The MeSpEN Resource for English-Spanish Medical Machine Translation and Terminologies: Census of Parallel Corpora, Glossaries and Term Translations

Marta Villegas[1,2], **Ander Intxaurrondo**[1,2], **Aitor Gonzalez-Agirre**[1,2], **Montserrat Marimon**[1], **Martin Krallinger**[1,2]*

[1]Barcelona Supercomputing Center. Carrer de Jordi Girona, 29, 08034 Barcelona.
[2] Spanish National Cancer Research Center. Calle de Melchor Fernández Almagro, 3, 28029 Madrid
martin.krallinger@gmail.com

Abstract

Clinical and biomedical text mining research efforts have so far focused mainly on documents written in English. These efforts benefited significantly from the availability, not only of domain-specific components such as a tokenizers or Part-of-Speech taggers, but particularly from the access to very large training corpora and terminological resources like UMLS. In order to exploit terminological resources currently restricted to English, it is necessary to promote more systematic translation efforts into other languages, be it manual or by means of machine translation techniques. An initial barrier not only for generating medical machine translation models is the actual identification of relevant datasets that could be exploited to derive glossaries and parallel corpora. Usually relevant datasets weren't constructed as a language technology resource and thus are often overseen by the natural language processing community. This article describes an exhaustive effort to identify and characterize heterogeneous types of documents and glossaries useful to build parallel corpora for Spanish-English medical machine translation systems, including: (1) the combination and harmonization of various bibliographic datasets of biomedical and clinical literature from Spain and Latin America, (2) technical specifications and package leaflets of medicines generated by the pharmaceutical industry, (3) medical and medicinal chemistry patent translations, (4) web-content with trusted information sources about diseases, conditions, and wellness issues for patients, (5) a joined medical multilingual glossary produced by over 500 professional translators and free online medical dictionaries, and (6) keywords derived from bilingual/multilingual medical questionnaires.

Keywords: medical language resources, glossaries, parallel corpora, machine translation, biomedicine

1. Introduction

Currently, a wealth of medical-related information resources do exist in English, not only for patients but also for healthcare professionals, biomedical researches or clinical language technology experts. Such resources include large infrastructures and knowledgebases providing terminologies, biomedical literature or medically related content produced or consumed by either patients or medical professionals. Not surprisingly, this scenario motivated the development of computational tools to improve access, processing and automatic extraction of relevant information by means of natural language processing techniques (Meystre et al., 2008, Krallinger et al., 2008), together with the development of Gold Standard corpora and associated shared tasks and evaluation campaigns (Neves et al., 2014).

The use of specialized medical machine translations techniques may represent a more systematic alternative to generate translations, not only of medical terms, but also of medically related natural language content in general. Efficient medical machine translation systems would be useful not only for generating term translations, but also to assist medical translators and interpreter services in their labor, while patients and healthcare professionals could make better use of clinical information locked in documents subjected to language barriers. Practical adoption of medical machine translation systems could ultimately result in potential improvements in patient safety, diagnostic aid, integration of multilingual clinical information sources and cross-language detection of cases of rare diseases. Moreover, errors in medical interpretation could potentially be reduced by means of medical machine translation assistance (Flores et al., 2003).

Parallel and comparable corpora are a key resource for the development of state-of-the-art corpus-based machine translation (MT), like statistical MT and example-based MT. However, MT systems trained on general-domain data perform poorly in the biomedical domain (Zeng-Treitler et al., 2010), and more recent works use biomedical corpus, sometimes combined with non-medical corpora, to produce higher quality translations (Wu et al., 2011;Yepes et al., 2013; Lui et al., 2015; Neves et al., 2016).

Here we present the MeSpEN (Medical Spanish English) Resource, to our knowledge the first attempt to systematically characterize heterogeneous, complementary, resources and relevant datasets for implementing medical English-Spanish machine translation technologies. The aim of MeSpEN was not to construct the MT technology itself, but to compile medically related bilingual datasets covering different scopes, end users and content types, including bibliographic databases (PubMed[1], Scielo[2] and IBECS[3] along with certified patient-oriented web-content like MedlinePlus[4]).

Figure 1 provides a general overview of the type of resources examined within MeSpEN, which will be detailed throughout this manuscript.

MeSpEN was constructed as part of a framework of the Plan de Impulso de las Tecnologías del Lenguaje de la Agenda Digital (PlanTL) (Plan for Promotion of Language

[1] https://www.ncbi.nlm.nih.gov/pubmed/
[2] http://scielo.org/php/index.php?lang=en
[3] http://ibecs.isciii.es
[4] https://medlineplus.gov/

Technologies), launched by Spanish Ministry of State for Telecommunications with the aim of providing specialized technical support to research and development of software solutions adapted to the field of biomedicine (Villegas et al., 2017). These resources are publicly available for download at http://temu.bsc.es/mespen.

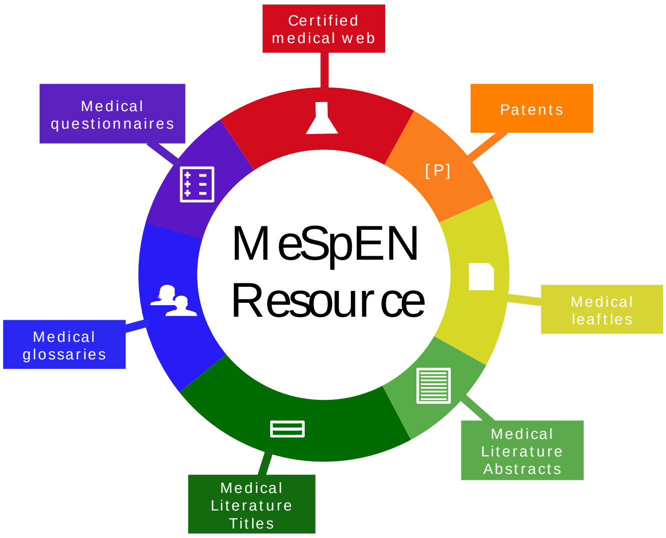


Figure 1: MedSpEN content overview.

2. Related Work

Several efforts were made to generate parallel biomedical corpora and to train statistical medical MT systems. A valuable resource for generating parallel biomedical corpora are medical publications, as it is common to provide, in addition to the full text medical article in any particular language, both the article title as well as its abstracts additionally in English. Although there is a general tendency to directly publish the entire manuscript in English, in case of Spanish medical articles, a growing number of publications published every year can be observed (see figure 2).

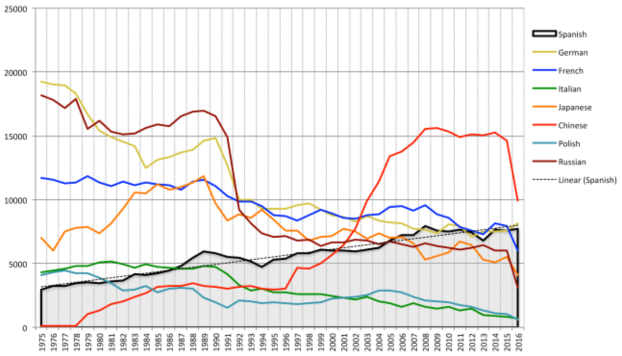


Figure 2: New records/year added to PubMed for non-English articles.

corpus from the foreign language titles (French, Spanish, German, Hungarian, Turkish and Polish) and their corresponding human English translations of Medline/PubMed articles. Yepes et al. (2013) obtained a parallel corpus of article titles from MEDLINE and abstract texts automatically retrieved from journal websites while Neves et al. (2016) presented a parallel corpus of biomedical titles and abstracts from the SciELO database in three pair languages: French/English, Portuguese/English and Spanish/English. Other efforts explored the use of EMEA (Eurpoean Medicines Agency) documents to derive multilingual parallel corpora (Tiedemann 2009) resulting in the OPUS corpus for machine translation. The access to biomedical parallel corpora also motivated carrying out machine translation shared tasks such as the WMT'16 Biomedical translation task (Yeped et al., 2017) and the CLEF-ER (Rebholz-Schuhmann et al., 2013).

3. Methodology and datasets

In order to construct the MeSpEN resource several datasets relevant for Spanish-English medical MT were retrieved, preprocessed and analyzed. A detailed description of each resource can be found in this section.

3.1 Biomedical and clinical literature

The MeSpEN resource includes, among others, bilingual (Spanish and English) data taken from different Biomedical and clinical literature sources, namely: IBECS, SciELO and PubMed. This section describes briefly the way these source data were collected and harmonized into the Dublin Core format

3.1.1 IBECS

The *Índice Bibliográfico Español en Ciencias de la Salud* (IBECS) (Spanish Bibliographic Index in Health Sciences) is a bibliographic database, maintained by the Biblioteca Nacional de Ciencias de la Salud (BNCS, National Health Sciences Library) of the *Instituto de Salud Carlos III* (Carlos III Health Institute) that collects scientific journals covering multiple fields in health sciences (including medicine, pharmacology, nursing, psychology, odontology and physiotherapy, among others) published in Spain from year 2000 onward.

Currently, IBECS includes mainly journals with monolingual content in Spanish[5]: 168,198 records, with an annual increase of more than 12,000 records. Database updates are made weekly and are freely accessible online. IBECS also includes 28,919 links to complete articles through SciELO Spain node[6]. To generate the IBECS subset of the MeSpEN resource we directly collaborated with the Carlos III Health Institute to obtain an XML file including all metadata records in IBECS in December 2017. The file was encoded following the LILACS model, a Virtual Health Library component and is composed by standards, manuals and software, which guide the identification, selection, bibliographic description, document indexing and databases development. The model is widely used in Latin American and the Caribbean countries for health documents indexation.

The mapping from LILACS into Dublin Core was quite straightforward. The following table shows the

[5] http://ibecs.isciii.es/iah/online/E/help/revistas.pdf

[6] http://scielo.isciii.es

correspondences between both models as shown in figure 3.

For some records, however, the language of the title and/or abstract was not specified in the original data. In these cases, we applied a simple language detection algorithm and included this information in the resulting records. Essentially, the system counts the occurrences of Spanish and English stop words in the title and abstract of a given record to discriminate between languages. In case of tie, the system checks if the texts have Spanish characters like accents (á, é,...) or inverted question and exclamation marks ("¿" and "¡").

ab_es	*<dc:description xml:lang="es">*
ab_en	*<dc:description xml:lang="en">*
au	*<dc:creator>*
id	*<dc:identifier>*
ti_es	*<dc:title xml:lang="es">*
ti_en	*<dc:title xml:lang="en">*
is	*<setSpec>*
type	*<dc:type>*
da	*<dc:date>*
mh	*<dc:subject>* (DeCS codes are kept here)
pu	*<dc:publisher>*
fo	*<dc:source>*
la	*<dc:language>*

Figure 3: LILACS Dublin Core mapping.

Initially the corpus had 9.699 titles and 10.252 abstracts with undefined language. After applying the language identification algorithm this number dropped to 487 titles and 62 abstracts. The java program used to convert MeSpEN-IBECS records into Dublin Core can be found in the MeSpEN website (http://temu.bsc.es/mespen).

3.1.2 SCIELO

The Scientific Electronic Library Online (SciELO) is an electronic library supported by the Sao Paulo Research Foundation (FAPESP) and the Brazilian National Council for Scientific and Technological Development (BIREME). This initiative gathers electronic publications of complete full text articles from scientific journals of Latin America, South Africa and Spain.

Currently, SciELO Network is present in 15 countries. Each country operates a collection of open access journals that are published nationally by scientific societies and associations, academic and research related institutions. The network operates over one thousand journals which publishes about 50 thousand articles per year and accumulate more than 700 thousand articles. The journals belong to different disciplines. The SciELO contents are used by different users, specially academic related communities, including students, teachers and researchers as well as professionals and the general public. During the first semester of 2017 the SciELO Network collections served a daily average of more than 1.5 million download. SciELO Spain is supported by the Spanish National Health Sciences Library (Biblioteca Nacional de Ciencias de la Salud, BNCS) .

Most SciELO nodes are available through their OAI-PMH servers and this allows the usage of standard metadata harvesting methods to collect the records. The Argentina, Chile, Peru and Venezuela nodes do not have OAI-PMH nodes and, therefore, they were not included in the corpus. Table 1 shows the number of journals currently available in SciELO for each country together with the number of articles and the number of articles written in Spanish

Although the web pages of the full text articles do have the corresponding translations into English, these translations are not included in the metadata records of the OAI-PMH servers. Thus, the source SciELO's records only contain titles and abstracts in the original language of the article. In order to collect the translations, we used a script that (i) gets the URL in the dc:identifier field of the metadata records pointing to the HTML page of the full article, (ii) retrieves the XML version and, finally, (iii) gets the translated title and abstract[7].

With this approach, the MeSpEN corpus extends the original Dublin Core metadata records with translated titles and abstracts. This enrichment implied adding xml:lang attributes in the original records.

Country	Journals	Publications	Publications in Spanish
Bolivia	23	4,728	4,568
Brazil	396	2,554	73
Colombia	228	53,987	50,783
Costa Rica	37	8,510	6,894
Cuba	70	32,702	31,745
Mexico	200	54,728	45,074
Paraguay	14	1,723	1,657
Portugal	70	10,704	482
South Africa	69	23,565	7
Spain	60	34,820	28,919
Uruguay	27	4,016	3,754
Total	**1,194**	**210,205**	**173,956**

Table 1: Publications in the SciELO network

Note however that about 12,000 out of the 171,000 XML files with full articles had some codification error and were not processed. This explains why the eventual MeSpEN-SciELO resource is smaller that the complete SciELO collection. Table 2 shows the number of current publications in the MeSpEN corpus by country:

Country	Publications in Spanish
Bolivia	4,034
Brazil	39
Colombia	45,658
Costa Rica	6,871
Cuba	25,694
Mexico	44,807
Paraguay	1,648
Portugal	1,648
South Africa	7
Spain	28,843
Uruguay	3,639
Total	**161,710**

[7] The html pages include a link to their XML version where titles and abstracts are in both the original language and English (when available) and the full text in the original language.

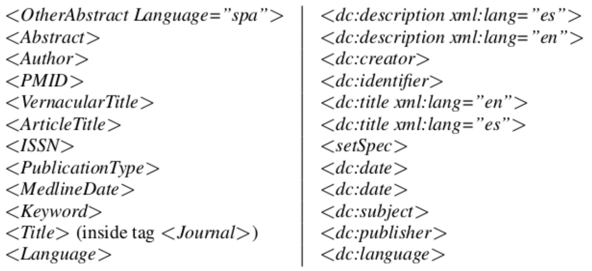
3.1.3 PubMed

PubMed is a free search engine used to access Medline database, a bibliographical database of references and abstracts on life sciences and biomedical topics. It is maintained by the U.S. National Library of Medicine. It contains more than 28 million publications up to 39 languages, including Spanish. The online PubMed search does not display directly non-English content, but it is provided in the actual XML record.

Table 2: the number of publications the MeSpEN in Spanish.

The Medline database stores more 330,000 records for articles published originally in Spanish. Most of the these records do additionally also provide title of the original publication in Spanish. It is possible to retrieve the abstract for at least 127,619 of the records.

PubMed allows downloading search results in XML format using the ‘send to’ option. We used this functionality to get all XML records in Spanish. These records were easily converted into Dublin Core as follows



<OtherAbstract Language="spa">	*<dc:description xml:lang="es">*
<Abstract>	*<dc:description xml:lang="en">*
<Author>	*<dc:creator>*
<PMID>	*<dc:identifier>*
<VernacularTitle>	*<dc:title xml:lang="en">*
<ArticleTitle>	*<dc:title xml:lang="es">*
<ISSN>	*<setSpec>*
<PublicationType>	*<dc:date>*
<MedlineDate>	*<dc:date>*
<Keyword>	*<dc:subject>*
<Title> (inside tag *<Journal>*)	*<dc:publisher>*
<Language>	*<dc:language>*

3.2 Trusted web-content for patient information: MedlinePlus

MedlinePlus is an online information service provided by the U.S. National Library of Medicine (NLM), and gives free information about health in both English and Spanish. MedlinePlus provides the following encyclopedic information:

- **health topics[8]:** summaries about disorders, therapies and body locations, among others.
- **drugs and supplements[9]:** prescription drugs, over-the-counter medicines, dietary supplements and herbal remedies, side effects, dosage and drug interactions.
- **laboratory test information[10]:** information about what the test is used for, or why doctors order it, among other**.**
- **medical encyclopedia11:** articles about diseases, tests, symptoms, injuries and surgeries. This is similar to health topics but with more extended and elaborated content

The current version of the MeSpEN-MedlinePlus corpus only includes the health topics part because it is the only one that includes Dublin Core metadata, while additional content will be added to future releases of MeSpEN

3.2.1 3.2.1 Health topics

MedlinePlus includes 1,063 documents on health topic. The structure of these documents is quite simple: they contain a summary of the topic with links to related sites that provide more information to the user about illnesses’ diagnosis, risk factors, treatments, etc. The summary is available in the English and Spanish versions of MedlinePlus.

The source code of the site stores metadata information about each topic. Metadata elements (meta) are used to include Dublin Core labels and this makes the conversion to Dublin Core records simple and fast.

To create the Dublin Core records, we checked the metadata in the English and Spanish versions of the topic, and joined them. Most of the information in the metadata has its translation to Spanish in the Spanish site. For instance, the title and keywords are always translated, but relation names and MeSH terms always remain untranslated.

3.2.2 MedlinePlus pages

For all the topics of the library, we used the TEI12 standard to create the parallel corpus. This standard is widely used to digitalize long texts in the academic field or digital humanities. The lack of Dublin Core data, and the high amount of information found in these articles motivated the use of this standard.

The corpus is composed by four different files per topic, instead of a unified one like in the previous resources:

- Clean raw text in English.
- TEI file with text in English.
- Clean raw text in Spanish.
- TEI file with text in Spanish.

The clean raw text is structured by sections and paragraphs. At the beginning of each line, a code indicates if the line belongs to a section title, section subtitle, paragraph or listed text. These files were created after extracting the complete text from the HTML source code of each article.

TEI files contain the same text content of the raw files, but structured in XML format. Each section is divided in subsection, paragraph and lists, following the TEI schema. This collection contains a total of 6,292 articles.

3.2.3 Incoherences in the MedlinePlus corpus

To check the quality of the corpus, we analyzed all articles of the library, comparing each article in English with their corresponding Spanish version. Unfortunately, this corpus is not totally parallel: there are situations where some sections are missing in one language, or a paragraph is splitted in two or more paragraphs in the other language, and occasionally the title of a new subsection is marked as paragraph in the other language. Table 3 shows the number of section titles, paragraphs, subsection titles and list elements found for each language, and the number of the documents where we can find these issues.

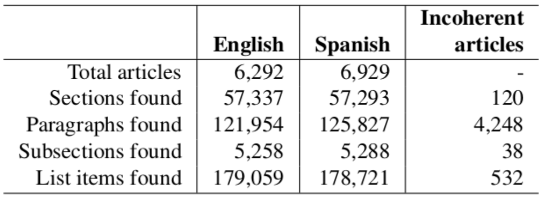
[8] Link: https://medlineplus.gov/healthtopics.html

[9] Link: https://medlineplus.gov/druginformation.html

[10] Link: https://medlineplus.gov/labtests.html

[11] Link: https://medlineplus.gov/encyclopedia.html

[12] Link: http://www.tei-c.org/index.xml



	English	Spanish	Incoherent articles
Total articles	6,292	6,929	-
Sections found	57,337	57,293	120
Paragraphs found	121,954	125,827	4,248
Subsections found	5,258	5,288	38
List items found	179,059	178,721	532

Table 3: MeSpEN-MedlinePlus subset document structure statistics. The last column shows the number of records with alignment issues.

3.3 EMEA corpus

The OPUS - EMEA corpus (Tiedemann, 2009) is a corpus of biomedical documents retrieved from the European Medicines Agency (EMEA). The corpus includes documents related to medicinal products and their translations into 22 official languages of the European Union. It contains roughly 1,500 documents for most of the languages. In particular, the English-Spanish language pair is composed of 1,667 documents, 998,015 sentences and 13,818,929 tokens.

The original EMEA corpus has been compiled out of PDF documents available online. After downloading these documents they were converted to text and sentence aligned. However, the authors did not evaluate the quality of the alignment, and an inspection of the corpus reveals that, unfortunately, the quality of the alignment is not very good. The corpus including all sentence alignments is available at: http://opus.nlpl.eu/download.php?f=EMEA.tar.gz

3.4 COPPA corpus

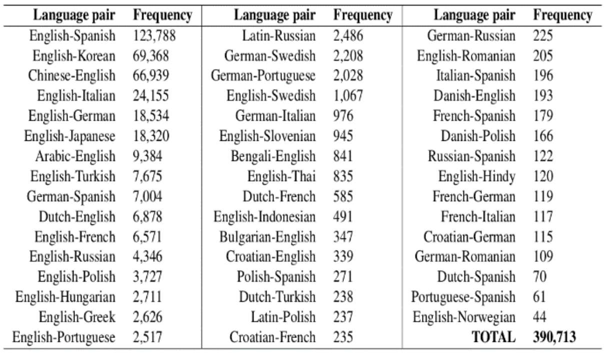
The COPPA corpus seeks to help users and researchers to overcome the language barrier when searching patents published in different languages and to stimulate research in Machine Translation and in language tools for patent texts. The segments included in the corpus are obtained by aligning the sentences of the abstracts and titles of published PCT applications with their translations, the translations having been produced by professional patent translators. The parallel corpus contains 18.303 documents, 62,057 sentences, 2,328,713 tokens and 14,624,745 characters for the English-Spanish language pair. The corpus is available for free for research purposes and for a nominal fee for other purposes, order form and details are available at: http://www.wipo.int/ patentscope/en/data/products.html#coppa

3.5 Bilingual glossaries

Hand crafted glossaries are a particularly valuable resource for the medical translator community and have shown to boost performance of MT systems. We generated 46 bilingual glossaries for various language pairs from free online medical glossaries and dictionaries made by over 500 professional translators. Glossaries were encoded in standard tab-separated values (tsv) format and 26 include English terms, 8 include Spanish terms and 13 files include other languages.

Table 7 shows the number of entries each glossary contains. As can be observed, the largest glossary is the English-Spanish one, with 123,788 terms, followed by English-Korean, with 69,368 terms and Chinese-English, with 66,939 terms.

To evaluate the quality of the glossaries we used a cTakes pipeline to identify UMLS concepts that appear in the glossaries. For time constrictions, we have only been able to process a randomly selected subset of 2% of the glossary in English. In this subset we identified 2,340 CUIs, of which 1,485 are unique (not repeated). We can estimate that in total, for the 100,245 unique terms in the English glossary, we will have about 116,000 UMLS concepts, although it is hard to anticipate how many of them will be unique (the multiplication gives 74,250, but as we increase the number of concepts it gets more difficult to get new/different ones).



Language pair	Frequency	Language pair	Frequency	Language pair	Frequency
English-Spanish	123,788	Latin-Russian	2,486	German-Russian	225
English-Korean	69,368	German-Swedish	2,208	English-Romanian	205
Chinese-English	66,939	German-Portuguese	2,028	Italian-Spanish	196
English-Italian	24,155	English-Swedish	1,067	Danish-English	193
English-German	18,534	German-Italian	976	French-Spanish	179
English-Japanese	18,320	English-Slovenian	945	Danish-Polish	166
Arabic-English	9,384	Bengali-English	841	Russian-Spanish	122
English-Turkish	7,675	English-Thai	835	English-Hindy	120
German-Spanish	7,004	Dutch-French	585	French-German	119
Dutch-English	6,878	English-Indonesian	491	French-Italian	117
English-French	6,571	Bulgarian-English	347	Croatian-German	115
English-Russian	4,346	Croatian-English	339	German-Romanian	109
English-Polish	3,727	Polish-Spanish	271	Dutch-Spanish	70
English-Hungarian	2,711	Dutch-Turkish	238	Portuguese-Spanish	61
English-Greek	2,626	Latin-Polish	237	English-Norwegian	44
English-Portuguese	2,517	Croatian-French	235	**TOTAL**	**390,713**

Table 4: Number of entries in bilingual glossaries

3.6 Keywords derived from bilingual/multilingual medical questionnaires

The MDM-Portal (Medical Data Models[13]) is a metadata registry for creating, analyzing, sharing and reusing medical forms. It contains forms with more than 350,000 data-elements and numerous core data sets, common data elements or data standards, code lists and value sets. Some of the source forms in the system include translations in other languages than English and constitute a potentially interesting multilingual resource. We have directly requested the support of language subset selection by teh MDM-Portal, which is now supported. The translations of concepts in MDM are provided by humans and sometimes they include a reference to the relevant UMLS CUI:

```
<ItemDef OID="I.101" Name="Toxicity" …
    <TranslatedText xml:lang="es">Toxicidad</
   <Alias Context="UMLS CUI" Name="C0013221"/>
</ItemDef>
```

Note however that there are only 199 data-element translated into Spanish and, in some cases, the translations are rather odd:

```
<ItemDef OID="I.104" Name="Comment" …
   <TranslatedText xml:lang="es">Comentarios, notas</
   <Alias Context="UMLS CUI [1]" Name="C0947611"/>
```

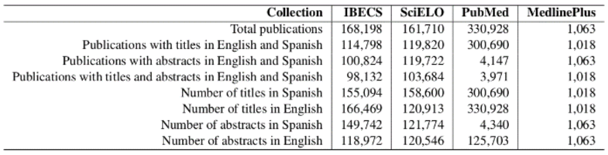
[13] https://medical-data-models.org/

4. MeSpEN statistics

In this section we show different statistics of the created corpus, and also classifying them by source.

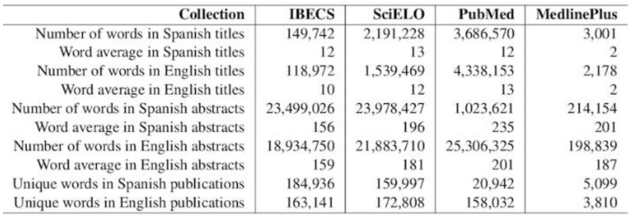
Table 5 displays the total number of MeSpEN literature subsets entries together with source information, and how many entries had titles and abstracts in different languages. This table also shows how many publications could be detected having titles and abstracts in both languages. In general, we can find more parallel titles in English and Spanish, while parallel abstracts were less frequent. It is important to stress that PubMed records generally lack parallel abstracts, as the number of abstracts in Spanish is minimal.

Table 6 shows the number of words of titles and abstracts from different corpora by language; the average number of words is also shown. We used the IXA pipeline (Agerri et al, 2017) tokenizer to detect words prior to count the their amounts in each title and abstract. We can find the word averages of IBECS, SciELO and PubMed to be alike between them, finding a similar nature with MedlinePlus abstracts. Meanwhile, the difference between MedlinePlus titles and other sources is bigger, which provides information about the different nature of this source and the other three; all articles in MedlinePlus use health terms as titles.



Collection	IBECS	SciELO	PubMed	MedlinePlus
Total publications	168,198	161,710	330,928	1,063
Publications with titles in English and Spanish	114,798	119,820	300,690	1,018
Publications with abstracts in English and Spanish	100,824	119,722	4,147	1,063
Publications with titles and abstracts in English and Spanish	98,132	103,684	3,971	1,018
Number of titles in Spanish	155,094	158,600	300,690	1,018
Number of titles in English	166,469	120,913	330,928	1,018
Number of abstracts in Spanish	149,742	121,774	4,340	1,063
Number of abstracts in English	118,972	120,546	125,703	1,063

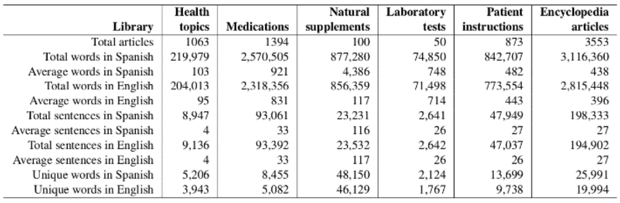
Table 5: Parallel corpus obtained from IBECS, SciELO, PubMed and MedlinePlus.



Collection	IBECS	SciELO	PubMed	MedlinePlus
Number of words in Spanish titles	149,742	2,191,228	3,686,570	3,001
Word average in Spanish titles	12	13	12	2
Number of words in English titles	118,972	1,539,469	4,338,153	2,178
Word average in English titles	10	12	13	2
Number of words in Spanish abstracts	23,499,026	23,978,427	1,023,621	214,154
Word average in Spanish abstracts	156	196	235	201
Number of words in English abstracts	18,934,750	21,883,710	25,306,325	198,839
Word average in English abstracts	159	181	201	187
Unique words in Spanish publications	184,936	159,997	20,942	5,099
Unique words in English publications	163,141	172,808	158,032	3,810

Table 6: Number of words and word average in titles and abstracts.

Table 7 the number of words, sentences, and their averages per document. Looking at the averages, we can find that these statistics are very similar for both the English and the Spanish collection.



Library	Health topics	Medications	Natural supplements	Laboratory tests	Patient instructions	Encyclopedia articles
Total articles	1063	1394	100	50	873	3553
Total words in Spanish	219,979	2,570,505	877,280	74,850	842,707	3,116,360
Average words in Spanish	103	921	4,386	748	482	438
Total words in English	204,013	2,318,356	856,359	71,498	773,554	2,815,448
Average words in English	95	831	117	714	443	396
Total sentences in Spanish	8,947	93,061	23,231	2,641	47,949	198,333
Average sentences in Spanish	4	33	116	26	27	27
Total sentences in English	9,136	93,392	23,532	2,642	47,037	194,902
Average sentences in English	4	33	117	26	26	27
Unique words in Spanish	5,206	8,455	48,150	2,124	13,699	25,991
Unique words in English	3,943	5,082	46,129	1,767	9,738	19,994

Table 7: Number of words, word averages and unique words in MedlinePlus documents.

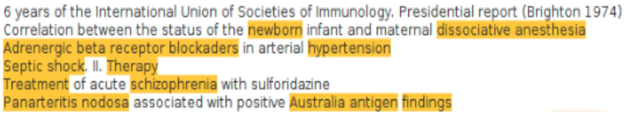
4.1 Explorative use of MeSpEN

One straightforward use of a parallel corpus is to employ it for the enrichment of terminological resources that are less developed in one of the languages. We propose to use the MeSpEN with the objective of enriching UMLS (https://www.nlm.nih.gov/research/umls/) automatically, that is generating candidate term pairs and medical vocabulary in Spanish, both suggesting completely new terms or synonyms of already existing terms in the original UMLS terminological resource.

To achieve this goal we tested titles of PubMed publications for which we had a title in English and its corresponding translated title in Spanish (total 298,040 pairs of titles). Our strategie comprised the following steps:

1. Identify UMLS terms in English titles using cTakes (http://ctakes.apache.org/).
2. Align the words of the titles in English to the words of the titles in Spanish.
3. Using the previous alignment we detected the terms in the titles in Spanish, and we assigned them to their corresponding candidate terms in English.

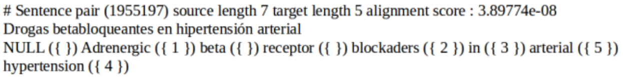
In the first step, to identify the UMLS terms in the English titles we will use a cTakes pipeline that queries UMLS. This pipeline identified the UMLS terms in the unstructured text and assigned to them the most likely Concept Unique Identifier (CUI). Thus, once the titles in English were processed, we were able to extract UMLS terms in English that appear in each title (see Figure 3).



6 years of the International Union of Societies of Immunology. Presidential report (Brighton 1974)
Correlation between the status of the newborn infant and maternal dissociative anesthesia
Adrenergic beta receptor blockaders in arterial hypertension
Septic shock. II. Therapy
Treatment of acute schizophrenia with sulforidazine
Panarteritis nodosa associated with positive Australia antigen findings

Figure 3: UMLS terms identified by the cTakes Clinical pipeline.

In the second step, we used our PubMed titles in English and Spanish as the parallel corpus. Using this parallel corpus we trained a word alignment model using GIZA++. After this process, we obtained the words in English aligned to their correspondences in Spanish. For example, for the titles "*Adregenetic beta receptor blockaders in arterial hypertension*" and "*Drogas betabloqueantes en hipertensión arterial*" we obtain the alignment shown in Figure 4.



```
# Sentence pair (1955197) source length 7 target length 5 alignment score : 3.89774e-08
Drogas betabloqueantes en hipertensión arterial
NULL ({ }) Adrenergic ({ 1 }) beta ({ }) receptor ({ }) blockaders ({ 2 }) in ({ 3 }) arterial ({ 5 })
hypertension ({ 4 })
```

Figure 4: Resulting alignment between "Adregenetic beta receptor blockaders in arterial hypertension" and "Drogas betabloqueantes en hipertensión arterial" using GIZA++.

In the last step, we used the UMLS terms in English detected in the first step, along with the alignments detected in the second step, to predict the candidate text span containing the potential translated terms in Spanish. Once the terms in Spanish were detected, they were extracted and assigned to the same UMLS concept identifier (CUI) as their counterpart in English. In the case of the previous example, the spans detected would be the ones shown in Figure 5. As we can see, using the spans detected by the alignment we can align "*Adregenetic beta receptor blockaders*" to "*Drogas betabloqueantes*" and "*arterial hypertension*" to "*hipertensión arterial*".

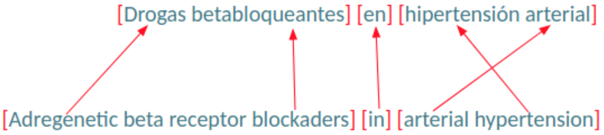

[Drogas betabloqueantes] [en] [hipertensión arterial]
[Adregenetic beta receptor blockaders] [in] [arterial hypertension]

Figure 5: Final translations from English to Spanish using PubMed titles.

A sample set of 200 candidate terms were manually validated by a domain expert. 47% were correct translations, 22% corresponded to either a more general term or more narrow term (hypernym/hyponym) and the remaining pairs were either substrings of the correct translation or wrong translations. The average validation time per term was of just 2.03 seconds, using the MyMiner (Salgado et al., 2012) annotation tool.

5. Discussion

We present a resource for machine translation that is unique in the sense of integrating heterogeneous types of resources for medical machine translation, and harmonizing all the medical literature resources to a common standardized format. Our corpus is composed of publications from three different sources. Right now not OAI-PMH servers work for the SciELO network. It also covers variants of medical Spanish, as it comprises resources from several countries including Spain and Latin American countries such as Argentina, Chile and Venezuela. Other co-official languages in Spain that is Basque, Catalan and Galician are currently not well covered in our resource. Note that we could only find very few publications in Catalan in SciELO (total of 8) and PubMed (total of 88). We are currently analyzing additional information sources to better cover parallel corpora for Galician and Catalan. Moreover, we also explore to directly derive medical glossaries from UMLS for Spanish and Basque. One additional aspect that will deserve a more in depth exploration is to actually compare the lexical characteristics and mentioned UMLS concepts across the various resources that we have gathered to characterize differences in the more formal, scientific language of medical publications, attributes of intellectual property texts found in medicinal chemistry patents and of language expressions in documents whose primary readers are patients, as in the case of MedlinePlus. Although the results of extracting candidate term pairs from our resource describe din section 4.1 is still very preliminary, it is already clear that improving the candidate terms detection followed by manual validation is extremely efficient to quickly expand terminological resources for the medical domain.

6. Acknowledgements

We acknowledge the Encomienda MINETAD-CNIO/OTG Sanidad Plan TL and OpenMinted (654021) H2020 project for funding.

7. Bibliographical References

Chiao, Y. and P. Zweigenbaum, Looking for French-English translations in comparable medical corpora, in: Proceedings of AMIA Symposium, 2002.

Deleger, L., M. Mergel, and P. Zweigenbaum, Translating medical terminologies through word alignment in parallel text corpora, Journal of Biomedical Informatics 42 (4) (2009) 692-701.

Deleger, L., T. Merabti, T. Lecrocq, M. Joubert, P. Sweigenbaum, and S. Darmoni, A Twofold Strategy for Translating a Medical Terminology into French, in: AMIA Annual Symposium Proceedings, 2010.

Flores, G., Laws, M. B., Mayo, S. J., Zuckerman, B., Abreu, M., Medina, L., & Hardt, E. J. (2003). Errors in medical interpretation and their potential clinical consequences in pediatric encounters. Pediatrics, 111(1), 6-14

Huang, C. C., & Lu, Z. (2015). Community challenges in biomedical text mining over 10 years: success, failure and the future. Briefings in bioinformatics, 17(1), 132-144.

Krallinger, M., Valencia, A., & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome biology, 9(2), S8.

Liu, W., S. Cai, B. P Ramesh, G. Chiriboga, K. Knight, and H. Yu. Translating Electronic Health Record Notes from English to Spanish: A Preliminary Study, in Proceedings of the 2015 Workshop on Biomedical Natural Language Processing (BioNLP 2015), pages 134–140, Beijing, China. Association for Computational Linguistics.

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., & Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform, 35(8), 128-144.

Neves, M. (2014). An analysis on the entity annotations in biological corpora. F1000Research, 3.

Neves, M., Yepes, A. J., and Névéol, A. (2016). The Scielo Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), pages 2942–2948. European Language Resources Association (ELRA).

Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E. M., Bui, C., ... & Jimeno-Yepes, A. (2013, September). Entity recognition in parallel multi-lingual biomedical corpora: The CLEF-ER laboratory overview. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 353-367). Springer, Berlin, Heidelberg.

Rodrigo Agerri, Josu Bermudez and German Rigau (2014): "IXA pipeline: Efficient and Ready to Use Multilingual NLP tools", in: Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014), 26-31 May, 2014, Reykjavik, Iceland.

Tiedemann, J. (2009). News from OPUS-A collection of multilingual parallel corpora with tools and interfaces. In Recent advances in natural language processing (Vol. 5, pp. 237-248).

Villegas, M., S. de la Peña, M. Krallinger, A. Intxaurrondo, J. Santamaría. Esfuerzos para fomentar la minería de textos en biomedicina más allá del inglés: el plan estratégico nacional español para las tecnologías del lenguaje, Procesamiento del Lenguaje Natural, Revista nº 59, septiembre de 2017, pp 141-144.

Wu, C., F. Xia, L. Deleger, and I. Solti. Statistical Machine Translation for Biomedical Text: Are We There Yet? AMIA Annual Symposium Proceedings:1290–1299 2011.

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).

Yepes, A.J., É. Prieur-Gaston, and A. Névéol. Combining MEDLINE and publisher data to create parallel corpora for the automatic translation of biomedical text. BMC Bioinformatics, 14(1):146, April. 2013.