ET1: Catálogo de los recursos lingüísticos disponibles en el dominio médico en lengua española y/o lenguas cooficiales

Plan de impulso de las Tecnologías del Lenguaje

Àlex Bravo, Horacio Saggion y Pablo Accuosto

07 2018

GOBIERNO DE ESPAÑA
MINISTERIO DE ECONOMÍA Y EMPRESA
SECRETARÍA DE ESTADO PARA EL AVANCE DIGITAL
ontsi observatorio nacional de las telecomunicaciones y de la SI
red.es
Plan TL
Plan de Impulso de las Tecnologías del Lenguaje

Este estudio ha sido realizado dentro del ámbito del Plan de Impulso de las Tecnologías del Lenguaje con financiación de la Secretaría de Estado para el Avance Digital y Red.es, que no comparten necesariamente los contenidos expresados en el mismo. Dichos contenidos son responsabilidad exclusiva de sus autores.

ÍNDICE

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

Estudio de viabilidad de una versión en español del sistema
Entregable 1: Catálogo de Recursos

El objetivo de este entregable es adquirir y documentar el estado de la cuestión en lo que respecta a los recursos léxicos para el análisis y la interpretación de lenguaje médico en español. Para ello, se va a documentar los principales recursos relacionados con el procesamiento de textos biomédicos en español como: repositorios de revistas científicas, corpora y ontologías en el dominio biomédico, herramientas de Procesamiento del Lenguaje Natural (PLN) en textos médicos en español y un repaso de los trabajos más destacables sobre el PLN en biomedicina relacionado con el dominio que vamos a trabajar:

- Construcción de corpora multilingüe en el campo biomédico.
- Creación de glosarios y ontologías multilingüe en el campo biomédico.
- Extensión de cobertura del Unified Medical Language System (UMLS) a otras lenguas que el inglés.
- Desarrollo de sistemas de traducción automática adaptados en el campo biomédico.

1. INTRODUCCIÓN

1.1 LA LITERATURA CIENTÍFICA ESPAÑOLA EN EL CAMPO BIOMÉDICO

Esta sección va a documentar los recursos disponibles relacionados con la literatura científica española centrada en el campo biomédico [1]. Las revistas científicas han sido el principal medio de difusión científica de cualquier investigación. A día de hoy, gracias a su adaptación a la era digital, las revistas científicas continúan siendo el principal recurso de difusión científica ofreciendo, además de los formatos impresos, formatos electrónicos fácilmente accesibles a través de internet.

Como ya se lleva comentando desde hace mucho tiempo, la frecuencia de publicaciones científicas no ha hecho más que aumentar, sobretodo a partir de la digitalización de las mismas. Esto ha hecho emerger un gran problema, los científicos se encuentran abrumados por una inmensa cantidad de información científica disponible en Internet. En Figura 1 se puede apreciar el incremento exponencial que está sufriendo la literatura científica biomédica. El gráfico se ha realizado a partir de las estadísticas dadas por Pubmed[1] de todas las publicaciones que se han indexado en la base de datos de Medline hasta el 2017, que a nivel internacional en el área de la biomedicina, es la base de datos de referencia en revistas científicas.

[1] *https://www.nlm.nih.gov/bsd/licensee/baselinestats.html*

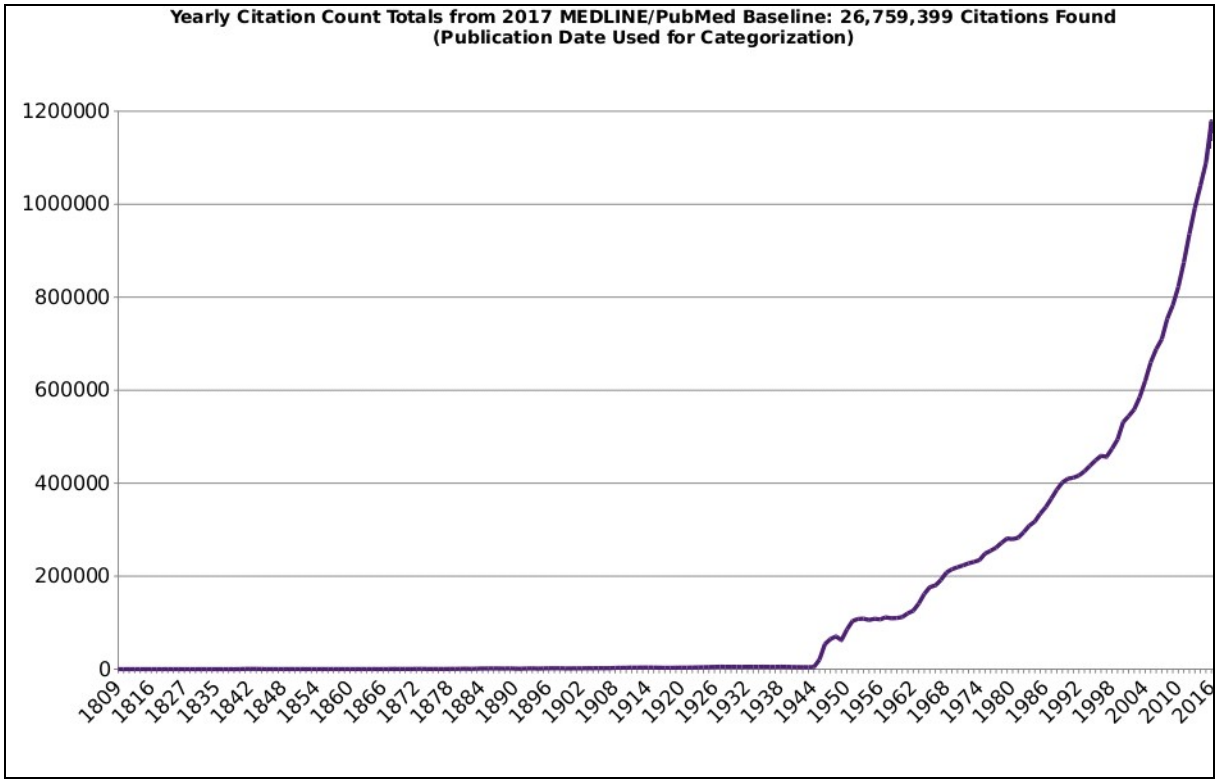


Figura 1: Número total de publicaciones indexadas por año.

Actualmente, el Inglés se ha consolidado como el idioma principal de la ciencia. Por este motivo, la gran mayoría de publicaciones se encuentran en este idioma, junto con una gran variedad de recursos como: como base de datos, ontologías, revistas o herramientas de procesado de texto.

Desde el punto de vista de la lengua española, nos encontramos que su participación en la divulgación científica es mínima, al igual que otras lenguas como francés y alemán. Fuera del ámbito científico, el español se ha establecido com una lengua muy utilizada:

- Es la segunda lengua materna más hablada en el mundo (después del chino mandarín).
- El segundo idioma de comunicación internacional (después del inglés).
- Es el tercer idioma más utilizado en internet[2].
- Si no cambia la tendencia, dentro de 3 o 4 generaciones el 10% de la población mundial se comunicará en español[3].

[2] *https://www.internetworldstats.com/stats7.htm*

[3] *http://www.cervantes.es/imagenes/File/prensa/EspanolLenguaViva16.pdf*

Aunque los investigadores españoles publican gran parte de su producción científica en revistas internacionales en inglés encontramos también algunas publicaciones en español, convirtiéndose en una valiosa información que debe ser recuperada y procesada para hacerla accesible a la comunidad científica internacional.

Estas producciones científicas se publican en un número muy reducido de revistas relevantes en el dominio biomédico iberoamericano. Para conocer las revistas biomédicas más relevante en nuestra lengua se van a comentar los resultados del trabajo de Bonfill et al. (2015) donde se presenta un estudio para identificar las principales revistas españolas dependiendo del país de procedencia [2]. Primeramente, identificaron un total de 2,457 revistas encontradas en el campo biomédico en español. Después, fueron seleccionadas las más relevantes mediante la aplicación de una serie de filtros como identificar la revista con un ISSN y la obtención de un impact factor acorde con el Journal Citation Reports. Finalmente un conjunto de 45 revistas fueron encontradas de las cuales, la mayoría procedían de España (26, 57.7%, ver ANEXO I), seguido por Argentina (5, 11.1%), Chile (4, 8.9%), Colombia (3, 6.7%), Mexico (3, 6.7%) and Venezuela (4, 8.9%).

1.1.1 Base de datos de revistas biomédicas

En esta sección vamos a describir las bases de datos más importantes de revistas biomédicas en español. Como hemos comentado anteriormente, a nivel internacional y en el área de la biomedicina, la base de datos de referencia en revistas científicas es Medline, elaborada por la National Library of Medicine de Estados Unidos. Como consecuencia de que el inglés es el idioma principal de divulgación científica, casi la totalidad de revistas incluidas en Medline recogen publicaciones en inglés. Muy escasamente podemos encontrar algunos artículos en otros idiomas como el español[4], por ejemplo, en el período 2005-2010, Medline recogió un porcentaje de publicaciones en español por debajo del 1% (similar al de otros idiomas como el alemán o el francés). Por ese motivo, Medline se convierte en un recurso muy pobre a la hora de obtener publicaciones científicas en español.

Afortunadamente, en España encontramos principalmente dos bases de datos activas (IBECS y MEDES) que recogen producción científica española en el área biomédica. También tenemos una base de datos muy importante inactiva (IME), pero que mantiene toda su información accesible desde 1971. Estas bases de datos que recogen publicaciones biomédicas en español son descritas a continuación:

[4] *https://www.ncbi.nlm.nih.gov/pubmed?term=%22spanish%22%5BLanguage%5D*

- IBECS[5] (Índice Bibliográfico Español de Ciencias de la Salud): es una base de datos bibliográfica mantenida por la Biblioteca Nacional de Ciencias de la Salud (BNCS, National Health Sciences Library) of the Instituto de Salud Carlos III (Carlos III Health Institute) que coleccionan revistas científicas cubriendo múltiples campos en ciencias de la salud publicadas en españa desde el año 2000. Actualmente incluye unos 170.000 registros y aumenta en 12.000 registros por año. IBECS también incluye unos 30.000 enlaces a artículos completos a través del nodo Scielo Spain.

- MEDES[6] (MEDicina en ESpañol): es una iniciativa de la Fundación Lilly que tiene como objetivo promover la utilización del español como lengua para la transmisión del conocimiento científico en general y de las Ciencias de la Salud en particular, entendiendo este propósito no solo orientado a la comunicación entre científicos y profesionales de la salud, sino también a la divulgación social del conocimiento entre todos los hispanohablantes.

- Por otro lado, IME (Índice Médico Español) es otra base de datos de revistas que lamentablemente dejó de actualizarse en 2012, pero mantiene la información de búsquedas de su base de datos, con una cobertura histórica desde 1971.

Aunque las anteriores bases de datos son las principales revistas en producción científica en España, podemos encontrar otras menos conocidas como:

- CUIDEN[7] es una Base de Datos Bibliográfica de la Fundación Index que incluye producción científica sobre Cuidados de Salud en el espacio científico Iberoamericano, tanto de contenido clínico-asistencial en todas sus especialidades y de promoción de la salud, como con enfoques metodológicos, históricos, sociales o culturales. Contiene artículos de revistas científicas, libros, monografías y otros documentos, incluso materiales no publicados, cuyos contenidos han sido evaluados previamente por un comité de expertos.

- ENFISPO[8] catálogo de artículos de la selección de revistas en español que se reciben en la Biblioteca de la Facultad de Enfermería, Fisioterapia y Podología de la Universidad Complutense de Madrid.

[5] *http://ibecs.isciii.es*

[6] *https://www.medes.com*

[7] *http://www.index-f.com/new/cuiden/*

[8] *http://alfama.sim.ucm.es/isishtm/enfispo/*

- Compludoc[9] contiene referencias de artículos publicados en más de 4.000 revistas científicas españolas y extranjeras recibidas en la biblioteca de la Universidad Complutense. El sistema permite seleccionar la búsqueda en revistas en español o en otros idiomas. El problema lo tenemos como en IME, desde el mes de marzo de 2012 Compludoc ha dejado de incorporar nuevas referencias, aún así seguirá estando disponible para su consulta.

1.1.2 Herramientas de acceso a los artículos biomédicos

Una vez visto las bases de datos que nos proporcionan artículos científicos en español, esta sección se va a centrar en las herramientas disponibles para acceder a estos artículos de manera óptima.

En la actualidad, las vías de acceso a la información científica son muy variadas y no siempre el investigador accede directamente a la página web en la que se encuentra alojada la información.

Algunas veces se accede directamente a una publicación o a una base de datos y se realiza una búsqueda de información a través de las herramientas que facilita la propia revista o base de datos, pero cada vez más frecuentemente, el acceso se realiza a través de bibliotecas virtuales, portales de revistas científicas, repositorios, buscadores generales o especializados, o incluso a través de redes sociales generales o de investigación.

1.1.2.1 Búsqueda de literatura científica en español

Los buscadores de literatura científica son herramientas que nos ayudan a encontrar información relevante dependiendo de nuestros criterios de búsqueda sobre una gran variedad de recursos, principalmente de revistas científicas.

A nivel internacional, los más utilizados serían PubMed Central (NLM) [10], BioMed Central (BioMed Central/Springer Nature) [11] y ScienceDirect (Elsevier) [12]. En el dominio de la lengua española incluyendo la extensión iberoamericana, podemos encontrar los siguientes portales:

[9] *http://europa.sim. ucm.es/compludoc/*

[10] *http://www.ncbi.nlm.nih.gov/pmc/*

[11] *https://www.biomedcentral.com/*

[12] *http://www.sciencedirect.com/*

- SciELO / SciELO España[13] [14] : Scielo España es una biblioteca virtual formada por una colección de revistas científicas españolas de ciencias de la salud seleccionadas de acuerdo a unos criterios de calidad preestablecidos.

- Redalyc[15]: Es un proyecto académico para la difusión en Acceso Abierto de la actividad científica editorial que se produce en y sobre Iberoamérica.

- Dialnet[16]: Dialnet es uno de los mayores portales bibliográficos del mundo, cuyo principal cometido es dar mayor visibilidad a la literatura científica hispana.

- Redib[17]: es una plataforma de agregación de contenidos científicos y académicos en formato electrónico producidos en el ámbito iberoamericano, relacionados con él en un sentido cultural y social más amplio y geográficamente no restrictivo.

Exclusivamente en España encontramos principalmente 3 buscadores de este tipo:

- Plataforma de Gestión del Conocimiento del Sistema Nacional de Salud del MSSSI[18]: El Ministerio de Sanidad, Servicios Sociales e Igualdad (MSSSI), como ente vertebrador y de cohesión del Sistema Nacional de Salud (SNS), se ha planteado habilitar, promover y facilitar a los profesionales sanitarios del SNS el acceso a una plataforma de Gestión del Conocimiento que permite realizar búsquedas simultáneas en bases de datos, revistas científicas, repositorios locales, buscadores web y catálogos de bibliotecas de Ciencias de la Salud. Esta iniciativa permite a los profesionales obtener más y mejores resultados que les servirán de apoyo en la toma de decisiones en el campo asistencial, docente y de investigación.

- BVS España[19]: Desde 1999, la Biblioteca Nacional de Ciencias de la Salud (BNCS) del Instituto de Salud Carlos III (ISCIII), consciente de la importancia para la comunidad científica e investigadora de nuestro país de la difusión de la producción científica, asume el papel de

[13] *http://scielo.isciii.es*

[14] *http://www.scielo.org/php/index.php?lang=es*

[15] *http://www.redalyc.org*

[16] *https://dialnet.unirioja.es/*

[17] *https://www.redib.org*

[18] *http://msssi.hosted.exlibrisgroup.com/primo_library/libweb/action/search.do?vid=34MDS_V1*

[19] *http://bvsalud.isciii.es*

Centro Coordinador del proyecto BVS en España, y comienza a desarrollarlo en colaboración con BIREME (OPS/OMS). BVS España permite acceder a distintas fuentes de información científica en Ciencias de la Salud que incluye bases de datos, catálogos colectivos, publicaciones electrónicas, noticias y herramientas de búsqueda.

- Biblioteca Virtual del CSIC[20]: En la Biblioteca Virtual del CSIC encontrarás todos los documentos impresos y electrónicos de la Red de Bibliotecas y Archivos del CSIC[21] y del Repositorio Digital.CSIC[22] y podrás enlazar a los textos completos o pedirlos al Servicio de Obtención de Documentos[23].

1.1.2.2 Repositorios temáticos e institucionales

Además de las bases de datos de revistas científicas, podemos encontrar repositorios que recogen producción científica de diferentes instituciones. Permiten acceso de forma abierta a los textos y permiten su reutilización de acuerdo con licencias creative commons[24]. En el ámbito de la salud en España podemos encontrar, por ejemplo, el repositorio del Sistema Sanitario Público de Andalucía[25] y repositorio de información digital del Departamento de Salud de la Generalitat de Catalunya llamado Scientia[26]. A parte de estos repositorios, existen herramientas que recogen información de diferentes repositorios institucionales, Entre ellas la más conocida es Recolecta[27], una plataforma creada por la FECYT y que agrupa a todos los repositorios científicos nacionales (83 repositorios de universidades y centros de investigación) y que provee de servicios a los gestores de repositorios, a los investigadores y a los agentes implicados en la elaboración de políticas (decisores públicos). En el ámbito internacional, OpenAIRE[28] recoge información de 781 repositorios y revistas, principalmente europeos.

[20] *http://bibliotecas.csic.es/biblioteca-virtual*

[21] *http://bibliotecas.csic.es*

[22] *http://digital.csic.es*

[23] *https://csic.gtbib.net/menu_usuario.php?centro=CSIC*

[24] *https://creativecommons.org/licenses/?lang=es_ES*

[25] *http://www.repositoriosalud.es*

[26] *http://scientiasalut.gencat.cat/*

[27] *http://recolecta.fecyt.es*

[28] *https://www.openaire.eu*

2. CORPORA MULTILINGÜE EN EL DOMINIO BIOMÉDICO

Muy recientemente, Villegas et al (2018) presentaron el conjunto de recursos MeSpEN[29] donde se describen los esfuerzos para identificar y caracterizar tipos heterogéneos de documentos y glosarios útiles para construir un corpus paralelo para traducción automática entre inglés y español en el dominio médico [3]. MeSpEN fue creado gracias al Plan de Impulso de las Tecnologías del Lenguaje de la Agenda Digital (Plan TL) [30] específicamente para el campo de la biomedicina.

Además del proyecto MeSpEN, otros esfuerzos se han llevado a cabo con el objetivo de crear corpora paralela entre español y otros idiomas en el campo biomédico. En la web del proyecto de MeSpEN nos proporciona el acceso fácil a varios corpora paralelos como el de Scielo, IBECS y Pubmed con más de 161.000, 168.000 y 127.000 títulos y resúmenes. A continuación, se describen más recursos de corpora paralelos:

- Mantra Gold Standard Corpus[31]: Es el resultado de un esfuerzo para crear un corpus multilingüe que se utilice como gold-standard en el campo del reconocimiento de entidades biomédicas. El corpus se ha creado a partir de diferentes corpora paralelos (como títulos y resúmenes de Medline, etiquetas de fármacos y patentes en el campo biomédico) en inglés, francés, alemán, español y holandés. El corpus fue manualmente anotado por tres expertos, que independientemente anotaron entidades biomédicas basadas en un subconjunto del UMLS y cubriendo una gran parte de los grupos semánticos. Finalmente el gold-standard incluye un total de 5.530 anotaciones con un Inter-annotator agreement de 0.79 de F-score.

- IULA[32]: Este corpus recoge textos escritos en cinco lenguas diferentes (català, castellano, inglés, francés y alemán) dentro de los dominios de especialidad de la economía, el derecho, el medio ambiente, la medicina, la informática y las ciencias del lenguaje. Este corpus ha sido desarrollado para la investigación de la detección de neologismos y términos, estudios sobre variación lingüística, análisis sintáctico parcial, alineación de textos, extracción de datos para la enseñanza de segundas lenguas y para la construcción de diccionarios electrónicos y elaboración de tesaurus.

[29] *http://temu.bsc.es/mespen/*

[30] *http://www.agendadigital.gob.es/tecnologias-lenguaje/Paginas/plan-impulso-tecnologias-lenguaje.aspx*

[31] *https://biosemantics.org/index.php/resources/mantra-gsc*

[32] *https://www.upf.edu/es/web/iula/corpus*

- MedlinePlus: Creado bajo el proyecto MeSpEN, incluye un total de 7.033 artículos en inglés y español procedentes de los campos de información que proporciona MedlinePlus: "Temas de salud" [33], "Medicinas, hierbas y suplementos" [34], "Información sobre pruebas de laboratorio" [35] y "Enciclopedia médica" [36].

- MedlinePlus - Temas de salud: Creado bajo el proyecto MeSpEN, 1.063 artículos en inglés y español procedentes únicamente centrados en la sección "Temas de salud" de MedlinePlus.

- Glosarios médicos: Creado bajo el proyecto MeSpEN, este dataset incluye 46 glosarios bilingües para varios pares de idiomas procedentes de glosarios médicos libres y diccionarios creados por más de 500 traductores profesionales. Estos glosarios se componen en: 26 glosarios con términos en inglés, 8 glosarios con términos en español y los restantes 13 glosarios incluyen diferentes idiomas.

- UFAL Medical Corpus[37]: es una colección de corpus paralelos creados durante el curso de los proyectos KConnect, Khresmoi y HimL con el objetivo de una traducción automática más confiable de textos médicos. El corpus ha sido creado mediante fuentes de dominio médico y fuentes fuera del dominio médico.

- EMEA[38]: es un corpus de documentos biomédicos recuperados de la European Medicines Agency (EMEA) [5]. El corpus incluye documentos relacionados con productos médicos y sus traducciones a los 22 idiomas oficiales de la Unión Europea. Concretamente está formado por más de 26 millones de frases repartidas en cerca de 42.000 documentos.

- COPPA[39]: Es un corpus compuesto por patentes publicadas en diferentes idiomas con la intención de fomentar y estimular la investigación en traducción automática y procesamiento de texto en patentes. COPPA incluye patentes de campos muy diversos, incluido patentes relacionadas con el campo biomédico. Las patentes pueden ser clasificadas por su

[33] *https://medlineplus.gov/spanish/healthtopics.html*

[34] *https://medlineplus.gov/spanish/druginformation.html*

[35] *https://medlineplus.gov/spanish/labtests.html*

[36] *https://medlineplus.gov/spanish/encyclopedia.html*

[37] *https://ufal.mff.cuni.cz/ufal_medical_corpus*

[38] *http://opus.nlpl.eu/EMEA.php*

[39] *http://www.wipo.int/patentscope/en/data/#coppa*

International Patent Classification (IPC), una categoría internacional para clasificar patentes. Dušek et al 2014 desarrollaron un sistema de MT para texto médico donde su dataset entre otros corpora se basó en COPPA [6]. Utilizaron el IPC para seleccionar las patentes relacionadas con el campo biomédico. Concretamente los siguientes códigos del IPC:

- A61*[40]: MEDICAL OR VETERINARY SCIENCE; HYGIENE.
- C12N[41]: Micro-organisms or enzymes; Compositions thereof; Propagating, preserving, or maintaining micro-organisms; Mutation or genetic engineering; Culture media.
- C12P41: Fermentation or enzyme-using processes to synthesise a desired chemical compound or composition or to separate optical isomers from a racemic mixture.

Utilizando estos códigos, se han seleccionado las patentes biomédicas del corpus COPPA y en la Tabla 1 se muestran las estadísticas sobre el número de documentos para español e inglés en el campo biomédico:

Idioma	A61*	C12P	C12N	TOTAL
Español	3.223	167	632	3.723
Inglés	413.940	18.059	70.084	467.422

Tabla 1: Documentos de patentes biomédicas en COPPA para inglés y español

- Khresmoi Summary Translation Test Data 2.0[42]: Este corpus incluye 1500 frases en diferentes idiomas (como inglés, francés, español o polaco) divididas en un set de desarrollo y otro de test para la evaluación de técnicas de traducción automáticas.

[40] *www.wipo.int/ipc/itos4ipc/ITSupport_and_download_area/20100101/pdf/scheme/core/en/ipcr_en_a_core_2010.pdf*

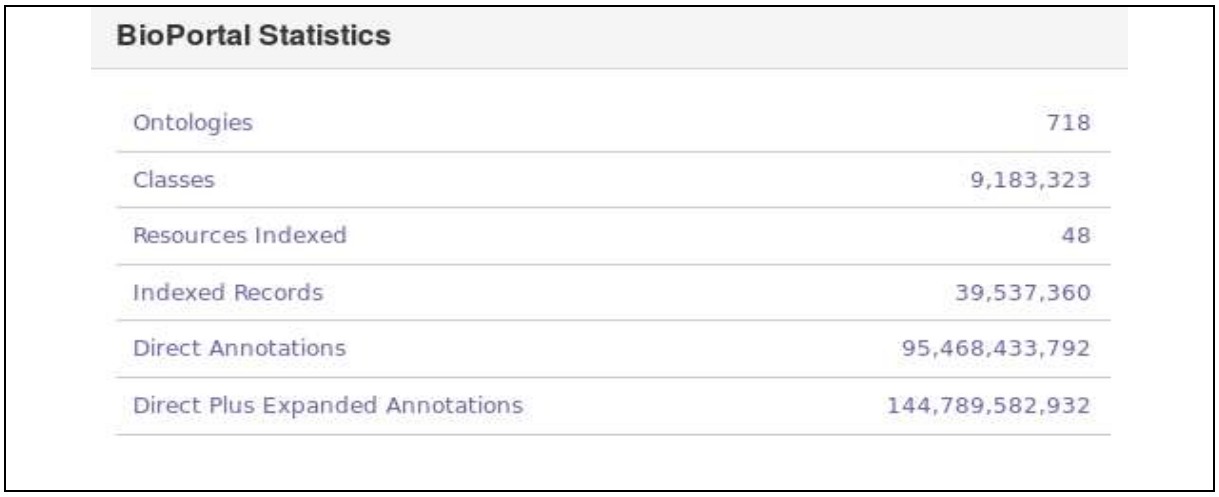
[41] *www.wipo.int/ipc/itos4ipc/ITSupport_and_download_area/20100101/pdf/scheme/core/en/ipcr_en_c_core_2010.pdf*

[42] *https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2122*

3. ONTOLOGÍAS EN EL DOMINIO BIOMÉDICO

En esta sección vamos a describir las principales ontologías en el dominio biomédico tanto en inglés como en español.

- BioPortal[43]: desarrollado bajo el proyecto National Center for Biomedical Ontology (NCBO), es el mayor repositorio de ontologías biomédicas, actualmente alberga más de 700 ontologías, vocabularios controlados y terminologías. BioPortal permite acceder a su información mediante una interfaz web, y programáticamente mediante Web Services. La siguiente tabla muestra unas estadísticas generales del contenido de esta ontología:



BioPortal Statistics

Ontologies	718
Classes	9,183,323
Resources Indexed	48
Indexed Records	39,537,360
Direct Annotations	95,468,433,792
Direct Plus Expanded Annotations	144,789,582,932

Figura 2: Número total de publicaciones indexadas por año.

- DBpedia[44]: es una enorme base de conocimiento (knowledge base) desarrollada a partir de un esfuerzo comunitario para extraer información de Wikipedia y representarla en un formato estructurado adecuado para el procesamiento automático. DBpedia es uno de los ejes centrales de Linked Life Data (ver abajo).

- Linked Life Data[45]: proporciona acceso a una enorme base de conocimiento que incluye y semánticamente relaciona conocimiento sobre genes, proteínas, interacciones moleculares,

[43] *https://bioportal.bioontology.org/*

[44] *https://wiki.dbpedia.org/*

[45] *http://linkedlifedata.com/*

pathways, fármacos, enfermedades, ensayos clínicos y otros tipos relacionados con entidades biomédicas. Es una parte de la inmensa Linked Open Data Cloud[46].

- NCBI BioSystems Database[47]: es un repositorio que proporciona acceso integrado a conocimiento y datos estructurados sobre sistemas biológicos y sus componentes: genes, proteínas y pequeñas moléculas.

- Open Biomedical Ontologies[48]: es el resultado del esfuerzo de una comunidad de desarrolladores de ontologías en crear lenguajes de indización para su uso compartido a través de diferentes dominios biológicos y médicos.

- Chemical Entities of Biomedical Interest (ChEBI) [49]: es una base de datos que es parte del esfuerzo de el Open Biomedical Ontologies, que describe la ontología de las entidades moleculares centrada en pequeños componentes químicos. El término de entidades moleculares son cualquier producto natural o de síntesis usado para intervenir en el proceso de organismos vivos. Las moléculas directamente relacionadas con el genoma, tales como los ácidos nucleicos, las proteínas y los péptidos, no están incluidos entre las ChEBI.

- Uniprot[50]: es un repositorio de secuencia de proteínas e información funcional, que contiene entradas manualmente anotadas. Las entradas son curadas por biólogos, actualizadas regularmente y enlazadas con numerosas bases de datos externas.

- SNOMED CT[51]: es considerada la terminología clínica integral, multilingüe y codificada de mayor amplitud, precisión e importancia desarrollada en el mundo. Snomed CT es un producto que nace de la fusión entre Snomed RT (Snomed Reference Terminology), creada por el College of American Pathologists (CAP) y el Clinical Terms Version 3 (CTV3), desarrollada por la National Health Service (NHS) del Reino Unido. Snomed-CT se está convirtiendo en un estándar

[46] *http://lod-cloud.net/*

[47] *https://www.ncbi.nlm.nih.gov/biosystems/*

[48] *http://www.obofoundry.org/*

[49] *https://www.ebi.ac.uk/chebi/*

[50] *https://www.uniprot.org/*

[51] *https://www.snomed.org/snomed-ct*

con el apoyo de muchos gobiernos incluido el de España, y es una de las ontologías más utilizadas para la normalización de términos biomédicos.

- Medical Dictionary for Regulatory Activities (MedDRA) [52]: contiene terminología médica en diferentes idiomas relacionadas con todas las fases de desarrollo de fármacos (excluyendo toxicología animal), los efectos terapeúticos y los dispositivos multifunción. Concretamente, el ámbito de utilización de MedDRA cubre tanto productos farmacéuticos, como biológicos, vacunas o combinaciones de fármacos con dispositivos médicos. MedDRA se ha traducido a las siguientes lenguas: Chino, Checo, Holandés, Francés, Alemán, Húngaro, Italiano, Portugués y Español. MedDRA Spanish es también conocida como MDRSPA.

- Medical Subject Headings (MeSH) [53]: es un vocabulario controlado para clasificar, indexar, categorizar y buscar artículos de más de 5400 revistas en el campo biomédico contenidas en la base de datos MEDLINE.

- WHO Adverse Drug Reaction (WHO-ART) [54]: Es una terminología estructurada a nivel jerárquico de 4 niveles que codifica la información clínica relacionada con las reacciones adversas producidas por fármacos.

- UMLS[55]: es la fuente de conocimiento biomédico más conocida y utilizada en el sector. Su principal objetivo es asignar un identificador único (CUI) a cada concepto biomédico y relacionarlos formando una estructura de red. Sus conceptos biomédicos vienen de una gran cantidad de terminologías y vocabularios, las cuales fueron manualmente integradas y distribuidas por la United States National Library of Medicine. El UMLS está formado principalmente por el Metathesaurus, la Red Semántica (Semantic Network) y el léxico especializado y herramientas léxicas (SPECIALIST Lexicon). Como hemos comentado, el UMLS integra las ontologías, terminológicas y vocabularios más importantes en el campo biomédico. Obviamente, la gran mayoría de fuentes están en inglés, pero UMLS también incluye terminologías en otros idiomas como el español y el vasco, concretamente son las siguientes:

    - CPTSP (Current Procedural Terminology Spanish)

[52] *https://www.meddra.org*

[53] *http://www.nlm.nih.gov/mesh*

[54] */https://www.who-umc.org/vigibase/services/learn-more-about-who-art/*

[55] *https://www.nlm.nih.gov/research/umls/*

  - CPCSPA (ICPC Spanish)
  - LNC-ES-AR (LOINC Linguistic Variant - Spanish, Argentina)
  - LNC-ES-CH (LOINC Linguistic Variant - Spanish, Switzerland)
  - LNC-ES-ES (LOINC Linguistic Variant - Spanish, Spain)
  - MDRSPA (MedDRA Spanish)
  - MSHSPA (MeSH Spanish)
  - SCTSPA (SNOMED CT Spanish Edition)
  - WHOSPA (WHOART Spanish)
  - ICPCBAQ (ICPC Basque)

- International Classification of Diseases (ICD) [56]: es un sistema de clasificación propuesto por los Centers for Medicare and Medicaid Service (CMS) de los Estados Unidos. El propósito de este recurso es clasificar y codificar causas orgánicas, condiciones externas, problemas y circunstancias que influyen en el estado de la salud y el contacto con los servicios médicos y las modalidades de prestación de servicios de salud como hospitalización y servicios ambulatorios. Está traducido al español con la convención CIE-10[57] (o ICD-10-ES).

- Bot PLUS[58]: es una aplicación informática elaborada por el Consejo General de Colegios Oficiales de Farmacéuticos, para la consulta de información homogénea y actualizada relativa a medicamentos, productos de parafarmacia, enfermedades e interacciones, así como para facilitar el ejercicio de la Atención Farmacéutica, en el ámbito de la farmacia comunitaria. Incluye una terminología de fármacos clínicos en español.

- CIMA[59]: es un centro de información online mantenido por la Agencia Española de Medicamentos y Productos Sanitarios (AEMPS). CIMA proporciona información de todos los

[56] *http://www.who.int/classifications/icd/en*

[57] *https://www.msssi.gob.es/estadEstudios/estadisticas/normalizacion/CIE10/home.htm*

[58] *http://www.portalfarma.com/inicio/botplus20*

[59] *https://www.aemps.gob.es/cima/publico/home.html*

fármacos autorizados en España. CIMA contiene un total de más de 16.400 marcas de fármacos y más de 2.200 fármacos genéricos.

4. HERRAMIENTAS DE PLN

En esta sección vamos a describir las herramientas de PLN más utilizadas en el campo biomédico tanto en inglés como en español y otras lenguas del estado.

- FreeLing[60]: es una librería de código abierto para el procesamiento multilingüe automático, que proporciona una amplia gama de servicios de análisis lingüístico para diversos idiomas [7]. FreeLing ofrece a los desarrolladores de aplicaciones de PLN, funciones de análisis y anotación lingüística de textos. FreeLing es altamente configurable, es decir, se pueden utilizar los recursos lingüísticos por defecto (diccionarios, lexicones, gramáticas, etc) o ampliarlos/adaptarlos a dominios particulares, o incluso desarrollar otros nuevos para idiomas específicos o necesidades especiales de las aplicaciones.

- FreelingMed[61]: está basado en una extensión del analizador Freeling para analizar documentos clínicos en español, a partir de nuevos datos lingüísticos como SNOMED CT, una lista de abreviaturas médicas, Bot PLUS[62] (terminología de fármacos clínicos en español) y ICD-9 [8, 10]. Esta herramienta es capaz de reconocer términos, pero no los identifica, es decir, no realiza la tarea de desambiguación, pero para cada término reconocido, devuelve la lista de todos los identificadores relacionados. Presenta una performance de un 90% de F-score en la detección de entidades médicas en un corpus manualmente etiquetado de 100 registros médicos.

- FreelingMed[63]: es probablemente el anotador biomédico más conocido y utilizado en el sector. Ha sido desarrollado bajo la National Library of Medicine (NLM) para detectar terminología biomédica en textos e identificarla con sus correspondientes conceptos en el UMLS Metathesaurus. Además, cada anotación incluye un score que refleja la relevancia que existe entre el texto y el concepto biomédico. Como Freeling, MetaMap es también altamente configurable y varios elementos del proceso de anotación pueden ser adaptados como el vocabulario utilizado (permite trabajar con otras ontologías usando el método de Data File

[60] *http://nlp.lsi.upc.edu/freeling/index.php/node/1*

[61] *http://ixa2.si.ehu.es/freeling2openbrat/*

[62] *http://www.portalfarma.com/inicio/botplus20*

[63] *https://metamap.nlm.nih.gov/*

Builder[64]), los filtros sintácticos aplicado al texto y el matching entre text y conceptos del UMLS. Uno de los puntos fuertes de MetaMap es su minucioso enfoque lingüístico para el análisis léxico y sintáctico del texto de entrada. Por contra, su minucioso análisis lingüístico causa que MetaMap demore mucho tiempo en el procesado de texto, haciendo imposible la anotación de un gran corpus. Otra debilidad que presenta MetaMap tiene que ver con su método de desambiguación que no es capaz de lidiar eficazmente con términos ambiguos. Para más información sobre MetaMap, se recomienda la lectura del trabajo presentado por Arondon et al. 2010, en el cual publican un overview muy interesante sobre MetaMap, su perspectiva histórica y los avances recientes [9].

- NCBO Annotator[65]: es un servicio web que anota texto en inglés con terminología procedente de UMLS y BioPortal [11]. Está basado en dos fases de anotación. La primera identifica conceptos biomédicos en el texto de entrada. Concretamente utiliza la herramienta llamada MGrep [12], la cual consiguió mejores resultados que MetaMap en diferentes evaluaciones [13]. En su segunda fase, las anotaciones identificadas con su correspondiente conceptos, son extendidas utilizando ontologías biomédicas, es decir, NCBO Annotator encuentra relaciones semánticas entre conceptos y da como resultados la asociación de un concepto encontrado con múltiples conceptos semánticamente relacionados a partir de ontologías, en vez de dar el concepto que mejor encaja con el contexto. Obviamente, no realiza ningún tipo de desambiguación.

- NOBLE Coder[66]: es otro sistema open-source para anotador text biomédico en inglés. Puede ser configurado para trabajar con diferentes vocabularios [14]. Una particularidad que incluye es la creación de una terminología personalizada a través de una interfaz gráfica, permitiendo seleccionar una o más ramas de un conjunto de vocabularios, y/o filtrar vocabularios por tipos semánticos. Gracias a su algoritmo de greedy, NOBLE Cocer puede procesar eficientemente grandes cantidades de corpora. También incluye un método de desambiguación basado en reglas heurísticas que dan preferencia a los conceptos candidatos presentes en un mayor número de vocabularios o que tengan una mayor coincidencia con un término en un vocabulario.

[64] *https://metamap.nlm.nih.gov/DataFileBuilder.shtml*

[65] *https://bioportal.bioontology.org/annotator*

[66] *http://noble-tools.dbmi.pitt.edu/*

- cTakes[67]: es una herramienta muy conocida para la anotación semántica de documentos biomédicos en general, y particularmente para textos de investigación clínica [15]. Está desarrollado mediante dos frameworks para PLN muy consolidados: UIMA[68] y OpenNLP[69]. Está desarrollado modularmente, formado por un conjunto de componentes de procesado de texto que aplican técnicas basadas en reglas y aprendizaje automático. cTakes reconoce conceptos biomédicos en textos y los relaciona con su identificador UMLS (como MetaMap). Una de las carencias de cTakes es que no implementa ningún tipo de método para desambiguar entidades, pero, al ser una herramienta modular, permite la fácil integración del componente YTEX[70], el cual está centrado en el análisis y procesamiento de texto biomédico con la capacidad de realizar la desambiguación entre conceptos UMLS.

- MedTagger[71]: es una herramienta para reconocer entidades biomédicas en textos en inglés, y esta basado en tres componentes principales: i) el MedTagger para la indexación basada en diccionarios; ii) el MedTaggerIE para la extracción de inofmarcón basada en patrones y iii) MedTaggerML para el reconocimiento de entidades utilizando un sistema basado en machine learning.

- Neji[72]: es un reconocedor de entidades biomédicas basado en machine learning [16,17]. Es una herramienta open-source centrada en cuatro características clave: modularidad, alto rendimiento, rapidez y usabilidad. Neji integra características para el PLN biomédico, desde la detección de frases a la construcción de dependencias siendo compatible con los formatos de entrada y salida más populares.

- BeCAS (2013) [73]: es una aplicación web, una APi y un widget para la identificación de conceptos biomédicos [18]. BeCAS ayuda en la identificación de más de 1.200.000 conceptos biomédicos en textos y resúmenes de pubmed.

[67] *http://ctakes.apache.org/*

[68] *https://uima.apache.org/*

[69] *https://opennlp.apache.org*

[70] *http://toolfinder.chpc.utah.edu/content/ytex*

[71] *http://ohnlp.org/index.php/MedTagger*

[72] *https://omictools.com/neji-tool and http://bioinformatics.ua.pt/neji*

[73] *http://bioinformatics.ua.pt/becas/*

- CLAMP (2017) [74]: es un software integral de PLN que permite el reconocimiento y la codificación automática de información clínica en informes narrativos de pacientes en inglés [19]. Está compuesto por una serie de componentes desarrollados independientemente y que se unen formando un pipeline el cual conforma a CLAMP. Estos componentes son:

    - Sentence boundary detection: proporciona dos sistemas para delimitar frases, un sistema de machine learning usando OpenNLP y un configurable componente basado en reglas para la detección de los extremos de una frase.

    - Tokenizer: CLAMP incluye 3 tipos de tokenizadores: OpenNLP tokenizer, un separador basado en delimitadores (como un espacio) y un sistema de reglas configurable para diferentes tokenizaciones.

    - Part-of-speech tagger: es un machine learning tagger entrenado en un corpus clínico para realizar la clasificación de palabras en su categoría gramatical.

    - Section header identification: es un sistema de diccionarios para la detección e identificación de secciones. El diccionario ha sido construido a partir de las secciones de documentos clínicos.

    - Abbreviation reorganization and disambiguation: CLAMP tamién implementa un sistema de reconocimiento y desambiguación de abreviaturas clínicas. Adicionalmente, los usuarios pueden especificar sus propias abreviaturas.

    - Named entity recognizer: CLAMP presenta 3 NERs: (1) un sistema de machine learning que utiliza un algoritmo de conditional random fields (CRF) del paquete CRFSuite library, (2) un sistema de diccionario basado en un recolección terminológica de múltiples fuentes como el UMLS and (3) un sistema de expresiones regulares con los patrones más comunes.

    - Assertion and negation: CLAMP proporciona un sistema de machine learning para la detección y clasificación de aserciones. Además incluye el algoritmo de NegEx para la detección de negaciones.

[74] *https://clamp.uth.edu/*

    - UMLS encoder: Los términos detectados pueden ser enlazados a su correspondiente identificador en el UMLS. Utiliza un sistema de rankeo basado en la similaridad entre la entidad y los conceptos candidatos utilizando modelos vectoriales.

- MOSTAS: El sistema MOSTAS pre-procesa historiales clínicos en español con el objetivo de facilitar el posterior tratamiento de los textos y recuperación de información de los mismos [20]. El sistema añade información morfo-semántica a los historiales, busca el significado de las siglas, acrónimos y abreviaturas que existen en los mismos y detecta conceptos biomédicos, utilizando para ello recursos biomédicos especializados (bases de datos, tesauros, un servidor de terminologías multilingüe en OWL, etc.). Además, MOSTAS es capaz de anonimizar y corregir los historiales clínicos.

- GALEN: propuso un MetaMap en español que combina Machine translation con el uso de MetaMap [21]. Primero procesan los textos en español con MetaMap, utilizando una base de datos personalizada que incluye términos en español e inglés del UMLS Metathesaurus. Una vez identificados los términos en español con los correspondientes CUIs, traducen de manera automática del español a inglés mediante el traductor de Google. Con esta traducción se busca en el MetaMap por defecto el correspondiente CUI en inglés, que será el CUI final que se le asignará a la mención encontrada.

- Spanish Metamap by Castro et al. (2010): Introdujo un sistema como MetaMap para documentos en español con el objetivo de obtener conceptos de SNOMED a partir de frases [22]. La normalización de términos es realizada mediante la obtención índices de Lucene en SNOMED-CT y posteriormente, se rankean según función desarrollada por los autores.

- IxaMedTagger[75]: Esta herramienta descargable permite analizar documentos médicos escritos en castellano y obtiene como resultados el reconocimiento de entidades del dominio médico, como medicamentos, alergias, estructuras corporales, calificadores y enfermedades.

- ADR2OpenBrat[76]: Esta herramienta permite analizar documentos médicos en español y obtiene como resultados el reconocimiento de entidades del dominio médico, como medicamentos (incluyendo nombre de medicamentos, sustancias y principios activos) y

[75] *http://ixa2.si.ehu.eus/ixamed-pertzeptroia/NER_Pertzeptroia.zip*

[76] *http://ixa2.si.ehu.es/adr2openbrat/*

enfermedades. Además, también reconoce relaciones adversas entre medicamentos, donde un medicamento, sustancia o principio activo produce una enfermedad.

A continuación se presenta una tabla listamos las herramientas descritas anteriormente y para qué idiomas pueden funcionar o ser adaptadas.

Herramienta	Inglés	Español (otros idiomas)
Freeling	✓	✓ Multilingüe: español, catalán y gallego (ver ANEXO II).
FreelingMed	✗	✓
MetaMap	✓	✓ Pero con adaptaciones
NCBO Annotator	✓	✗
NOBLE Coder	✓	✗
cTakes	✓	✓ Puede adaptarse a otros idiomas como el español [38].
MedTagger	✓	✗
Neji	✓	✗
CLAMP	✓	✗
MOSTAS	✗	✓
GALEN	✗	✓
Spanish MetaMap	✗	✓
IxaMedTagger	✗	✓
ADR2OpenBrat	✗	✓

Tabla 2: Listado de herramientas para el PLN en el dominio biomédico para inglés y español

5. TRABAJOS RECIENTES DESTACABLES SOBRE PLN EN TEXTO BIOMÉDICO EN ESPAÑOL

El text mining en textos clínicos y biomédicos son tareas muy populares en el campo de la investigación del PLN. Como se ha comentado anteriormente, actualmente el inglés es el idioma principal para la difusión científica, y esto ha empujado a que la gran mayoría de los trabajos relacionados con el PLN han sido desarrollados para la lengua inglesa, incluyendo los trabajos en el campo biomédico.

Para promover el procesamiento de texto clínico/biomédico y traducción automática en idiomas diferentes al inglés se han creado diferentes retos y workshops a nivel mundial. Muy recientemente ha tenido lugar el workshop MultilingualBIO (Multilingual Biomedical Text Processing)[77] durante el LREC 2018[78], del cual han surgido trabajos muy interesantes en el PLN en textos biomédicos para varios idiomas. En este entregable vamos a destacar aquellos que tienen mayor interés para nuestro objetivo.

- Cross-lingual Candidate Search for Biomedical Concept Normalization [23]: En este trabajo encaran las dificultades que existen para la normalización de terminología biomédica para textos médicos no escritos en inglés. Otras lenguas que el inglés tienen una menor cobertura de vocabularios y terminologías que dificultan la normalización de conceptos biomédicos. Para superar estas dificultades, en este trabajo proponen una búsqueda de candidatos inter-lingües para la normalización del concepto utilizando un modelo de traducción neuronal (basado en caracteres) entrenado en terminología biomédica multilingüe. Concretamente, han entrenado el modelo en las versiones española, francesa, holandesa y alemana del UMLS. Su elemento central para la normalización de conceptos biomédicos es el modelo de traducción neuronal basado en caracteres de Lee et al., 2016 [24]. Muchas traducciones de conceptos biomédicos se pueden resolver mediante pequeñas enmiendas (sufijos y prefijos), debido al origen común de muchas palabras, y este sistema puede capturar tales características. Dado un término candidato, si el término no puede ser emparejado con algún concepto multilingüe (no inglés) del UMLS, se utiliza su sistema de traducción para transformar el término al inglés y buscarlo en el UMLS de lengua inglesa. Las evaluaciones son realizadas en el caso del español por el corpus Mantra, anteriormente descrito, obteniendo una performance de 0.691 de F-score en la normalización de conceptos en español.

[77] *https://multilingualbio.bsc.es/*

[78] *http://lrec2018.lrec-conf.org/en/*

- English-Catalan Neural Machine Translation in the Biomedical Domain through the cascade approach [25]: Este trabajo trata sobre la construcción de un sistema de traducción neuronal en el campo biomédico entre inglés y catalán. Este trabajo encara la baja cantidad de fuentes en catalán implementado una estrategia en cascada utilizando el Scielo corpus Inglés-Español y la base de datos de El Periódico en Español-Catalán. Su trabajo se basa en uno de los últimos modelos neuronales de traducción llamado Transformer [26]. Crea dos modelos de traducción como hemos comentado antes y utiliza el español como lengua pivote entre el inglés y el catalán.

- KabiTermICD: Nested Term Based Translation of the ICD-10-CM into a Minor Language[79]: En este trabajo se describe el esfuerzo de traducir automáticamente la versión en inglés del ICD-10-CM (Clinical Modification) a una lengua con muy escasos recursos como el vasco. Su sistema ha sido llamado KabiTermICD y es una extensión del sistema KabiTerm que se realizó para traducir SNOMED-CT en vasco [27]. Para el análisis de la información lingüística utilizan AnaMed que también identifica teŕminos de SNOMED CT y epónimos de un texto dado. AnaMed ha sido desarrollada para inglés y vasco. La información extraída por AnaMed es utilizada para el proceso automático de traducción de KabiTermICD.

- The MeSpEN Resource for English - Spanish Medical Machine Translation and Terminologies: Census of Parallel Corpora, Glossaries and Term Translations [3]: Otro trabajo a destacar es el esfuerzo realizado en el proyecto MeSpEN para la creación de recursos que ayuden a la traducción automática de textos médicos entre inglés y español, y que ya ha sido comentado en la sección de Corpora en el dominio biomédico.

- Improving the accessibility of biomedical texts by semantic enrichment and definition expansion [28]: Este trabajo busca facilitar la comprensión de textos médicos en un corpus paralelo inglés-español mediante la anotación semántica de conceptos y la expansión automática de definiciones. Consideran la limitación de recursos disponibles para el español, y proponen explotar herramientas dirigidas al inglés para obtener anotaciones que luego se transfieren a los textos en español. Las evaluaciones realizadas muestran la viabilidad de este enfoque. Se hace público un conjunto de textos enriquecidos que se pueden recuperar, visualizar y descargar mediante una interfaz web.

[79] *https://multilingualbio.bsc.es/wp-content/uploads/2018/05/4_W3.pdf*

- Biomedical term normalization of EHRs with UMLS [29]: Este trabajo presenta un novedoso normalizador de terminología biomédica en historias clínicas con el UMLS. Aunque es un sistema multilingüe, se centrar en el procesado de texto clínico en español. La herramienta se basa en un Apache Lucene para indexar el Metathesaurus y generar candidatos de mapeo a partir del texto de entrada. Para el procesado básico del texto utiliza el pipeline IXA y para resolver ambigüedades utiliza el kit de herramientas UKB. El sistema es evaluado utilizando dos corpus paralelos inglés-español comparándolo con los resultados obtenidos por MetaMap.

- The Biomedical Abbreviation Recognition and Resolution (BARR) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to Spanish biomedical abstracts [30]: BARR[80] track es un reto propuesto en la campaña de evaluación de IberEval 2017 con el objetivo de promover el desarrollo de sistemas para extraer tanto abreviaturas como su correspondiente forma extensa a partir de resúmenes en español en el campo biomédico. Desde una enorme colección de resúmenes y títulos conteniendo un total de 155.538 registros. Los organizadores seleccionaron un subconjunto para anotar manualmente y crear un conjunto para entrenar y evaluar. Concretamente, el conjunto para entrenar contiene 1.050 resúmenes, mientras que el Gold Standard para evaluar presenta un total de 600 resúmenes. Finalmente, de manera resumida describe los 7 sistemas que participaron con buenos resultados en el BARR track.

- A machine learning approach for semantic recognition and normalization of clinical term descriptions [31]: El objetivo del trabajo es desarrollar un sistema de aprendizaje automático para el reconocimiento y normalización semántica de términos clínicos en las secciones de texto libre de las historias clínicas (descripciones). Su trabajo explora un sistema híbrido basado en métodos tradicionales de soft-matching con máquinas de aprendizaje usando los clasificadores MaxEnt and XGBoost.

- Extracting drug indications and adverse drug reactions from Spanish health social media [32]: Presentan un sistema basado en co-ocurrencias de pares de efectos producidos por fármacos como un primer paso en el estudio de la detección de efectos secundarios e indicaciones de textos de un foro médico en español. Este método es utilizado para la construcción automática de una base de datos de efectos secundarios e indicaciones de fármacos en español.

[80] *http://temu.inab.org*

- Clinical text mining for efficient extraction of drug-allergy reactions [33]: El objetivo del trabajo es procesar EHR en español para encontrar eventos relacionados con alergias derivadas de fármacos. Primero crean un corpus anotado a partir de historias clínicas del Hospital de Galdakao-Usansolo. Las anotaciones no solo se centran en encontrar alergias y los fármacos que las producen, sino también en relaciones negativas entre ellos. Utilizan FreeLing-Med como analizador adecuado para el lenguaje médico español. Experimentan con un método de anotación basado en reglas y otro de aprendizaje automático (random forest), el cual obtiene los mejores resultados de la evaluación con un 88% de F-score.

- Extracción terminológica en el dominio médico a partir del reconocimiento de sintagmas nominales [34]: En este trabajo se describen dos aproximaciones para la extracción automática de términos médicos a partir del reconocimiento automático de sintagmas nominales (SN) realizado sobre un corpus de textos médicos en español. En el primero de ellos, considerado como baseline, se extrajeron todos los SN encontrados. En el segundo experimento, en cambio, la extracción estuvo dirigida a SN específicos, que fueron determinados sobre la base de criterios sintácticos y posicionales, entre otros. La segunda aproximación obtuvo mejores resultados.

- Findings of the WMT 2017 biomedical translation shared task [35]: La segunda edición de la Biomedical Translation[81] task in the Conference of Machine Translation centrada en la traducción automática de documentos relacionados con el campo biomédico entre inglés y varias lenguas europeas como el español. En esta publicación se describen los todos los sistemas participantes en la tarea.

- Savana: A Global Information Extraction and Terminology Expansion Framework in the Medical Domain [36]: este trabajo describe un sistema para enriquecer una ontología con nuevo conocimiento, es decir, dado un término médico nuevo en la literatura, le asigna su hiperónimo más probable, constituyendo así un facilitador en tareas de enriquecimiento y expansión de bases terminológicas médicas. Concretamente se presenta Savana, un sistema de extracción de información biomédica que, combinado con validación por parte de profesionales médicos, es utilizado para popular la rama española de Snomed CT con nuevo conocimiento.

[81] *http://www.statmt.org/wmt17/biomedical-translation-task.html*

- Translating electronic health record notes from English to Spanish: A preliminary study [37]: este trabajo se centra en la traducción de historias clínicas entre el inglés y el español. Para ellos, primero construyen un corpus paralelo de historias clínicas el cual es accesible bajo demanda. Este corpus lo utilizan para entrenar un sistema de traducción automática estadístico llamado NoteAidSpanish.

6. CONCLUSIONES

En este entregable nos hemos centrado en documentar todos los recursos léxicos que existen para el procesado de texto biomédico en inglés, español y otras lenguas del estado. Hemos podido reafirmar, que aunque el español es la segunda lengua más hablada en el mundo, el inglés sigue siendo el idioma oficial de la difusión científica alcanzando millones de publicaciones en inglés. Este enorme número de publicaciones ha motivado y promovido la implementación y desarrollo de técnicas y estratégias para obtener de manera automática información estructurada de estas publicaciones. Como la gran mayoría están en inglés, lógicamente, la gran mayoría de trabajos, herramientas, recursos lingüísticos, etc.. también están en inglés.

Por otro lado, en menor medida hemos podido encontrar recursos léxicos que hay para español y otras lenguas del estado. Además, se han publicado varios estudios, herramientas y trabajos centrados en la literatura biomédica en español, incluso en otras lenguas co-oficiales del estado como el catalán y el vasco.

Esta observación nos permite afirmar que hay una buena base en el desarrollo de técnicas y estrategias en el campo del PLN en el dominio biomédico en español intentando afrontar el hecho de los pocos recursos que hay para el español en el campo biomédico mediante diferentes perspectivas. Desde la construcción de corpora, glosarios y ontologías multilingüe en el campo biomédico, extensión de cobertura terminologías en inglés como el UMLS en otro idioma como el español y el desarrollo de sistemas de traducción automática adaptados en el campo biomédico.

Con todos los recursos/trabajos descritos, podemos enumerar una serie de ideas que nos ayudarán a realizar diferentes técnicas y estrategias para lograr nuestro objetivo final, que será la extensión del sistema UMLS al español:

- A partir de un corpus paralelo de textos biomédicos (inglés-español), procesar los textos en inglés con todos los recursos y herramientas necesarios en este idioma para identificar terminología UMLS. Esto nos garantiza un mayor rango de encontrar más terminología biomédica que utilizando recursos en español. Una vez procesados los textos en inglés, la terminología encontrada puede ser alineada con su correspondiente texto en el documento en español (corpus paralelo). Una vez alineados los conceptos biomédicos encontrados se pueden codificar con diferentes características, así como traducirlos en un espacio vectorial (word embeddings), para encontrar una función de transferencia entre la terminología en inglés español. Gracias a esta función de transferencia, podríamos transferir el conocimiento

encontrado en inglés a la lengua española extendiendo así la terminología y vocabularios biomédicos en español.

- Utilizando la técnica anterior, entrenada a partir de un corpus paralelo, se podría traducir automáticamente un texto de español a inglés, procesarlo con herramientas en inglés para encontrar toda la terminología UMLS y utilizar la transferencia para encontrar las entidades normalizadas en el texto en español.

- Otra técnica sería procesar directamente los textos en español con los recursos existentes e intentar enlazar cada concepto biomédico con su correspondiente identificador UMLS. Si el término no está en las ontologías y vocabularios en español del UMLS, se puede realizar una traducción y encontrar los mejores candidatos en el UMLS. Esta traducción se puede hacer de manera literal o mediante características computadas para cada concepto encontrado, como su transformación a un espacio vectorial.

Nuestro siguiente paso, a parte de documentar todo el sistema UMLS en detalle en el próximo entregable, será escoger un corpus de documentos biomédicos en inglés y en español y procesarlos independientemente con herramientas de cada idioma, para conocer el grado de discrepancias que hay a la hora de encontrar y normalizar conceptos biomédicos.

7. REFERENCIAS

[1] Primo-Peña, E. (2016). Las bases de datos de información biomédica, ¿en español? Presente y futuro. Educación Médica, 17(Supl. 2), 39-44.

[2] Bonfill, X., Osorio, D., Posso, M., Solà, I., Rada, G., Torres, A., ... & Gandarilla, O. (2015). Identification of biomedical journals in Spain and Latin America. Health Information & Libraries Journal, 32(4), 276-286.

[3] Marta Villegas, Ander Intxaurrondo, Aitor Gonzalez-Agirre, Montserrat Marimon, Martin Krallinger. (2018). The MeSpEN Resource for English-Spanish Medical Machine Translation and Terminologies: Census of Parallel Corpora, Glossaries and Term Translations. In LREC MultilingualBIO: Multilingual Biomedical Text Processing. ELRA.

[4] Neves M, Yepes AJ, Névéol A. The scielo corpus: a parallel corpus of scientific publications for biomedicine In: Chair NCC, Choukri K, Declerck T, Goggi S, Grobelnik M, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, editors. Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). Paris, France: European Language Resources Association (ELRA): 2016.

[5] Tiedemann, J. (2012, May). Parallel Data, Tools and Interfaces in OPUS. In Lrec (Vol. 2012, pp. 2214-2218).

[6] Dušek, O., Hajič, J., Hlaváčová, J., Novák, M., Pecina, P., Rosa, R., ... & Zeman, D. (2014). Machine translation of medical texts in the khresmoi project. In Proceedings of the Ninth Workshop on Statistical Machine Translation (pp. 221-228).

[7] Padró, L. (2011). Analizadores multilingües en freeling. Linguamática, 3(1), 13-20.

[8] Oronoz, M. & Casillas, A. & Gojenola, K. & P ́erez, A.: Automatic Annotation of Medical Records in Spanish with Disease, Drug and Substance Names. Lecture Notes in Computer Science, 8259. Progress in Pattern Recognition, ImageAnalysis, ComputerVision, and Applications 18th Iberoamerican Congress, CIARP 2013 Havana, Cuba, November 20-23. (2013)

[9] Aronson, A. R., & Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association, 17(3), 229-236.

[10]Gojenola, K., Oronoz, M., Pérez, A., & Casillas, A. (2014). IxaMed: Applying freeling and a perceptron sequential tagger at the shared task on analyzing clinical texts. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) (pp. 361-365).

[11]Jonquet, C., Shah, N., Youn, C., Callendar, C., Storey, M. A., & Musen, M. (2009, October). NCBO annotator: semantic annotation of biomedical data. In International Semantic Web Conference, Poster and Demo session (Vol. 110).

[12]Dai, M., Shah, N. H., & Xuan, W. (2008). An efficient solution for mapping free text to ontology terms. AMIA Summit on Translational Bioinformatics. San Francisco CA.

[13]Shah, N. H., Bhatia, N., Jonquet, C., Rubin, D., Chiang, A. P., & Musen, M. A. (2009, September). Comparison of concept recognizers for building the Open Biomedical Annotator. In BMC bioinformatics (Vol. 10, No. 9, p. S14). BioMed Central.

[14]Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., & Jacobson, R. S. (2016). NOBLE–Flexible concept recognition for large-scale biomedical natural language processing. BMC bioinformatics, 17(1), 32.

[15]Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., & Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association, 17(5), 507-513.

[16]Campos, D., Matos, S., & Oliveira, J. L. (2013). A modular framework for biomedical concept recognition. BMC bioinformatics, 14(1), 281.

[17]Campos, D., Matos, S., & Oliveira, J. L. (2013). Neji: a tool for heterogeneous biomedical concept identification. Proceedings of BioLINK SIG, 20, 1178.

[18]Nunes, T., Campos, D., Matos, S., & Oliveira, J. L. (2013). BeCAS: biomedical concept recognition services and visualization. Bioinformatics, 29(15), 1915-1916.

[19]Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., & Xu, H. (2017). CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. Journal of the American Medical Informatics Association, 25(3), 331-336.

[20]Iglesias Maqueda, A., Castro Galán, E., Pérez Láinez, R., Castaño Zabaleta, L., Martínez Fernández, P., Gómez Pérez, J. M., ... & Melero, R. (2008). MOSTAS: un etiquetador morfo-

semántico, anonimizador y corrector de historiales clínicos. Procesamiento del lenguaje natural. N. 41 (septiembre 2008); pp. 299-300.

[21]Carrero, F. M., Cortizo, J. C., Gómez, J. M., & De Buenaga, M. (2008, October). In the development of a spanish metamap. In Proceedings of the 17th ACM conference on Information and knowledge management (pp. 1465-1466). ACM.

[22]Castro, E., Iglesias, A., Martínez, P., & Castano, L. (2010, November). Automatic identification of biomedical concepts in spanish-language unstructured clinical texts. In Proceedings of the 1st ACM International Health Informatics Symposium (pp. 751-757). ACM.

[23]Roller, R., Kittner, M., Weissenborn, D., & Leser, U. (2018). Cross-lingual Candidate Search for Biomedical Concept Normalization. arXiv preprint arXiv:1805.01646.

[24]Lee, J., Cho, K., & Hofmann, T. (2016). Fully character-level neural machine translation without explicit segmentation. arXiv preprint arXiv:1610.03017.

[25]Costa-jussà, M. R., Casas, N., & Melero, M. (2018). English-Catalan Neural Machine Translation in the Biomedical Domain through the cascade approach. arXiv preprint arXiv:1803.07139.

[26]Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008).

[27]Perez-de-Viñaspre, O., & Oronoz, M. (2015, December). SNOMED CT in a language isolate: an algorithm for a semiautomatic translation. In BMC medical informatics and decision making (Vol. 15, No. 2, p. S5). BioMed Central.

[28]Accuosto, P., & Saggion, H. (2018). Improving the accessibility of biomedical texts by semantic enrichment and definition expansion.

[29]Perez, N., Cuadros, M., & Rigau, G. (2018). Biomedical term normalization of EHRs with UMLS. arXiv preprint arXiv:1802.02870.

[30]Intxaurrondo, A., Pérez-Pérez, M., Pérez-Rodríguez, G., López-Martín, J. A., Santamaria, J., de la Pena, S., ... & Krallinger, M. (2017). The Biomedical Abbreviation Recognition and Resolution (BARR) track: benchmarking, evaluation and importance of abbreviation recognition systems applied to Spanish biomedical abstracts.

[31]Castano J, Gambarte ML, Park HJ, Avila Williams MdP, Perez D, Campos F, Luna D, Benitez S, Berinsky H, Zanetti S. A machine learning approach to clinical terms normalization. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing. Berlin, Germany: Association for Computational Linguistics: 2016. p. 1–11. http://anthology.aclweb.org/W16-2901.

[32]Segura-Bedmar I, de la Peña González S, Martínez P. Extracting drug indications and adverse drug reactions from Spanish health social media. In: Proceedings of BioNLP 2014. Baltimore, Maryland: Association for Computational Linguistics: 2014. p. 98–106.

[33]Casillas, A., Gojenola, K., Pérez, A., & Oronoz, M. (2016, December). Clinical text mining for efficient extraction of drug-allergy reactions. In Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on (pp. 946-952). IEEE.

[34]Koza, W., Solana, Z., Conrado, M. D. S., Rezende, S. O., Pardo, T. A., Díaz-Labrador, J., & Abaitua, J. Extracción terminológica en el dominio médico a partir del reconocimiento de sintagmas nominales [0]. 2011

[35]Yepes, A. J., Névéol, A., Neves, M., Verspoor, K., Bojar, O., Boyer, A., ... & Pecina, P. (2017). Findings of the WMT 2017 biomedical translation shared task. In Proceedings of the Second Conference on Machine Translation (pp. 234-247).

[36]Espinosa-Anke, L., Tello, J., Pardo, A., Medrano, I., Ureña, A., Salcedo, I., & Saggion, H. (2016). Savana: A Global Information Extraction and Terminology Expansion Framework in the Medical Domain.

[37]Liu W, Cai S. Translating electronic health record notes from English to Spanish: A preliminary study. In: Proceedings of BioNLP 15. Beijing, China: Association for Computational Linguistics: 2015. p. 134–140.

[38]Costumero, R., García-Pedrero, Á., Gonzalo-Martín, C., Menasalvas, E., & Millan, S. (2014, August). Text analysis and information extraction from Spanish written documents. In International Conference on Brain Informatics and Health (pp. 188-197). Springer, Cham.

8. GLOSARIO DE SIGLAS Y ACRÓNIMOS

ACTH	Hormona adrenocorticotropa
AEMPS	Agencia Española de Medicamentos y Productos Sanitarios
AUI	Atom Unique Identifiers
BARR	Biomedical Abbreviation Recognition and Resolution
BNCS	Biblioteca Nacional de Ciencias de la Salud
CAP	College of American Pathologists
CBOW	Continuous Bag-of-Words
ChEBI	Chemical Entities of Biomedical Interest
CIMA	Centro de información online de medicamentos de la AEMPS
CMS	Centers for Medicare and Medicaid Service
CoNLL	The SIGNLL Conference on Computational Natural Language Learning
CPCSPA	ICPC Spanish
CPTSP	Current Procedural Terminology Spanish
CRF	conditional random field
CTV3	Clinical Terms Version 3
CUI	Concept Unique Identifiers
CUIDEN	Base de Datos Bibliográfica de la Fundación Index
DRI	Dr. Inventor
EHR	Electronic health record
EMEA	European Medicines Agency
ENFISPO	Biblioteca de la Facultad de Enfermería, Fisioterapia y Podología de la Universidad Complutense de Madrid

FECYT	Fundación Española para la Ciencia y la Tecnología
HUGO	Human Genome Organisation
IBECS	Índice Bibliográfico Español de Ciencias de la Salud
ICD	International Classification of Diseases
ICPC	International Classification of Primary Care
ICPCBAQ	ICPC Basque
IHTSDO	International Health Terminology Standards Development Organisation
IME	Índice Médico Español
IPC	International Patent Classification
ISCIII	Instituto de Salud Carlos III
ISSN	International Standard Serial Number
IULA	Institut de Lingüística Aplicada
JSON	JavaScript Object Notation
LNC-ES-AR	LOINC Linguistic Variant - Spanish, Argentina
LNC-ES-CH	LOINC Linguistic Variant - Spanish, Switzerland
LNC-ES-ES	LOINC Linguistic Variant - Spanish, Spain
LREC	Language Resources and Evaluation Conference
LUI	Lexical (term) Unique Identifiers
MDRSPA	MedDRA Spanish
MedDRA	Medical Dictionary for Regulatory Activities
MEDES	MEDicina en ESpañol
MeSH	Medical Subject Headings
MSHSPA	MeSH Spanish

MSSSI	Ministerio de Sanidad, Servicios Sociales e Igualdad
NCBI	National Center for Biotechnology Information
NCBO	National Center for Biomedical Ontology
NHS	National Health Service
NLM	National Library of Medicine
NLTK	Natural Language Toolkit
PLN	Procesamiento del Lenguaje Natural
POS	Part-of-speech
SciELO	Scientific Electronic Library Online
SCN	Nombre Científico
SCTSPA	SNOMED CT Spanish Edition
Snomed RT	Snomed Reference Terminology
SNS	Sistema Nacional de Salud
SUI	String Unique Identifiers
TUI	Type Unique Identifier
UIMA	Unstructured Information Management Architecture
UMLS	Unified Medical Language System
UTS	Servicios de Terminología UMLS
WHOART	WHO Adverse Drug Reaction
WHOSPA	WHOART Spanish
WSD	Word Sense Disambiguation
XML	Extensible Markup Language

ANEXO

ANEXO I: Listado de las 26 revistas biomédicas españolas proporcionadas por [1] y ordenadas por IF.

Country	Journal title (abbreviated)	ISSN	I.F.	Category	Ranking within category	Quartile within category
Rev Esp Cardiol		0300-8932	3.204	Cardiac and Cardiovascular Systems	39/124	Q2
Emergencias		1137-6821	2.578	Emergency Medicine	3/25	Q1
Rev Esp Enferm Dig		1130-0108	1.652	Gastroenterology and Hepatology	53/74	Q3
Enferm Infec Micr Cl		0213-005X	1.478	Infectious Diseases	58/70	Q4
				Microbiology	89/116	Q4
Med-Clin Barcelona		0025-7753	1.399	Medicine, General and Internal	65/155	Q2
Arch Bronconeumol		0300-2896	1.372	Respiratory System	41/50	Q4
Med Intensiva		0210-5691	1.323	Critical Care Medicine	23/27	Q4
Neurologia		0213-4853	1.322	Clinical Neurology	143/193	Q4
Nutr Hosp		0212-1611	1.305	Nutrition and Dietetics	57/76	Q4
Nefrologia		0211-6995	1.274	Urology and Nephrology	55/73	Q4
Allergol Immunopath		0301-0546	1.229	Allergy	16/23	Q3
				Immunology	120/137	Q4
Rev Neurologia		0210-0010	1.179	Clinical Neurology	154/193	Q4
Actas Urol Esp		0210-4806	1.144	Urology and Nephrology	57/73	Q4
Gac Sanit		0213-9111	1.116	Public, Environ. And Occupational Health	115/161	Q3
Med Oral Patol Oral		1698-6946	1.017	Dentistry, Oral Surgery and Medicine	53/83	Q3
Aten Prim		0212-6567	0.957	Primary Health Care	13/18	Q3
				Medicine, General and Internal	86/155	Q3
Cir Espan		0009-739X	0.871	Surgery	137/199	Q3
An Pediatr		1695-4033	0.867	Pediatrics	91/122	Q3
Rev Esp Med Nucl Ima		2253-654X	0.863	Radiology, Nuclear Medicine, and Medical Imaging	100/120	Q4
Gastroent Hepat-Barc		0210-5705	0.567	Gastroenterology and Hepatology	69/74	Q4
An Sist Sanit Navar		1137-6627	0.351	Public, Environ. And Occupational Health	153/161	Q4
Neurocirugia		1130-1473	0.343	Neurosciences	242/252	Q4
				Surgery	177/199	Q4
Med Paliativa		1134-248X	0.326	Health Care Sciences and Services	83/83	Q4
Rev Int Androl		1698-031X	0.256	Andrology	6/6	Q4
Rev Int Med Cienc Ac		1577-0354	0.205	Sports Sciences	81/84	Q4
Aten Farm		1139-7357	0.125	Pharmacology and Pharmacy	257/261	Q4
Invest Clin		0535-5133	0.394	Medicine, Research and Experimental	108/121	Q4
Arch Latinoam Nutr		0004-0622	0.241	Nutrition and Dietetics	73/76	Q4
Kasmera		0075-5222	0.071	Tropical Medicine	21/22	Q4
Arch Latinoam Nutr		0004-0622	0.241	Nutrition and Dietetics	73/76	Q4

ANEXO II: Listado de todos los módulos implementados en Freeling y para qué idiomas.

	as	ca	cy	de	en	es	fr	gl	hr	it	nb	pt	ru	sl
Tokenization	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Sentence splitting	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Number detection		X		X	X	X	X	X		X		X	X	
Date detection		X		X	X	X	X	X				X	X	
Morphological dictionary	X	X	X	X	X	X	X	X		X	X	X	X	X
Affix rules	X	X	X	X	X	X	X	X		X	X	X		
Multiword detection	X	X	X		X	X	X	X		X		X		
Basic named entity detection	X	X	X		X	X	X	X		X		X	X	X
B-I-O named entity detection		X			X	X		X				X		
Named Entity Classification		X			X	X						X		
Quantity detection		X			X	X		X				X	X	
PoS tagging	X	X	X	X	X	X	X	X		X	X	X	X	X
Phonetic encoding					X	X								
WN sense annotation		X			X	X	X	X	X					X
UKB sense disambiguation		X			X	X	X	X	X					X
Shallow parsing	X	X			X	X		X				X		
Full/dependency parsing	X	X			X	X		X	X					X
Semantic Role Labelling		X		X	X	X								
Coreference resolution					X	X								