemas de recuperación de información de entidades biológicas, donde la precisión se mueve en el rango del 80-90% y la cobertura entre el 50-60%[49].
- La aproximación estadística consigue la mejor tasa de precisión (92%) pero a costa de una cobertura muy baja (14%). Esto compromete la media armónica (25%) y deja el funcionamiento global muy por debajo de los otros dos sistemas.
- El método basado en la lista de un corpus general se queda en una posición intermedia aunque con la peor precisión de los tres.

Nos preguntamos por el grado de compartición de términos entre los tres subconjuntos de candidatos para comprobar si hay un alto grado de redundancia entre ellos. Así, la Tabla 5 recoge los términos que comparten dos métodos entre sí. Los datos se han calculado sobre las listas de candidatos filtrados ya por los afijos. Queremos llamar la atención sobre el hecho de que las listas basadas en el CREA y en test estadístico no comparten ningún caso, ya que precisamente se empleó el corpus como datos de contraste (ver explicación en 3.2.1.).

Efectivamente, los datos muestran que las listas generadas con el analizador morfológico y el corpus comparten casi un 40% del *gold standard*. Por tanto, hay cierto grado de redundancia entre ellas. Por el contrario, entre el método de GRAMPAL y el test estadístico solo comparten 1203 términos, pero es un tercio de los términos generados por la medida estadística.

TABLA 5. Ítems compartidos entre dos listas de candidatos

Listas de candidatos	*Términos compartidos del* gold standard	
	Términos	*Porcentaje del* gold standard *(24639)*
No en GRAMPAL y No en CREA	9222	37.42%
No en GRAMPAL y en *LLH*	1203	4.88%
No en CREA y en *LLH*	0	0.00%

[49] GORAN NENADIC, IRENA SPASIC Y SOPHIA ANANIADOU, "Terminology-driven mining of biomedical literature", *Bioinformatics*, 19(8) (Oxford, 2003), págs. 938-943.

Con el fin de estimar la aportación del filtrado de afijos a la selección de candidatos, contrastamos la lista de afijos con el *gold standard*. Como método independiente, llega a reconocer 16 117 términos, es decir, un 65.41% del total. Por tanto, la lista de afijos por sí misma da una buena tasa de precisión, que mejora en combinación con cualquiera de los otros métodos. Esto se explica porque la creación terminológica en el dominio médico está muy dirigida por los temas grecolatinos. En contrapartida, la mejora de precisión disminuye la cobertura: los neologismos que se forman sin base grecolatina (p. ej., *VIH* o *Alzhéimer*) se descartan de la lista inicial aportada por cada método. En las Tablas 6 y 7 mostramos las diferencias en precisión y cobertura entre las listas sin y con filtrado de afijos. Recordemos que el *gold standard* se ha creó verificando cada término propuesto en las listas originales antes del filtrado de afijos.

TABLA 6. *Precisión sin y con filtro de afijos*

Método	*Cobertura*	*Precisión con filtro de afijos*	*Diferencia*
No en GRAMPAL	73.96%	87.40%	+13.45%
No en CREA	60.21%	82.99%	+22.75%
Log-Likelihood (*LLH*)	79.57%	91.99%	+12.42%

TABLA 7. *Cobertura sin y con filtro de afijos*

Método	*Cobertura*	*Cobertura con filtro de afijos*	*Diferencia*
No en GRAMPAL	67.28%	51.56%	-15.71%
No en CREA	56.79%	41.45%	-15.34%
Log-Likelihood (*LLH*)	27.99%	14.28%	-13.71%

5.2. *Análisis de los errores en la extracción de términos*

Tras analizar las cifras generales, nos interesó conocer los problemas de cada método. Los métodos lingüísticos (basados en un corpus o un analizador morfológico) superaron al enfoque estadístico en nuestro experimento. Aunque tienen algo menos de precisión, la cobertura com-pensa su uso. Por su parte, el test de Dunning (*LLH*) obtuvo una baja cobertura y por sí mismo no es recomendable (a pesar de que es el método más usado, por su bajo coste y disponibilidad).

Una posible explicación de la menor precisión obtenida por el método basado en un corpus general es que el *CREA* incluye textos del dominio. Cuando comparamos la lista de palabras del *CREA* con las del corpus

MultiMedica, nos encontramos que muchos términos válidos no se seleccionaban porque aparecían en la lista general del *CREA*. Estos falsos negativos son, sobre todo, términos biomédicos muy extendidos: por ejemplo, *médula, mitocondria, occipital* o *cirrosis*.

Análogamente, la lista obtenida a partir de las palabras desconocidas en GRAMPAL carece de aquellos términos que están incluidos en lexicón del analizador, como *vacuna, enzima* o *virus*. Tampoco se reconocieron epónimos como *Alzheimer* o *Down*, pues en un caso fueron eliminados (lista *CREA* depurada) o fueron reconocidos como nombres propios (GRAMPAL). Si comparamos ambos métodos, GRAMPAL es más preciso que el basado en el *CREA*. Como se comentó anteriormente, GRAMPAL reconoce muchas más formas que las incluidas en el *CREA*, lo cual afecta a la precisión, porque el analizador rechaza formas del lenguaje no médico que no aparecen en el corpus. Por otra parte, GRAMPAL contiene menos palabras especializadas que el *CREA*. Por tanto, es posible que los resultados se puedan acercar mucho o incluso mejorar a los de GRAMPAL empleando un corpus general más amplio que el *CREA*.

El test estadístico de Dunning (*LLH*) reconoce satisfactoriamente términos distintivos del corpus. A diferencia de los otros dos métodos lingüísticos, el test *LLH* elige términos biomédicos generales y frecuentes (*cuello, salud, microbio*). Otro aspecto reseñable es su capacidad para seleccionar términos polisémicos con significados biomédicos y generales. Por ejemplo, *mejoría* (clínica), *diferenciarse* (una célula), *reducir* (una sustancia) o *practicar* (una operación). Esta técnica es la única que permite esta selección y puede ayudar a explicar por qué se obtiene una mejor precisión.

La contrapartida del test de Dunning es su baja cobertura. La principal razón es su predisposición por palabras muy generales y frecuentes, distintivas en el corpus MultiMedica, aunque no estén relacionadas con el dominio biomédico. Por ejemplos, verbos como *pueden* o *suele* aparecen frecuentemente en los textos médicos para indicar posibilidad o tendencia general.

Un problema que presentan los tres métodos es su dificultad para reconocer la mayoría de los epónimos y términos referidos a entidades biomédicas como sustancias farmacológicas, acrónimos o nombres de proteínas. El test de Dunning es el que peor trata este problema, pues necesita tener un número suficiente de ejemplos del mismo token para seleccionarlo.

Con respecto al filtro de afijos y raíces, no se aprecian diferencias entre las listas. Los errores (falsos positivos) se deben básicamente a homografía. Es decir, la coincidencia formal entre afijos y palabras que contienen la misma cadena: *-AMINO* aparece en *amino**ácido*** (verdadero positivo) pero también en *amino**ra*** (falso positivo). Por esta razón, no usamos afijos con menos

de dos caracteres. Por ejemplo, *AN-* puede aparecer en *anaerobio* pero también en *Andalucía*.

Los afijos también seleccionan candidatos con errores orto-tipográficos. Por ejemplo, *CARDIO-* filtra *cardiovascular* y **cardiovacular*. Tampoco permiten descartar extranjerismos, ya que muchos afijos son homógrafos en español e inglés: *angiograma*, *angiography*. Los terminólogos son muy sensibles a estos calcos de términos extranjeros al español. Los nombres comerciales de fármacos que incluyen un afijo médico tampoco deben recogerse en los glosarios terminológicos y nosotros no los aceptamos en la lista. La Tabla 8 proporciona distintos ejemplos.

TABLA 8. *Ejemplos de homografías con afijos y raíces*

Afijo/raíz	*Verdaderos positivos*	*Falsos positivos*
aden-	*adenomatoso*	*adenomatous*
angio-	*angiograma*	*angiography*
bacteri-	*bacteriófago*	*bacterial*[50]
cardio-	*cardiopatía*	**cardiovacular*
hemo-	*hemoglobina*	*Hemovas®*
neuro-	*neurólogo*	*Neurontin®*
-geno	*broncógeno*	**cangerígenos*
-manía	*potomanía*	*Rumanía*
-ismo	*aldosteronismo*	*asimismo*

Es interesante analizar la productividad de los afijos para filtrar términos médicos correctamente. La Tabla 9 muestra los veinte afijos más productivos, entre los que destacan por orden de eficacia *CITO- /-CITO* (90.97% de casos correctos), *NEURO-* (89.86%), *MICRO-* (87.74%), *ANTI-* (84.91%), *-OSIS* (84.44%), *-ARIO/A* (84.23%), *SUB-* (78.60%), *TRANS-* (77.54%) y *DIS-* (74.63%).

[50] Calco del inglés, en español se prefiere *bacteriano*.

TABLA 9. *Los veinte afijos más productivos con ejemplos positivos y negativos*

Verdaderos positivos		*Falsos positivos*	
Afijo	*Casos*	*Afijo*	*Casos*
anti-	529	***anti-***	94
hiper-	335	***dis-***	69
hipo-	291	***trans-***	69
neuro-	257	*auto-*	65
cito- /-cito	252	*inter-*	50
trans-	238	*per-*	48
-asa	235	***sub-***	46
-ario/a	219	***-ario/a***	41
-osis	217	***-osis***	40
dis-	203	*multi-*	37
inmun-	197	*prop-*	35
micro-	186	*pred-*	30
peri-	181	*-itis*	30
intra-	175	***neuro-***	29
poli-	169	***prote-***	29
sub-	169	*-dor(a)*	26
prote-	159	***micro-***	26
endo-	155	***-cito/cito-***	25
epi-	144	*intra-*	24

El problema del filtrado de afijos es el rechazo de términos que no contienen afijos o raíces; por ejemplo, nombres, adjetivos y verbos muy comunes en el dominio (*salud, agudo, cicatrizar*) o términos anatómicos (*rodillas*). Como con los otros métodos, no se identifican genes (*RAS*), acrónimos (*SIDA*), epónimos (*Parkinson*), sustancias farmacológicas (*codeína*) o nombres científicos ("*Vibrio*"). El remedio puede ser emplear tesauros como los incluidos en el *UMLS*, aunque su extensa cobertura de términos obliga a restringir la clase de entidades según su tipo semántico.

6. CONCLUSIONES

Hemos comparado tres métodos para seleccionar candidatos de un corpus. Los resultados pueden resumirse de la siguiente manera: a) independientemente del método empleado, el filtrado con afijos mejora la precisión, b) por el contrario, en todos los casos el filtrado de afijos y raíces hace bajar la cobertura; c) el método basado en el analizador morfológico es el que consigue mejor cobertura; d) en el funcionamiento global (medida F) gana de nuevo el analizador, pero sin filtrado de afijos y raíces.

A continuación, hemos analizado las causas de los falsos positivos. Cada método tiene sus propios problemas, así como el filtrado de afijos y raíces. Los principales errores se cometen por no reconocer abreviaturas, epónimos, nombres de genes, compuestos farmacológicos y sustancias bioquímicas. La mejor manera de resolverlo es recurrir a listas fiables de estos términos.

Finalmente, esta investigación ha producido un recurso lingüístico digital: una lista de casi 25 000 términos a partir de la combinación de los tres métodos. Cada candidato fue comprobado manualmente mediante su aparición en diccionarios especializados, libros y revistas biomédicos. El recurso digital se ha empleado en la evaluación de los tres métodos y también es la base del extractor terminológico automático[50] para español desarrollado en el proyecto MultiMedica.

[51] Leonardo Campillos Llanos, Antonio Moreno Sandoval y José Mª. Guirao Miras, "An automatic term extractor for biomedical terms in Spanish", en *Proc. of the 5th International Symposium on Languages in Biology and Medicine* (LBM 2013), 12 y 13 de diciembre de 2013, Tokio, Japón. El extractor se encuentra disponible en la dirección <https//cartago.lllf.uam.es/corpus3/extractor.pl?menu=extractor> [11/09/2015].

REFERENCIAS BIBLIOGRÁFICAS

ANANIADOU, SOPHIA, "A methodology for Automatic Term Recognition", *COLING'94 – Proc. of the 15th Int. Conf. on Computational Linguistics*, 1994, págs. 1034-1038.

ANANIADOU, SOPHIA; MCNAUGHT, JOHN (eds.), *Text Mining for Biology and Biomedicine*, Artech House, Boston, MA, 2006.

ANANIADOU, SOPHIA; NENADIC, GORIN, "Automatic terminology management in biomedicine", en SOPHIA ANANIADOU Y JOHN MCNAUGHT (eds.), *Text Mining for Biology and Biomedicine*, Artech House, Boston, MA, 2006, págs. 67-98.

BALLESTER, ALMUDENA; MARTÍN MUNICIO, ÁNGEL; PARDOS, FERNANDO; PORTA, JORDI; RUIZ, RAFAEL J.; SÁNCHEZ, FERNANDO, "Combining statistics on *n*-grams for automatic term recognition", en *Proc. of the 3rd Intl. Conference on Language Resources and Evaluation (LREC'02),* Universidad de Las Palmas, 29-31 de mayo de 2002, págs. 1209-1213.

BARRÓN CEDEÑO, ALBERTO; SIERRA, GERARDO; DROUIN, PATRICK; ANANIADOU, SOPHIA, "An Improved Automatic Term Recognition Method for Spanish", en ALEXANDER GELBUKH (ed.), *CICLing 2009, LNCS 5449*. Springer-Verlag, Berlín, 2009, págs. 125-136.

BODENREIDER, OLIVIER, "The Unified Medical Language System (UMLS): integrating biomedical terminology", *Nucleic Acids Research*, 32(Database issue) (Oxford, 2004), págs. 267-270.

BOURIGAULT, DIDIER; JACQUEMIN, CHRISTIAN; L'HOMME, MARIE-CLAUDE (eds.), *Recent Advances in Computational Terminology*, John Benjamins, Ámsterdam, 2001.

CABRÉ, Mª. TERESA; ESTOPÁ, ROSA; VIVALDI, JORGE, "Automatic term detection: A review of current systems", en DIDIER BOURIGAULT, CHRISTIAN JACQUEMIN Y MARIE-CLAUDE L'HOMME (eds.), *Recent Advances in Computational Terminology,* John Benjamins, Ámsterdam, 2001, vol. 2, págs. 53-87.

CABRÉ, Mª. TERESA, *Terminology: Theory, methods and applications*, J. Benjamins, Ámsterdam, 1999.

CAMPILLOS LLANOS, LEONARDO; MORENO SANDOVAL, ANTONIO; GUIRAO MIRAS, JOSÉ M.ª, "An automatic term extractor for biomedical terms in Spanish", en *Proc. of the 5th International Symposium on Languages in Biology and Medicine (LBM 2013)*, 12 y 13 de diciembre de 2013, Tokio, Japón.

COHEN, KEVIN B., "Biomedical Text Mining", en NITIN INDURKHYA y FRED J. DAMERAU (eds.), *Handbook of natural language processing,* 2ª ed., Chapman and Hall/CRC, Boca Raton, 2010, págs. 605-625.

CORTÉS GABAUDÁN, FRANCISCO (coord.), *Dicciomed. Diccionario médico-biológico, histórico y etimológico,* 2007-2013. <http://dicciomed.eusal.es> [11/09/2015]

DAGAN, IDO; CHURCH, KEN, "TERMIGHT: Identifying and Translating Technical Terminology", en *4th Conference on Applied Natural Language Processing*, págs. 34-40, 1994.

DORLAND, *Diccionario enciclopédico ilustrado de medicina Dorland*, 30ª ed., Elsevier, Madrid, 2005; y 32ª versión en línea, 2012, <https://dorlandsonline.com> [11/09/2015].

DUNNING, TED, "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics*, 19(1) (Cambridge, MA, 1993), págs. 61-74.

ESTOPÀ, ROSA; VIVALDI, JORGE; CABRÉ, Mª. TERESA, "Use of Greek and Latin forms for term detection", en *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, Atenas, Grecia, 31 de mayo-2 de junio del 2000.

GONZALO SANZ, LUIS Mª. (coord.), *Diccionario Espasa Medicina*, Espasa S.L., Madrid, 1999.

HERRERO ZORITA, CARLOS; MOLINA, CLARA; MORENO SANDOVAL, ANTONIO, "Medical term formation in English and Japanese: A study of the suffixes –gram, -graph and –graphy", *Review of Cognitive Linguistics*, 13(1) (Ámsterdam, 2015), págs. 81-101.

INDURKHYA, NITIN; DAMERAU, FRED J. (eds.), *Handbook of natural language processing*, 2ª Ed., Chapman and Hall/CRC, Boca Raton, 2010.

JIMÉNEZ, Mª. ELENA, "Afijos grecolatinos y de otra procedencia en terminos medicos", *MEDISAN*, 16(6) (Santiago de Cuba, 2012), págs. 1005-1021.

JURAFSKY, DAN; MARTIN, JAMES, *Speech and Language Processing*, 2ª ed., Prentice Hall Series in Artificial Intelligence, New Jersey, 2009.

JUSTESON, JOHN S.; KATZ, SLAVA M., "Technical terminology: some linguistic properties and an algorithm for identification in text", *Natural Language Engineering*, 1(1) (Cambridge, 1995), págs. 9-27.

KAGEURA, KYO; UMINO, BIN, "Methods of automatic term recognition: A review", *Terminology*, 3(2) (Ámsterdam, 1996), págs. 259-289.

KILGARRIFF, ADAM, "Which words are particularly characteristic of a text? A survey of statistical approaches", en *Proc. AISB Workshop on Language Engineering for Document Analysis and Recognition,* Sussex University, abril de 1996, págs. 33-40.

KILGARRIFF, ADAM; RYCHLY, Pavel; SMRZ, Pavel; TUGWELL, DAVID, "The Sketch Engine", en *Proceedings of EURALEX 2004*, Lorient, France, 2004, págs. 105-116.

KIT, CHUNYU; LIU, XIAOYUE, "Measuring mono-word termhood by rank difference via corpus comparison", *Terminology*, 14(2) (Ámsterdam, 2008), págs. 204-229.

KOZA, WALTER; SOLANA, ZULEMA; CONRADO, MERLEY DA S.; REZENDE, SOLANGE O.; PARDO, THIAGO A.; DÍAZ-LABRADOR, JOSUKA; ABAITUA, JOSEBA, "Extracción terminological en el dominio médico a partir del reconocimiento de sintagmas nominales", *INFOSUR*, 5 (Rosario, Argentina, 2011), págs. 27-40.

KRAUTHAMMER, MICHAEL; NENADIC, GORAN, "Term identification in the biomedical literature", *Journal of Biomedical Informatics*, 37 (Ámsterdam, 2004), págs. 512-526.

LÓPEZ PIÑERO, JOSÉ M.ª; TERRADA FERRANDIS, M.ª LUZ, *Introducción a la terminología médica*, Masson, Barcelona, 2005.

MORENO SANDOVAL, ANTONIO; GUIRAO MIRAS, JOSÉ M.ª, "Frecuencia y distintividad en el uso lingüístico: casos tomados de la lematización verbal de corpus de distintos registros", en *Actas del I Congreso Intnal. de Lingüística de Corpus (CILC-09)*, Universidad de Murcia, Murcia, 2009.

MORENO SANDOVAL, ANTONIO; GUIRAO MIRAS, JOSÉ M.ª, "Morpho-syntactic Tagging of

the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation", en YUJI KAWAGUCHI, SUSUMU ZAIMA Y TOSIHIRO TAKAGAKI (eds.), *Spoken Language Corpus and Linguistic Informatics,* John Benjamins, Ámsterdam, 2006, págs. 199-218.

MORENO SANDOVAL, ANTONIO; CAMPILLOS LLANOS, LEONARDO, "Design and annotation of MultiMedica - a multilingual text corpus of the biomedical domain", en C. VARGAS-SIERRA (ed.), *Procedia - Social and Behavioral Sciences*, 95, Elsevier, Berlín, 2013, págs. 33-39.

NAKAGAWA, HIROSHI; MORI, TATSUNORI, "Automatic term recognition based on statistics of compound nouns and their components", *Terminology*, 9(2) (Ámsterdam, 2003), págs. 201-219.

NAZAR, ROGELIO; CABRÉ, M.ª TERESA, "Un experimento de extracción de terminología utilizando algoritmos estadísticos supervisados", *Debate Terminológico*, 7 (2010), págs. 36-55.

NAZAR, ROGELIO; VIVALDI, JORGE; WANNER, LEO, "Automatic taxonomy extraction for specialized domains using distributional semantics", *Terminology*, 18(1) (Ámsterdam, 2012), págs. 188-225.

NENADIC, GORAN; SPASIC, IRENA; ANANIADOU, SOPHIA, "Terminology-driven mining of biomedical literature", *Bioinformatics*, 19(8) (Oxford, 2003), págs. 938-943.

ORGANIZACIÓN MUNDIAL DE LA SALUD (OMS), "The use of stems in the selection of International Nonproprietary Names (INN) for pharmaceutical substances", 2011a, <http://apps.who.int/medicinedocs/documents/s19117en/s19117en.pdf> [11/09/2015]

ORGANIZACIÓN MUNDIAL DE LA SALUD (OMS), "International Nonproprietary Names (INN) for biological and biotechnological substances (a review)", 2011b, <http://apps.who.int/medicinedocs/documents/s19119en/s19119en.pdf> [11/09/2015]

PÉREZ-DE-VIÑASPRE, OLATZ; ORONOZ, MAITE; AGIRREZABAL, MANEX; LERSUNDI, MIKEL, "A Finite-State Approach to Translate SNOMED CT Terms into Basque Using Medical Prefixes and Suffixes", en *Proceedings of the 11th International Conference on Finite State Methods and Natural Language Processing,* St Andrews-Scotland, 15-17 de julio de 2013, págs. 99-103.

REAL ACADEMIA NACIONAL DE MEDICINA, *Diccionario de términos médicos,* Editorial Médica Panamericana, Madrid, 2011.

SÁNCHEZ GONZÁLEZ, MIGUEL A., *Historia de la medicina y humanidades médicas,* 2ªed., Elsevier/Masson, Barcelona, 2012.

SEGURA-BEDMAR, ISABEL; MARTÍNEZ FERNÁNDEZ, PALOMA; SAMY, DOAA, "Detección de fármacos genéricos en textos biomédicos", *Procesamiento del Lenguaje Natural*, 40 (Jaén, 2008), págs. 27-34.

VIVALDI, JORGE, *Extracción de candidatos a término mediante la combinación de estrategias heterogéneas,* Tesis doctoral, Universidad Politécnica de Cataluña, 2001.