First Steps towards a Medical Lexicon for Spanish with Linguistic and Semantic Information

Leonardo Campillos-Llanos
Computational Linguistics Laboratory, Universidad Autónoma de Madrid
leonardo.campillos@uam.es

Abstract

We report the work-in-progress of collecting MedLexSp, an unified medical lexicon for the Spanish language, featuring terms and inflected word forms mapped to Unified Medical Language System (UMLS) Concept Unique Identifiers (CUIs), semantic types and groups. First, we leveraged a list of term lemmas and forms from a previous project, and mapped them to UMLS terms and CUIs. To enrich the lexicon, we used both domain-corpora (e.g. Summaries of Product Characteristics and MedlinePlus) and natural language processing techniques such as string distance methods or generation of syntactic variants of multi-word terms. We also added term variants by mapping their CUIs to missing items available in the Spanish versions of standard thesauri (e.g. Medical Subject Headings and World Health Organization Adverse Drug Reactions terminology). We enhanced the vocabulary coverage by gathering missing terms from resources such as the Anatomical Therapeutical Classification, the National Cancer Institute (NCI) Dictionary of Cancer Terms, OrphaData, or the Nomenclátor de Prescripción for drug names. Part-of-Speech information is being included in the lexicon, and the current version amounts up to 76 454 lemmas and 203 043 inflected forms (including conjugated verbs, number and gender variants), corresponding to 30 647 UMLS CUIs. MedLexSp is distributed freely for research purposes.

1 Introduction

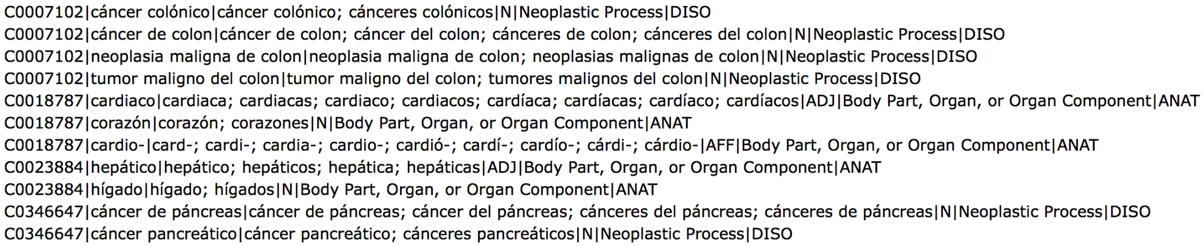
Current machine-learning and deep-learning-based methods are *data-intensive*; however, in domains such as Medicine, sufficient data are not always available—due to ethical concerns or privacy issues, especially when dealing with Patient Protected Information. Moreover, some tasks demand high precision outcomes, which either need supervised approaches with annotated data or hybrid methods (e.g. rule-based and dictionary-based). In order to overcome the data bottleneck, richly-structured terminological thesauri enhance the annotation and concept normalization of domain corpora to be used subsequently in supervised models. More importantly, to achieve comparable benchmarks, domain resources should integrate standard terminologies and coding schemes.

In this context, we aim at providing a computational lexicon to be used in the pre-processing of text data used in more complex Natural Language Processing (NLP) tasks. The work here presented reports the first steps towards building the Medical Lexicon for Spanish (MedLexSp). MedLexSp is conceived as an unified resource with linguistic information (lemmas, inflected forms and part-of-speech), concepts mapped to Unified Medical Language System® (hereafter, UMLS) (Bodenreider, 2004) Concept Unique Identifiers (CUIs), and semantic information (UMLS types and groups). Figure 1 is a sample of the lexicon. MedLexSp is firstly aimed at named entity recognition (NER), and it can be used in the pre-annotation step of an NER pipeline. It can also help lemmatization and feed general-purpose Part-of-Speech taggers applied to medical texts—as done in previous works (Oronoz et al., 2013).[1] Because it gathers semantic data of terms, it can ease relation extraction tasks.

Our work makes several contributions. We provide a resource to be distributed for research purposes in the BioNLP community. MedLexSp includes inflected forms (singular/plural, masculine/feminine) and conjugated verb forms of term lemmas, which are mapped to UMLS Concept Unique Identifiers. Verb terms are also mapped to Concept Unique Identifiers; this is the line of current works for expanding terminologies by in-

[1] https://zenodo.org/record/2621286

Proceedings of the BioNLP 2019 workshop, pages 152–164
Florence, Italy, August 1, 2019. ©2019 Association for Computational Linguistics



```
C0007102|cáncer colónico|cáncer colónico; cánceres colónicos|N|Neoplastic Process|DISO
C0007102|cáncer de colon|cáncer de colon; cáncer del colon; cánceres de colon; cánceres del colon|N|Neoplastic Process|DISO
C0007102|neoplasia maligna de colon|neoplasia maligna de colon; neoplasias malignas de colon|N|Neoplastic Process|DISO
C0007102|tumor maligno del colon|tumor maligno del colon; tumores malignos del colon|N|Neoplastic Process|DISO
C0018787|cardiaco|cardiaca; cardiacas; cardiaco; cardiacos; cardíaca; cardíacas; cardíaco; cardíacos|ADJ|Body Part, Organ, or Organ Component|ANAT
C0018787|corazón|corazón; corazones|N|Body Part, Organ, or Organ Component|ANAT
C0018787|cardio-|card-; cardi-; cardia-; cardio-; cardió-; cardí-; cardío-; cárdi-; cárdio-|AFF|Body Part, Organ, or Organ Component|ANAT
C0023884|hepático|hepático; hepáticos; hepática; hepáticas|ADJ|Body Part, Organ, or Organ Component|ANAT
C0023884|hígado|hígado; hígados|N|Body Part, Organ, or Organ Component|ANAT
C0346647|cáncer de páncreas|cáncer de páncreas; cáncer del páncreas; cánceres del páncreas; cánceres de páncreas|N|Neoplastic Process|DISO
C0346647|cáncer pancreático|cáncer pancreático; cánceres pancreáticos|N|Neoplastic Process|DISO
```

Figure 1: Sample of the MedLexSp lexicon. In each entry, field 1 is the UMLS CUI of the entity; field 2, the lemma; field 3, the variant forms; field 4, the Part-of-Speech; field 5, the semantic types(s); and field 6, the semantic group.

cluding verb terms (Thompson et al., 2011; Chiu et al., 2019). We also added inflected terms from MedlinePlus terms, OrphaData (INSERM, 2019), the National Cancer Institute (NCI) Dictionary of Cancer Terms, or the Nomenclator de prescripción (AEMPS, 2019), a knowledge base of medical drugs prescribed in Spain.

Section 2 gives an overview of medical thesauri, and Section 3 describes the methods used to gather terms (both corpora and NLP techniques), map them to UMLS CUIs, and enrich the lexicon. Section 4 reports descriptive statistics of the current version, and Section 5, the results of an evaluation conducted during development. We discuss some limitations and conclude in Section 6.

2 Background and Context

2.1 Health thesauri and taxonomies

Medical thesauri and controlled vocabularies aggregate listings of domain terms, and also gather information about the type of term (e.g. synonym or preferred term), a semantic descriptor (e.g. DRUG or FINDING), an unique concept identifier, and very often a term definition or hierarchical relations between concepts (e.g. IS_A). Thesauri are essential for indexing and populating databases, domain-specific information retrieval, and standardized codification (Cimino, 1996).

Medical thesauri vary according to the application (we only give examples related to our work). The Systematized Nomenclature of Medicine Clinical Terms (SNOMED-CT) (Donnelly, 2006) aims at encoding verbatim mentions in clinical texts, and gathers ontological relations between concepts. To report drug reactions in pharmacovigilance, the World Health Organization created the Adverse Reactions Terminology (WHO ART), although the Medical Dictionary for Regulatory Activities (MedDRA) (Brown et al., 1999) is now preferred. The Medical Subject Headings (MeSH) are developed by the National Library of Medicine for indexing biomedical articles. Lastly, the World Organization of Family Doctors produced the International Classification of Primary Care (ICPC) to classify data aimed at family and primary care physicians (WONCA, 1998).

Medical taxonomies or classifications gather essential domain knowledge.Some examples are the International Classification of Diseases vs. 10 (ICD-10) (WHO, 2004), or the Anatomical Therapeutical Chemical (ATC) classification of pharmacological substances (WHO, 2019).

2.2 Medical Lexicons

Medical lexicons provide a structured representation of terms and their linguistic information (lemmas, inflection, or surface variants); hence, they are essential for NLP tasks. Unlike medical thesauri or classifications, they do not register term hierarchies, classifications nor ontological relations, but they can encode semantic information and, occasionally, argument structure and corpus-based frequency data (Thompson et al., 2011).

Initiatives to collect medical lexicons have been conducted for English (McCray et al., 1994; Johnson, 1999; Davis et al., 2012), German (Weske-Heck et al., 2002), French (Zweigenbaum et al., 2005) or Swedish, even in multilingual initiatives (Markó et al., 2006). For Spanish, some efforts were sparked when a team at the National Library of Medicine (Divita et al., 2007) started to build an equivalent of the MetaMap tool (Aronson, 2001). Other teams conducted experiments to automate the creation of a Spanish MetaMap by applying machine translation and domain ontologies (Carrero et al., 2008). These initiatives, to the best of our knowledge, did not achieve a Spanish lexicon available for medical NLP.

Besides medical lexicons, domain-specific vocabularies were collected for Biology (Thompson et al., 2011). With a different perspective and goal, Consumer Health Vocabularies have been collected to bridge the gap between patients' expressions and healthcare professionals' jargon (Zeng and Tse, 2006; Keselman et al., 2007).

2.3 The Unified Medical Language System

The Unified Medical Language System® (UMLS) (Bodenreider, 2004) MetaThesaurus includes thesauri. The version we used (2018AB) gathers 210 sources and over 3.82 millions of concepts in 23 languages. Synonym terms are encoded with Concept Unique Identifiers (CUIs); and concepts are assigned a semantic type and group (McCray et al., 2001).

2.4 Methods for Creating Medical Lexicons

We will restrict us here to a shallow overview of approaches and will not consider taxonomy nor ontology building. Methods for widening medical vocabularies range from generating syntactic-level variants of multi-word terms (Jacquemin, 1999), inferring derivation rules from string similarity matches and morphological relations between derivational variants (Grabar and Zweigenbaum, 2000), gathering inflected variants semi-automatically (Cartoni and Zweigenbaum, 2010), or deriving terms from corpora (more below).

Graeco-Latin components are very productive for coining medical terms; thus, several BioNLP systems integrate morphology-based lexical resources. For example, for decomposing terms morphosemantically and deriving their definitions (Namer and Zweigenbaum, 2004), or mapping queries to concepts and indexing documents in cross-lingual information retrieval, based on a subword-based morpheme thesaurus (Markó et al., 2005). In this line, generating paraphrase equivalents of neoclassical compounds (e.g. *thyromegalia* → *enlarged thyroid*) is an approach with potential for deriving new terms, and concept normalization systems (Thompson and Ananiadou, 2018) already implement it. Because string similarity measures and edit distance patterns are used for normalization—e.g (Tsuruoka et al., 2007; Kate, 2015)—and terminology mapping (Dziadek et al., 2017), these approaches are also powerful for expanding medical lexicons from a set of reference terms. Decomposition of multi-word terms and synonym expansion of their components are also alternative strategies applied in normalization systems (Tseytlin et al., 2016).

Corpus-derived medical terminology construction requires collecting domain texts and applying term extraction methods, among others: computing graphs of relations between parse trees and word dependency similarities (Nazarenko et al., 2001), using parallel corpora to map cognates or aligned words (Sbrissia et al., 2004; Deléger et al., 2009), linking terms or abbreviations to their definitions or expanded word forms in the text where they occur (Yu and Agichtein, 2003; McCrae and Collier, 2008), using dictionary features to identify polysemy (Pezik et al., 2008), combining text mining techniques with databases (Thompson et al., 2011), or having experts review terms, a method which has been used to build disease-specific vocabularies (Wang et al., 2016).

Approaches based on the Firthian *Distributional hypothesis* exploit distributional similarity metrics (Carroll et al., 2012). Among them, more recent distributional semantics methods represent terms in the vector space, or calculate word-embeddings to compute similarity measures between vectors, thus allowing the unsupervised expansion of domain terms (Pyysalo et al., 2013; Skeppstedt et al., 2013; Henriksson et al., 2014; Wang et al., 2015; Ahltorp et al., 2016; Segura-Bedmar and Martínez, 2017) or concept normalization (Limsopatham and Collier, 2016).

Lastly, to develop Consumer Health Vocabularies (CHV), a variety of techniques have been used: analysis by experts of Medline queries (Zeng and Tse, 2006), term recognition methods and collaborative review of user logs in medical sites (Zeng et al., 2007), hybrid methods combining n-grams extraction, the C-value, and dictionary look-up (Doing-Harris and Zeng-Treitler, 2011), co-occurrence analysis of terms and seed words (Jiang and Yang, 2013), or approaches based of similarity measures between CHV lexicons and reference lexicons (Seedorff et al., 2013).

3 Methods

Figure 2 depicts the methods used to collect the MedLexSp lexicon. In a first step (left part of Figure 2), we leveraged the lemmas and word forms obtained from a Spanish medical lexicon, mostly corpus-derived; we will refer to it as the *base list*. We only used the subset of lemmas and forms that could be mapped authomatically to UMLS CUIs (exact string match). In a second step, we added missing variants of terms using different methods:

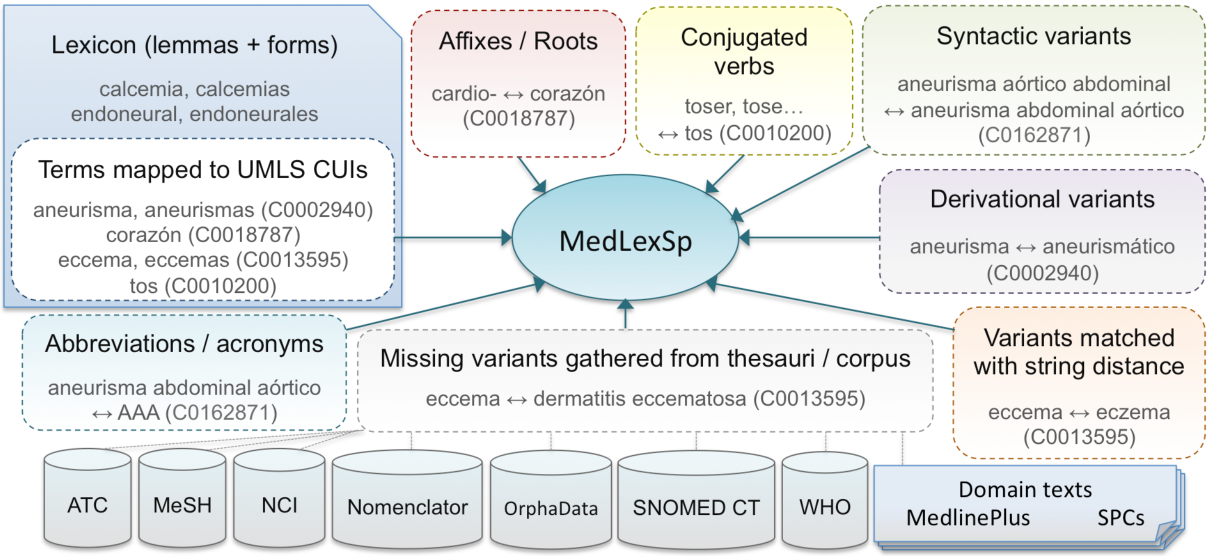


Figure 2: Methods to collect the MedLexSp lexicon.

- Testing **string distance metrics** to match terms in the base list to variants that remained unmatched: e.g. *eccema* ↔ *eczema* ('eczema', C0013595).

- Incorporating **derivational variants** to the base list: e.g. *aneurisma* ('aneurysm') ↔ *aneurismático* ('aneurysmatic', C0002940).

- Including **conjugated verbs** corresponding to the noun terms with CUIs selected in the base list: e.g. *tos* ('cough', C0010200) → *toser*, *tosiendo*... ('to cough', 'coughing'...).

- Matching **affixes and roots** to those terms in the base list with CUIs: e.g. *corazón* ('heart, C0018787) → *cardio-* ('cardio-').

- Adding **syntactic variants** of the multi-word terms in the base list: e.g. *aneurisma aórtico abdominal* ('aortic abdominal aneurysm') ↔ *aneurisma abdominal aórtico* ('abdominal aortic aneurysm', C0162871).

- Adding **acronyms and abbreviations** of the terms included in the base list: e.g. *aneurisma abdominal aórtico* ('abdominal aortic aneurysm', C0162871) → *AAA*.

- Extending the base list by mapping the CUIs of the terms in the subset to gather **missing variants of synonymous terms**: e.g. *eccema* ('eczema', C0013595) ↔ *dermatitis eccematosa* ('eczematous dermatitis', C0013595). We considered several sources from the UMLS—e.g. Spanish Medical Subject Headings (MeSH), SNOMED CT or the WHO ART terminology—and external sources such as the Anatomical Therapeutical Classification, the National Cancer Institute (NCI) Dictionary of Cancer Terms,[2] the Nomenclator de prescripción (AEMPS, 2019), OrphaData (INSERM, 2019), or the Spanish Drug Effect database (SD-Edb) (Segura-Bedmar et al., 2015).

- Including subsets of **missing terms from thesauri if attested in domain texts**. And vice versa, extracting **corpus-derived terms** from domain texts: synonymous terms from MedlinePlus,[3] and terms from Summaries of Product Characteristics (Segura-Bedmar and Martínez, 2017).

The next subsections explain each method.

3.1 Leveraging an Inflected Lexicon

We started using a list of medical terms collected in a previous project on Spanish medical terminology;[4] we will refer to it as the *base list*. We collected this resource by combining different methods (Moreno Sandoval and Campillos Llanos, 2015) applied on a corpus of 4204 Spanish medical texts (around 4 million tokens) (Moreno-Sandoval and Campillos-Llanos, 2013). To extract candidate medical

[2] https://www.cancer.gov/espanol/publicaciones/diccionario

[3] https://medlineplus.gov/spanish/

[4] http://labda.inf.uc3m.es/multimedica/

terms for the base list, we combined rule-based techniques (Part-of-Speech tagging and filtering through medical affixes), corpus-based methods (comparing word forms from a general corpus and from the domain corpus), and statistical methods, namely the Log-Likelihood ratio (Dunning, 1993). We checked in medical sources—e.g. the dictionary published by the Spanish Royal Academy of Medicine (RANME, 2011)—the terms selected by means of those three methods, before being included in the list. This base list was used to build an automatic term extractor (Campillos Llanos et al., 2013), and amounted to 38 354 entries.

Because one of the goals of MedLexSp is concept normalization by using standard domain terminologies, we did not include the full base list. We only used terms that could be assigned UMLS Concept Unique Identifiers (CUIs) in the UMLS MetaThesaurus version 2018AB, namely from those terminologies of special biomedical or clinical interest (e.g. SNOMED CT, WHO ART or Medical Subject Headings) with available Spanish translations. We mapped 18 263 lemmas to CUIs, which means 47.61% entries of the original lexicon. CUIs were assigned according to an *exact match* criterion. For example, *donación* ('donation') is not matched with *donación de tejido* ('Tissue Donation', C0080231), because the latter makes reference to a donation subtype. Note that the current version of MedLexSp does not include the full list of terms from MeSH or SNOMED CT, but only those which were originally mapped from the base list to UMLS terms with CUIs.

3.2 Enriching the Lexicon

String distance metrics We tested mapping terms from the subset of entities with CUIs to terms in the UMLS by applying distance metrics (Levenshtein, 1966) of less than 2. This allowed us mapping hyphenated variants to terms without hyphen (e.g. *creatina-cinasa* ↔ *creatina cinasa*, 'creatine kinase', C0010287), compound terms that are often written as single-words (*dietil éter* ↔ *dietiléter*, 'diethyl ether', C0014994), or matching terms with minimal morphological variation (*eccema* ↔ *eczema*, 'eczema', C0013595). A total of 1463 terms with CUIs were matched to the original base list.

Derivational variants In line with previous work (Grabar and Zweigenbaum, 2000), we collected a list of equivalent derivational variants of terms. Using this list, we assigned a CUI to the corresponding derivational variant: e.g. the CUI of *páncreas* (C0030274) was also ascribed to *pancreático* ('pancreatic'). The current version gathers a total of 801 derivational variants with CUIs.

Conjugated verbs Most terms in the UMLS or standard terminologies are noun or adjective phrases. This limits the named entity recognition of medical concepts expressed with verbs in free text; given a context such as *el paciente tose* ('the patient coughs'), the concept of 'coughing' would not be identified. To widen the scope of concept normalization, verb terms were mapped to CUIs from derived nouns: e.g. *tos* ('coughing', C0010200) → *toser* ('to cough', C0010200). We again used a list of correspondences between verbs and deverbal nouns. We included the conjugated forms of verb lemmas in each verb entry of the lexicon. We used a python script that relies on the lexicon of a Spanish Part-of-Speech tagger (Moreno Sandoval and Guirao, 2006) to generate all conjugated forms of verb terms: e.g. *toser* ('to cough') → *tose* ('he/she coughs'), *tosiendo* ('coughing'), etc. The current version includes a total of 295 single- or multi-word verb items.

Affixes and lexical roots In a first step, we collected affixes and roots from several sources. Firstly, we leveraged a list used in a previous experiment (Sandoval et al., 2013). This list amounts to 1719 forms and considers morphological variants of affixes (e.g. prefix *cardio-* may have accented variant forms in Spanish, such as *cardió-*). Secondly, we translated to Spanish several affixes and roots from the Specialist Lexicon® (McCray et al., 1994) and then added variant forms. In a second step, we assigned UMLS CUIs to affixes and roots in the list. The current list gathers a total of 161 entries (82 prefixes and 79 suffixes) with 134 different CUIs and 386 variant forms. Note that many affixes and roots were not included because they are too underspecified to be assigned to a CUI, or are not restricted to the medical domain (e.g. *kilo-* expresses a quantitative concept).

Abbreviations and acronyms Firstly, we gathered a list of equivalences between full forms and abbreviations and acronyms; we used three sources: 1) the collection of Spanish abbreviations and acronyms used in hospitals, collected by medical doctors (Yetano and Alberola, 2003); 2) abbreviations and acronyms used in the 2nd IberEval Challenge 2018 on Biomedical Abbreviation Recognition and Resolution (Intxaurrondo et al., 2018); and 3) Spanish abbreviations and acronyms from Wikipedia.[5] Secondly, we matched the resulting list of equivalent terms (acronyms and full forms) to UMLS terms, adding the corresponding CUIs to those missing acronyms. For example, the full term *virus de Epstein-Barr* ('Epstein-Barr virus') has CUI C0014644, and we also assigned this code to the corresponding acronym in Spanish (*VEB*). With this method, we assigned CUIs to 1225 items.

Syntactic variants of terms To widen the coverage of terms mapped to CUIs, we generated variants of multiword entities by swapping the word order of their components. Then, we tried to match each new variant to entities with CUIs. For example, *aneurisma aórtico abdominal* ('aortic abdominal aneurysm') has CUI C0162871, and we assigned the same CUI to the generated variant *aneurisma abdominal aórtico* ('abdominal aortic aneurysm'). With this method, we gathered a total of 154 variants of terms with CUIs in the base list.

Mapping UMLS term variants through CUIs We gathered synonymous variants referring to each corresponding concept by using the UMLS CUIs from the terms included in the base list. To avoid including noisy terms adequate for biomedical natural language processing, we first cleaned the terms from the terminologies we used. To do so, we applied methods for cleaning term strings (Aronson et al., 2008; Hettne et al., 2010; Névéol et al., 2012; Hellrich et al., 2015). We deleted paraphrastic terms that include a description or specification of the entity type in the term string. These terms commonly come from Spanish SNOMED CT. For example, we deleted *tos (hallazgo)*, 'cough (finding)' (CUI C0010200) and kept the term (*cough*, 'cough'). Likewise, we removed most anatomic terms beginning with *estructura de* ('structure of'): e.g. regarding term *estructura del ojo* ('structure of eyeball', C0015392), we only kept the synonym *ojo* ('eyeball'). Lastly, terms in the WHO ART terminology needed to be accented and reversed regarding word order: e.g. *disociativa, reaccion* → *reacción disociativa* ('dissociative reaction', C0012746).

We also applied an exact-match mapping of Spanish terms from the base list to the English component of the UMLS. This method allowed us to obtain the CUIs of terms unavailable in Spanish terminologies, which remain unchanged in the Spanish language. Namely, Latin scientific names (e.g. *Campylobacter fetus*, C0006814), compound terms with Graeco-Latin roots (e.g. *abdominalgia*, C0000737), English acronyms that are broadly used in the medical discourse without Spanish translation (e.g. *GABA*, 'gamma-aminobutyric acid', C0016904), or international brand drug names (e.g. *abilify*®). In these cases, the same word is used in both English and Spanish. We manually revised the list of mapped terms to discard homonymous terms with a different meaning in English (e.g. *TIP*® is a brand name of a medical drug, but it also means 'point' or 'suggestion' in English).

We extended the list of terms by extracting the information related to rare diseases from OrphaData (INSERM, 2019).[6] We also added terms of pharmacological substances and international non-proprietary names from the Spanish Drug Effect database (SDEdb) (Segura-Bedmar et al., 2015) and the Nomenclator de prescripción (AEMPS, 2019), a resource published and updated regularly by the Spanish Agency of Drugs and Food Products.[7]

For all these procedures and sources, we applied semiautomatic methods to generate the singular and plural inflected forms of the missing terms that were mapped through CUIs. We used the Pattern python library (Smedt and Daelemans, 2012) to create plural forms of terms, which were revised manually before being included in MedLexSp.

Corpus-derived terms When we started adding variant terms from thesauri, the question of where to stop adding terms came up. In the first version, we decided not to include all terms available in MeSH or SNOMED CT terminologies, given that these thesauri contain terms that are often not necessary in clinical or biomedical NER tasks (e.g. names of trees, wild animals, professions or abstract concepts). On the other hand, to make the

[5] https://es.wikipedia.org/wiki/Anexo:Acrnimos_en_medicina

[6] http://www.orphadata.org/data/xml/es_product1.xml We make available the script used to extract terms from OrphaData: https://github.com/lcampillos/bionlp2019 The code can be adapted to process OrphaData in other languages (e.g. English, French, Italian or Portuguese).

[7] http://listadomedicamentos.aemps.gob.es/prescripcion.zip.

resource comprehensive, we needed to complement the base list with supplementary terms from thesauri. Hence, in order to decide which items to include in a first version, we computed term frequencies using a medical corpus from a previous project (4 million tokens) (Moreno-Sandoval and Campillos-Llanos, 2013). We currently include terms from the Spanish MeSH and SNOMED CT that were missing in the base list, if they were documented in that corpus. By limiting the inclusion of such subset of terms, we aim at providing quality enriched data (i.e. with revised inflected forms) in a reasonable time and manner.

In a different vein, and similarly to former work (Calleja et al., 2017), we extracted terms from Summaries of Product Characteristics (SPCs). We used Easy Drug Package Leaflets (EasyDPL), a corpus of 306 texts annotated with medical drugs and pathological entities (1400 drug effects) (Segura-Bedmar and Martínez, 2017). We annotated these texts and compared our output annotation with regard to this dataset. We used a purely dictionary-based named-entity recogniser with modules for normalization (e.g. lowercasing), tokenization and lemmatization, implemented in spaCy;[8] then, the MedLexSp lexicon was used for exact string matching. We did not use pre- or post-processing rules in the current version (e.g. rules of term composition).

In several iterative rounds, we annotated the texts, identified the unannotated entities, and added them to the lexicon. We did not add (although annotated in the corpus) entities without a CUI, e.g. coordinated entities (e.g. *pies y manos frías*, 'cold hands and feet') or too specific, post-modified terms (e.g. *dolor de cabeza intenso*, 'intense headache'; only 'headache' has CUI C0018681). By using SPCs, we added 837 term entries to MedLexSp, and we ensure that it includes common terms referring to adverse drug reactions and medical drugs.

Lastly, for Consumer Health Vocabulary terms, we extracted synonyms in MedlinePlus Spanish. This resource provides terms in patient language that were missing: e.g. *ojo vago* ('lazy eye') is a synonym of *ambliopía* ('amblyopia', C0002418). We added 783 term entries from this resource. In addition, we collected 6110 cancer-related terms from the Spanish version of the National Cancer Institute Dictionary.

3.3 Semantic and linguistic information

We added to each CUI and lexical entry the corresponding semantic type(s) and group from the UMLS. To avoid noise when annotating biomedical texts semantically, we disfavoured semantic types of the semantic group Concepts and Ideas (CONC, e.g. Quantitative Concept, Functional Concept or Qualitative Concept), which are rather unspecified. We only included terms from that group if no other semantic label was available. If a concept or term can be assigned to two different groups, the element labelled with CONC is not included in our lexicon. For example, the term *inhalación* ('inhalation') can be related to concept C0004048 (semantic type Organism Function, and group PHYS) and also to concept C4521689 (semantic type Intellectual Product, and group CONC). In this case, we only preserve the lexical entry of concept C0004048 and we rule out the entry of concept C4521689.

We have also started adding the Part-of-Speech (PoS) category of each entry in the lexicon. For multiword terms, the category of the head term is selected; e.g. *enfermedad de Crohn* ('Crohn's disease') is categorized as N ('noun'). We are currently testing different techniques to predict the PoS and automate the assignment of categories to each entry, which is still not fully satisfactory.

4 Statistics

Table 1 shows the count of entries in the lexicon according to each source or procedure applied to map terms to UMLS CUIs. Note that the full count exceeds the count of term entries in the current version of MedLexSp, given that some terms were gathered through different methods simultaneously. Table 2 shows the descriptive statistics of the lexicon: counts of lemmas and word forms, and total number of CUIs. Lastly, Table 3 shows a preliminary count of PoS categories in the current version of the lexicon. Note that most entries are nouns or need revision (UNKN stands for 'unknown'); this task is currently being undertaken.

Finally, Figure 3 depicts the distribution of semantic groups. Of note, some groups are underrepresented, due to the corpora and thesauri used to collect terms. For example, few entities belong to the GENE group, which implies that the coverage of the current version of MedLexSp is not adequate for tasks in the Genomics domain.

The amount of lemmas/word forms is lower

[8] https://spacy.io/

Method	# entries
Abbreviations / acronyms	1225
Affixes / roots	161
Derived adjectives	801
Conjugated verb forms	295
Base list mapped to UMLS CUIs:	
Exact match to Spanish UMLS	18 263
Exact match to English UMLS	2534
String distance method	1463
Syntactic variants	134
Terms from thesauri and corpora:	
ATC + Nomenclátor + SDEdb	2931
ICD-10	1299
ICPC	55
MedDRA	5015
MedlinePlus	783
MeSH	6831
NCI	6110
OrphaData	10 741
SNOMED CT	23 096
SPCs (EasyDLP corpus)	837

Table 1: Count of lexical entries according to each source or procedure to map terms to UMLS CUIs.

	Lemmas	Forms	CUIs
Single-words	23 572	23 592	-
Multi-words	52 882	179 451	-
Total	76 454	203 043	30 647

Table 2: Descriptive statistics of the lexicon

PoS	Example	Count
N	*pancreas*	58 830
UNKN	-	13 618
ADJ	*abdominal*	2283
ADJ/N	*gemelo* ('twin')	700
NPR	*Filoviridae*	549
V	*toser* ('to cough')	295
AFF	*cardio-*	161
ADV	*gravemente* ('severely')	20

Table 3: Preliminary counts of Part-of-Speech (PoS) categories. N: 'noun'; UNKN: 'unknown'; ADJ: 'adjective'; ADJ/N: a term that can be an adjective or a noun (depending on the context); NPR: 'proper name'; AFF: 'affix'; ADV: 'adverb'

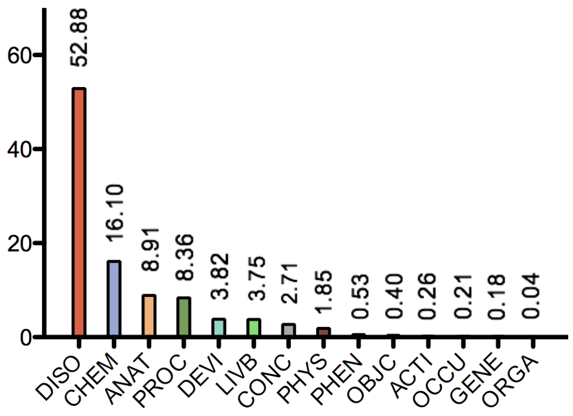


Figure 3: Distribution of semantic groups in the lexicon

than in other UMLS-based resources because: 1) we did not include the full thesauri, but only terms from the original base list that were mapped to UMLS CUIs; and 2) we cleaned noisy terms. As explained, descriptors and qualifiers were removed: e.g. SNOMED CT term *fiebre (hallazgo)* ('fever (finding)', C0015967) was shortened to *fiebre*. We also ruled out some concepts belonging to semantic groups that we can be noisy for clinical or medical NER tasks, such as CONC or GEOG; e.g. *hierro* is related to concept C0302583, 'iron', CHEM; or to concept C0454671, 'Island of Hierro', GEOG (the latter concept was discarded).

5 Development Evaluation

We analysed the coverage of the lexicon with regard to UMLS semantic groups. We applied the dictionary-based NER tool explained below to a gold standard available in the community. We focused on analysing the annotation of few UMLS groups (DISO, CHEM, PROC and ANAT) and assessed how well the lexicon annotated them with regard to the gold standard. We quantified the matched annotations in terms of precision, recall and F1-measure by using the BRAT-Eval script (Verspoor et al., 2013).

A first version of MedLexSp was evaluated with the Spanish texts from the MANTRA corpus (Kors et al., 2015), which gathers 100 texts from the European Medicines Agency (1961 tokens) and 100 texts from Medline (1087 tokens). These texts are available in BRAT format and were annotated with UMLS CUIs, semantic types and groups. We preprocessed the annotated texts for mapping reference annotations to UMLS semantic groups.

With this dataset, we achieved an overall F-measure of 0.83 (exact match) and of 0.87 (approximate match), although the performance var-

	Exact match			Approx. match		
	P	**R**	**F1**	**P**	**R**	**F1**
ANAT	0.64	0.91	0.75	0.67	0.98	0.80
CHEM	0.85	0.96	0.90	0.88	0.97	0.92
DISO	0.83	0.83	0.83	0.89	0.92	0.90
PROC	0.73	0.84	0.78	0.76	0.88	0.82
OA	0.79	0.87	0.83	0.83	0.93	0.87

Table 4: Evaluation of the lexicon; *P*: precision; *R*: recall; *F1*: F-measure; *OA*: overall

ied across semantic groups (Table 4). In our error analysis, we observed that unmatched entities were misspellings (e.g. **deteción* instead of *detección*, 'detection'), discontinuous entities (e.g. *hinchazón de la piel* in *hinchazón y hormigueo de la piel*, 'swelling and tingling of skin'), or entities whose scope was wrongly annotated.

6 Discussion and Conclusions

The lexicon is being developed by means of hybrid NLP methods and corpus-derived terms. We combine the mapping of corpus terms to available thesauri, and viceversa, terms missing in the lexicon were attested in domain texts, so that only a subset of attested terms be included in a first version. Interestingly, searching terms from thesauri in a corpus showed us that many of those terms show low frequencies. From a subset of 56 813 MeSH terms missing in the base list, only 6 676 (11.75%) occurred in the corpus we used (Moreno-Sandoval and Campillos-Llanos, 2013). Although this is due to the influence of the text types, it also reflects the difference betweem terms from thesauri and in real usage. This is another argument that stands for the need for dedicated lexicons combined with NLP methods to achieve successful NER results.

A limitation of our evaluation procedure is the restriction to a very small set of texts; hence, results are not comparable to other tasks or text types. To provide more generalizable results, we need to evaluate the MedLexSp lexicon with another annotated medical corpus in Spanish, but such resource is not freely available to date.

We assume the lexicon is not task-independent. To avoid ambiguity, terms would need to be filtered according to the semantic types needed. For example, terms from the Occupation or Discipline group could be removed for most NER tasks. We are also aware of the limits of a purely lexicon-based approach. Contexts of variation occur in multiwords with coordinated terms (e.g. *cáncer de mama y ovario*, 'breast and ovarian cancer') and adjective modifiers. For example, MedLexSp includes the term *cáncer de mama* ('breast cancer'), but not common variants such as *cáncer de mama derecha* ('right breast cancer') or *cáncer de una mama* ('cancer of one breast'). Both phenomena need specific processing techniques.

Mapping concepts to terms differing across varieties of the Spanish language was not exhaustive. As we departed mainly from a set of corpus-derived terms, most terms belong to the variety used in the texts (i.e. Peninsular Spanish). However, since we used other terminological sources, terms from other varieties were included: e.g. *virus sincitial respiratorio* ('respiratory syncytial virus', C0035236) is a term preferred in Spain or Colombia, but we have the variant *virus sincicial respiratorio* (most frequent in Chile or Argentina). These aspects need nonetheless improvement in future versions, in the same way as the coverage of terms from Consumer Health Vocabularies.

Lastly, we are interested in exploring embedding-based methods for term expansion, and in evaluating the lexicon with a broader set of domain texts.

Acknowledgments

This work has been done under the NLPMedTerm project,[9] funded by the European Union's Horizon 2020 research programme under the Marie Skodowska-Curie grant agreement no. 713366 (InterTalentum UAM). We greatly thank the institutions who gave permission to include their data in MedLexSp, and also thank the anonymous reviewers for their valuable comments.

References

AEMPS. 2019. Nomenclátor de prescripción. `www.aemps.gob.es` *[accessed 2019-03-09].*

Magnus Ahltorp, Maria Skeppstedt, Shiho Kitajima, Aron Henriksson, Rafal Rzepka, and Kenji Araki. 2016. Expansion of medical vocabularies using distributional semantics on Japanese patient blogs. *Journal of Biomedical Semantics*, 7(1):58.

Alan R Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap

[9] `http://www.lllf.uam.es/ESP/nlpmedterm_en.html`

program. In *Proc. of the AMIA Symposium*, pages 17–21. American Medical Informatics Association.

Alan R Aronson, James G Mork, Aurélie Névéol, Sonya E Shooshan, and Dina Demner-Fushman. 2008. Methodology for creating UMLS content views appropriate for biomedical natural language processing. In *Proc. of the AMIA Annual Symposium*, pages 21–25. American Medical Informatics Association.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Elliot G Brown, Louise Wood, and Sue Wood. 1999. The Medical Dictionary for Regulatory Activities (MedDRA). *Drug safety*, 20(2):109–117.

Pablo Calleja, Raúl García-Castro, Guadalupe Aguado-de Cea, and Asunción Gómez-Pérez. 2017. Expanding SNOMED-CT through Spanish Drug Summaries of Product Characteristics. In *Proc. of the Knowledge Capture Conference*, pages 29–37. ACM.

L Campillos Llanos, A Moreno Sandoval, and JM Guirao. 2013. An automatic term extractor for biomedical terms in Spanish. In *Proc. of the 5th Int. Symposium on Languages in Biology and Medicine*, Tokyo, Japan.

Francisco Carrero, José Carlos Cortizo, and José María Gómez. 2008. Building a Spanish MMTx by using automatic translation and biomedical ontologies. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 346–353. Springer.

John Carroll, Rob Koeling, and Shivani Puri. 2012. Lexical acquisition for clinical text mining using distributional similarity. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 232–246. Springer.

Bruno Cartoni and Pierre Zweigenbaum. 2010. Semi-Automated Extension of a Specialized Medical Lexicon for French. In *Proc. of LREC*, Valletta, Malta.

Billy Chiu, Olga Majewska, Sampo Pyysalo, Laura Wey, Ulla Stenius, Anna Korhonen, and Martha Palmer. 2019. A neural classification method for supporting the creation of BioVerbNet. *Journal of Biomedical Semantics*, 10(1):2:1–2:12.

James J Cimino. 1996. Coding systems in health care. *Methods of information in medicine*, 35(04/05):273–284.

Allan Peter Davis, Thomas C Wiegers, Michael C Rosenstein, and Carolyn J Mattingly. 2012. MEDIC: a practical disease vocabulary used at the Comparative Toxicogenomics Database. *Database*, 2012.

Louise Deléger, Magnus Merkel, and Pierre Zweigenbaum. 2009. Translating medical terminologies through word alignment in parallel text corpora. *Journal of Biomedical Informatics*, 42(4):692–701.

Guy Divita, Graciela Rosemblat, and Allen C Browne. 2007. Building a Medical Spanish Lexicon. In *Proc. of the AMIA Symposium*, page 941.

Kristina M Doing-Harris and Qing Zeng-Treitler. 2011. Computer-assisted update of a consumer health vocabulary through mining of social network data. *Journal of Medical Internet Research*, 13(2):e37.

Kevin Donnelly. 2006. SNOMED-CT: The advanced terminology and coding system for eHealth. *Studies in health technology and informatics*, 121:279–290.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1):61–74.

Juliusz Dziadek, Aron Henriksson, and Martin Duneld. 2017. Improving terminology mapping in clinical text with context-sensitive spelling correction. *Informatics for Health: Connected Citizen-Led Wellness and Population Health*, 235:241.

Natalia Grabar and Pierre Zweigenbaum. 2000. A general method for sifting linguistic knowledge from structured terminologies. In *Proc. of the AMIA Symposium*, pages 310–314. American Medical Informatics Association.

Johannes Hellrich, Stefan Schulz, Sven Buechel, and Udo Hahn. 2015. Jufit: A configurable rule engine for filtering and generating new multilingual UMLS terms. In *Proc. of the AMIA Symposium*, pages 604–610. American Medical Informatics Association.

Aron Henriksson, Hans Moen, Maria Skeppstedt, Vidas Daudaravičius, and Martin Duneld. 2014. Synonym extraction and abbreviation expansion with ensembles of semantic spaces. *Journal of Biomedical Semantics*, 5(1):6–.

Kristina M Hettne, Erik M van Mulligen, Martijn J Schuemie, Bob JA Schijvenaars, and Jan A Kors. 2010. Rewriting and suppressing UMLS terms for improved biomedical term identification. *Journal of Biomedical Semantics*, 1(1):5.

INSERM. 2019. Orphadata: Free access data from Orphanet. Data version (XML data version). `http://www.orphadata.org` *[accessed 2019-05-10]*.

Ander Intxaurrondo, Montserrat Marimón, Aitor González-Agirre, José Antonio López-Martín, H Rodríguez Betanco, J Santamaría, Marta Villegas, and Martin Krallinger. 2018. Finding mentions of abbreviations and their definitions in Spanish Clinical Cases: the BARR2 shared task evaluation results. In *Proc. of IberEval@SEPLN 2018*. SEPLN.

Christian Jacquemin. 1999. Syntagmatic and paradigmatic representations of term variation. In *Proc. of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 341–348.

Ling Jiang and Christopher C Yang. 2013. Using co-occurrence analysis to expand consumer health vocabularies from social media data. In *2013 IEEE International Conference on Healthcare Informatics*, pages 74–81. IEEE.

Stephen B Johnson. 1999. A semantic lexicon for medical language processing. *Journal of the American Medical Informatics Association*, 6(3):205–218.

Rohit J Kate. 2015. Normalizing clinical terms using learned edit distance patterns. *Journal of the American Medical Informatics Association*, 23(2):380–386.

Alla Keselman, Tony Tse, Jon Crowell, Allen Browne, Long Ngo, and Qing Zeng. 2007. Assessing consumer health vocabulary familiarity: an exploratory study. *Journal of Medical Internet Research*, 9(1):e5.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Vladimir Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet Physics Doklady*, volume 10, pages 707–710.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1014–1023.

Kornél Markó, Robert Baud, Pierre Zweigenbaum, Lars Borin, Magnus Merkel, and Stefan Schulz. 2006. Towards a multilingual medical lexicon. In *Proc. of the AMIA Annual Symposium*, pages 534–538. American Medical Informatics Association.

Kornél Markó, Stefan Schulz, and Udo Hahn. 2005. MorphoSaurus. Design and evaluation of an interlingua-based, cross-language document retrieval engine for the medical domain. *Methods of information in medicine*, 44(04):537–545.

John McCrae and Nigel Collier. 2008. Synonym set extraction from the biomedical literature by lexical pattern discovery. *BMC Bioinformatics*, 9(1):159.

Alexa T McCray, Anita Burgun, and Olivier Bodenreider. 2001. Aggregating UMLS semantic types for reducing conceptual complexity. *Studies in health technology and informatics*, 84(0 1):216–220.

Alexa T McCray, Suresh Srinivasan, and Allen C Browne. 1994. Lexical methods for managing variation in biomedical terminologies. In *Proc. of the Annual Symposium on Computer Application in Medical Care*, pages 235–239. American Medical Informatics Association.

Antonio Moreno-Sandoval and Leonardo Campillos-Llanos. 2013. Design and Annotation of MultiMedica–A Multilingual Text Corpus of the Biomedical Domain. *Procedia-Social and Behavioral Sciences*, 95:33–39.

Antonio Moreno Sandoval and Leonardo Campillos Llanos. 2015. Combinación de estrategias léxicas y estadísticas para el reconocimiento automático de términos: aplicación a un corpus de medicina. *Lingüística Española Actual*, 37:173–197.

Antonio Moreno Sandoval and José María Guirao. 2006. Morphosyntactic tagging of the Spanish C-ORAL-ROM corpus: Methodology, tools and evaluation. *Spoken language corpus and linguistic informatics*, 5:199–218.

Fiammetta Namer and Pierre Zweigenbaum. 2004. Acquiring meaning for French medical terminology: contribution of morphosemantics. In *Proc. of MedInfo*, pages 535–539.

Adeline Nazarenko, Pierre Zweigenbaum, Benoît Habert, and Jacques Bouaud. 2001. Corpus-based extension of a terminological semantic lexicon. *Recent Advances in Computational Terminology*, pages 327–351.

Aurélie Névéol, Jiao Li, and Zhiyong Lu. 2012. Linking multiple disease-related resources through UMLS. In *Proc. of the 2nd ACM SIGHIT international health informatics symposium*, pages 767–772. ACM.

Maite Oronoz, Arantza Casillas, Koldo Gojenola, and Alicia Perez. 2013. Automatic annotation of medical records in Spanish with disease, drug and substance names. In *Iberoamerican Congress on Pattern Recognition*, pages 536–543. Springer.

Piotr Pezik, A Jimeno-Yepes, V Lee, and D Rebholz-Schuhmann. 2008. Static dictionary features for term polysemy identification. In *Proc. of Building and Evaluating Resources for Biomedical Text Mining LREC Workshop*, Marrakech, Morocco.

Sampo Pyysalo, Filip Ginter, Hans Moen, Salakoski Tapio, and Sophia Ananiadou. 2013. Distributional semantics resources for biomedical text processing. *Proc. of Languages in Biology and Medicine*, pages 39–44.

RANME. 2011. *Diccionario de Términos Médicos*. Editorial Panamericana.

A Moreno Sandoval, L Campillos Llanos, A Gonzlez Martnez, and JM Guirao. 2013. An affix-based method for automatic term recognition from a medical corpus of Spanish. In *Proc. of the 7th Corpus Linguistics Conference 2013*, Lancaster University.

Eduardo Sbrissia, Percy Nohama, Stefan Schulz, and Kornél Markó. 2004. Semi-Supervised Acquisition of a Spanish Lexicon from a Portuguese Seed Lexicon. *Proc. of COLING*.

Michael Seedorff, Kevin J Peterson, Laurie A Nelsen, Cristian Cocos, Jennifer B McCormick, Christopher G Chute, and Jyotishman Pathak. 2013. Incorporating expert terminology and disease risk factors into consumer health vocabularies. In *Biocomputing 2013*, pages 421–432. World Scientific.

Isabel Segura-Bedmar and Paloma Martínez. 2017. Simplifying drug package leaflets written in Spanish by using word embedding. *Journal of Biomedical Semantics*, 8(1):45.

Isabel Segura-Bedmar, Paloma Martínez, Ricardo Revert, and Julián Moreno-Schneider. 2015. Exploring Spanish health social media for detecting drug effects. In *BMC medical informatics and decision making*, volume 15, page S6. BioMed Central.

Maria Skeppstedt, Magnus Ahltorp, and Aron Henriksson. 2013. Vocabulary expansion by semantic extraction of medical terms. *Proc. of Languages in Biology and Medicine*, pages 63–67.

Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.

Paul Thompson and Sophia Ananiadou. 2018. Hyphen. *Terminology.*, 24(1):91–121.

Paul Thompson, John McNaught, Simonetta Montemagni, Nicoletta Calzolari, Riccardo Del Gratta, Vivian Lee, Simone Marchi, Monica Monachini, Piotr Pezik, Valeria Quochi, et al. 2011. The BioLexicon: a large-scale terminological resource for biomedical text mining. *BMC Bioinformatics*, 12(1):397.

Eugene Tseytlin, Kevin Mitchell, Elizabeth Legowski, Julia Corrigan, Girish Chavan, and Rebecca S Jacobson. 2016. Noble–flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics*, 17(1):32.

Yoshimasa Tsuruoka, John McNaught, Junichi Tsujii, and Sophia Ananiadou. 2007. Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics*, 23(20):2768–2774.

Karin Verspoor, Antonio Jimeno Yepes, Lawrence Cavedon, Tara McIntosh, Asha Herten-Crabb, Zoë Thomas, and John-Paul Plazzer. 2013. Annotating the biomedical literature for the human variome. *Database*, 2013.

Chang Wang, Liangliang Cao, and Bowen Zhou. 2015. Medical synonym extraction with concept space models. In *Proc. of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, pages 989–995.

Liqin Wang, Bruce E Bray, Jianlin Shi, Guilherme Del Fiol, and Peter J Haug. 2016. A method for the development of disease-specific reference standards vocabularies from textual biomedical literature resources. *Artificial intelligence in medicine*, 68:47–57.

Gesa Weske-Heck, Albrecht Zaiss, Matthias Zabel, Stefan Schulz, Wolfgang Giere, Michael Schopen, and Rüdiger Klar. 2002. The German specialist lexicon. In *Proceedings of the AMIA Symposium*, pages 884–888. American Medical Informatics Association.

WHO. 2004. *International Statistical Classification of Diseases and Related Health Problems*. World Health Organization.

WHO. 2019. *Anatomical Therapeutical Chemical classification*. Uppsala: Nordic Council on Medicines.

WONCA. 1998. *International Classification of Primary Care 2nd ed.* Oxford: Oxford University Press, 1998.

Javier Yetano and Vincent Alberola. 2003. *Diccionario de siglas médicas y otras abreviaturas, epónimos y términos médicos relacionados con la codificación de las altas hospitalarias*. SEDOM.

Hong Yu and Eugene Agichtein. 2003. Extracting synonymous gene and protein terms from biological literature. *Bioinformatics*, 19(suppl_1):i340–i349.

Qing Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29.

Qing Zeng, Tony Tse, Guy Divita, Alla Keselman, Jonathan Crowell, Allen Browne, Sergey Goryachev, and Long Ngo. 2007. Term identification methods for consumer health vocabulary development. *Journal of Medical Internet Research*, 9(1):e4.

Pierre Zweigenbaum, Robert Baud, Anita Burgun, Fiammetta Namer, Éric Jarrousse, Natalia Grabar, Patrick Ruch, Franck Le Duff, Jean-Francois Forget, Magaly Douyère, and Stéfan Darmoni. 2005. A unified medical lexicon for French. *International Journal of Medical Informatics*, 74(2–4):119–124.

A Appendix - Copyright and Usage

MedLexSp is distributed freely for research purposes; contact for a license at the email address provided or through the project page. Some thesauri included in MedLexSp were obtained through a distribution and usage agreement from the corresponding institutions who develop them. In addition, some material in the UMLS Metathesaurus is from copyrighted sources of the respective copyright holders. Users of the UMLS Metathesaurus are solely responsible for compliance with any copyright, patent or trademark restrictions and are referred to the copyright, patent or trademark notices appearing in the original sources, all of which are hereby incorporated by reference.

The version of MedLexSp freely available for research does not include terms nor coding data from terminological sources with copyright rights; only the subset of data in MedLexSp without usage restrictions is accessible.

We acknowledge the intellectual property rights of the institutions who develop the sources from which we extracted subsets of terms to compile the lexicon, and who gave permission (or provide a licence to reuse their data) to distribute these subsets of terms: the National Library of Medicine maintains the MedLinePlus resource and the Medical Subject Headings, and BIREME/OPS (Latin-American and Caribbean Center on Health Sciences Information) is in charge of the Spanish translation (Descriptores en Ciencias de la Salud, DeCS); the National Cancer Institute publishes the Dictionary of Cancer Terms; the French National Institute of Health and Medical Research (INSERM) supports OrphaNet and gathers the information provided in OrphaData; the World Health Organization produces the Adverse Drug Reactions terminology, the International Classification of Diseases vs. 10, and the Anatomical Therapeutical Classification; the Spanish translation of the International Classification of Primary Care (ICPC) is supported by the World Organization of Family Doctors; and the Spanish Agency of Drugs and Food Products (AEMPS) publishes the Nomenclátor de prescripción. MedLexSp also gathers some terms from the Spanish version of the Medical Dictionary for Regulatory Activities (MedDRA), which is maintained by the Maintenance and Support Services Organization (MSSO). However, the distributed version of MedLexSp does not include terms coming solely from the MedDRA sources, because of copyright restrictions. In addition, MedLexSp includes a subset of the Spanish version of SNOMED Clinical Terms®, which is used by permission of the International Health Terminology Standards Development Organization (IHTSDO; all rights reserved). SNOMED CT® was originally created by The College of American Pathologists.