An online tool for enhancing NLP of a biomedical corpus[1]

(Antonio Moreno Sandoval)
Computational Linguistics Laboratory
Universidad Autónoma Madrid
Tel. no.: +34 91 4974109
Email: antonio.msandoval@uam.es
(Leonardo Campillos Llanos)
Computational Linguistics Laboratory
Universidad Autónoma Madrid
Tel. no.: +34 91 4974109
Email: leonardo.campillos@uam.es
(Carlos Herrero Zorita)
Computational Linguistics Laboratory
Universidad Autónoma Madrid
Tel. no.: +34 91 4974109
Email: carlos.herrero@uam.es
(José María Guirao Miras)
Department of Computer Languages and Systems
University of Granada, Spain
Tel. no.: +34 958 242813
Email: jmguirao@ugr.es
(Alicia González Martínez)
Instituto de Ingeniería del conocimiento
Universidad Autónoma Madrid
Tel. no.: +34 91 4973710
Email: alicia.gonzález@iic.uam.es
(Doaa Samy)
Department of Spanish
Cairo University
Email: doaasamy@cu.edu.eg
(Emi Takamori)
Computational Linguistics Laboratory
Universidad Autónoma Madrid
Tel. no.: +34 91 4974109
Email: emi.takamori@uam.es

Abstract

This work presents an online interface that allows the user to search words and medical terms in the MultiMedica corpus, which gathers 51,476 texts in Spanish, Japanese and Arabic. In order to develop the tool, several natural language processing (NLP) techniques were applied: firstly, a number of corpora were processed and Part-of-Speech-tagged using morphological analysers for each language; then, the tagged texts were indexed to enhance online queries; thirdly, lists of medical terms were collected for each language. The online tool features word query system, a term query system, and a medical term extractor. The word query system makes it possible to look up items according to word form, lemma, category or string. The medical term query system features an autocomplete function to enhance the input of the query, which is based on the 5000 more frequent terms in the corpus. Finally, the term extractor detects candidate medical terms in an input text, and highlights them according to their level of reliability (medium and high). The list of detected terms is downloadable, and those items that are found in the BabelNet dictionary contain a hyperlink to this resource. The interface is freely available and may be useful for linguists, terminologists, translators and other professionals who work in this domain.

Keywords: Biomedicine, Spanish, Japanese, Arabic, language resource.

1. Introduction

MultiMedica (Multilingual Information Extraction in Health Domain and its application to Scientific and Informative Documents) is a coordinated project between the LABDA research group

[1] This project was funded by the Spanish Government under the grant TIN2010-20644-C03-03.

(UC3M), the GSI group (UPM) and the LLI (UAM). The LLI-UAM team has been in charged of the following tasks:

- Compilation of a specialised corpus of informative texts about health topics. The corpus gathers documents in three languages with different genetic and typological features: Arabic, Japanese and Spanish.
- Morphosyntactic tagging of the corpora.
- Contrastive research on term formation.
- Development of an automatic term extractor.
- Design of a web-based search tool.

This paper presents the online interface for the MultiMedica corpus, which gathers 51,476 biomedical texts in Spanish, Japanese and Arabic (Moreno-Sandoval & Campillos-Llanos, 2013). The tool features two main functions: queries in the medical corpus, and medical term extraction of an input text. The system is freely available at http://cartago.lllf.uam.es/corpus3/index.pl.

The paper is organised as follows. Section 2 summarises the data included in the three subcorpora. Section 3 describes the morphosyntactic annotation of the corpus and the term collection. Section 4 is devoted to the presentation of the methodology for developing the medical term extractor in every language. Section 5, the largest of this paper, shows the web interface to query the corpus and to access the term extractor. Finally, some practical applications are suggested in the conclusions.

2. The data

The MultiMedica corpus (Moreno Sandoval and Campillos Llanos, 2013) is a suitable resource for performing terminological and contrastive linguistic studies. It gathers texts in Spanish, Japanese and Arabic, including different registers and genres (popularisation and technical texts). Table 1 outlines the composition of the corpus (number of texts and words/characters):

Subcorpus	Documents	Word or characters
Japanese	3,746	1,131,304
Arabic	43,526	2,559,323
Spanish	4,204	4,031,174
TOTAL	51,476	7,721,801

Table 1. Summary of the MultiMedica corpus data

The Spanish corpus is made up of three subcollections, each of them reflecting a different type of text. The *Harrison* subcorpus assembles professional and scientific texts written by medical doctors. The *OCU-Salud* subcollection gathers journalistic texts written by medical doctors and edited by journalists. Finally, the *Tu otro médico* subcorpus collects popularization texts from encyclopaedic articles written by professional doctors for non-specialists. Regarding the Arabic corpus, several difficulties were found to gather documents due to the fact that most medical doctors in the Arabic-speaking world write articles in English. Most documents in this subcorpus were articles and popularisation news collected from *Altibbi*, a Jordanian medical website equivalent to Healthline in the United States. The remaining texts were drawn from the health sections of the following journals: *Al-Awsat* (from Saudi Arabia), *Youm7* (from Egypt), and *El Khabar* (from Algeria).
In relation to the Japanese corpus, only abstracts of five medical journals were collected, due, again, to the lack of availability of the data. Nonetheless, texts gather contents on different specialties: Oriental Medicine in Japan (from the journal *Kampo Medicine*), infectious diseases (*Kansenshogaku Zasshi*), liver diseases (*Kanzo*), otolaryngology (*ORLTokyo*) and obstetrics

(*Sanfujinka no shinpo*).

3. Part-of-Speech tagging and creation of lists of medical terms

Several natural language processing (NLP) techniques were undertaken to develop the tool. Firstly, each corpus was processed and Part-of-Speech-tagged. The Spanish subcorpus was tagged by using GRAMPAL (Moreno and Guirao, 2005),[2] a morphological analyser for Spanish with a lexicon of over 50,000 lemmas. The tagging process is automatic, but requires manually revision to ensure annotation quality. To date, two linguists revised the tags corresponding to the popularization texts of the Spanish corpus, even though a further stage of the project envisages revising the technical texts. A random sample representing the 5% of the popularization texts in Spanish was revised twice to compute the inter-annotator agreement (IAA) value. This was assessed by computing the F-measure, as exposed in Hripcsak and Rothschild (2005), and both annotators agreed in about 98% of the texts.

Herrero et al. (2014) explain the methodology followed in the creation of the morphological tagging for the Japanese corpus. After considering three different taggers (ChaSen, Mecab and Juman), we finally chose the last one for the tagging, because Juman[3] provides a good segmentation and a wider range of morphological information. Similarly, the Arabic corpus was automatically annotated using a state-of-the-art PoS tagger, MADA+Tokan (Habash, Rambow and Roth 2009). Then, for all languages, the tagged texts were indexed to enhance online queries.

The next step was to create lists of medical terms for each language. The Spanish list was compiled semi-automatically combining rule-based, tagger-based, and statistical approaches (Moreno-Sandoval et al., 2013). A gold standard list included terms that appeared in leading medical dictionaries (e.g. RANM*,* 2011; Dorland*,* 2005). A silver-standard list gathered terms that were just found in biomedical books and journals.

Regarding Japanese, a single list was compiled with terms from several medical dictionaries: *Online Life Science Dictionary*, 2013, and *Japanese-English-Chinese Dictionary*, 1994. As for the Arabic language, the final list is a combination of full terms translated from English resources (SNOMED and UMLS) and a list of Arabic words equivalent to Spanish prefixes and suffixes, such as *–itis*, *cardio-*, etc. (Samy et al., 2012).

4. Developing a term extractor for each language

One of the project goals was to offer a medical term extractor. This functionality required a different procedure for each language. The Spanish extractor a uses lists of terms, medical roots and affixes, the GRAMPAL tagger, and rules for multi-words and context patterns (Campillos Llanos et al., 2013). The processing of the input text to detect candidate terms is as follows. First, a dictionary-based method that relies on pattern matching is applied. Each item found in the gold standard list will be marked as a highly reliable candidate term (e.g. *pulmón,* 'lung'). Likewise, each term found in the silver standard list will be selected as a medium reliable candidate term (e.g. *secundario,* 'secondary'). In a third stage, those words that were not found in any list are PoS-tagged through the GRAMPAL tagger. Unrecognised items (i.e. words not included in the lexicon of the tagger, which was designed for the general language) are then filtered using a list of biomedical roots and affixes (e.g. *hemat(o)-,* an affix related to blood). In this way, for example, an adverb such as *hematológicamente* ('hematologically') may be recognised as a term and highlighted with medium reliability. The last stage involves applying multi-word formation rules to the previous list of candidate terms. If any element of the multi-word candidate term has medium reliability, the whole unit will be highlighted as such. For example, if the term *complejo* ('complex', medium reliability) and *amigdalino* ('tonsillar', high reliability) are recognised, a multi-word rule will join both terms in *complejo amigdalino* ('tonsillar complex') and mark it as a medium reliability candidate term. Figure 1 outlines the architecture of the system.

[2] http://www.lllf.uam.es/ING/Grampal.html

[3] http://nlp.ist.i.kyoto-u.ac.jp/EN/index.php?JUMAN

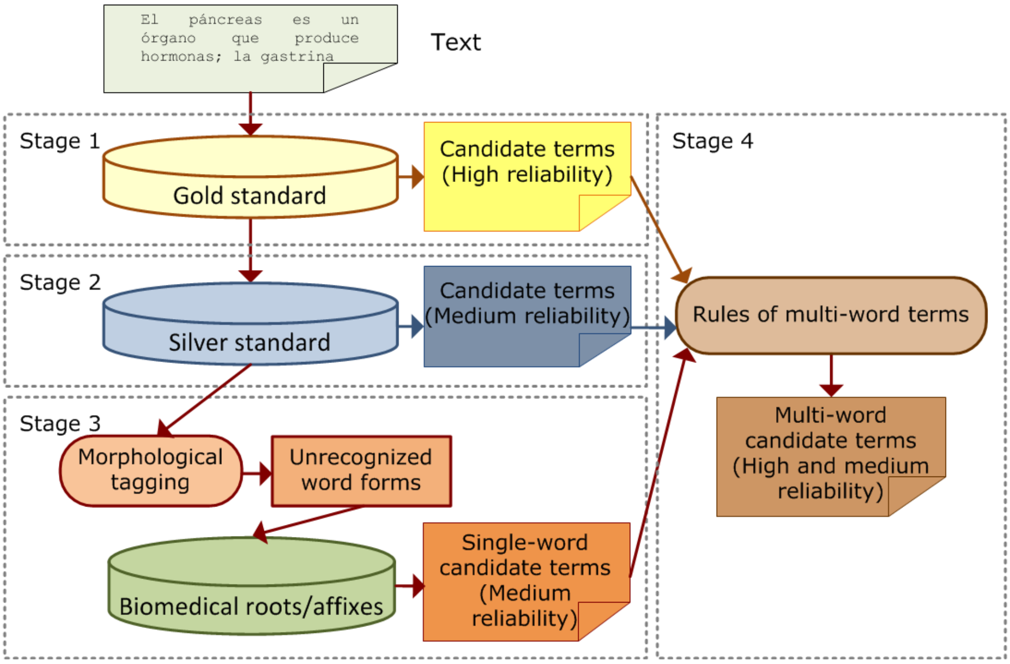

Figure 1. Architecture of the Spanish term extractor

The extractors for Japanese and Arabic follow a simpler procedure. Regarding Japanese, the extractor performs an initial pattern matching throughout the dictionary, identifying those terms as highly reliable. Secondly, a series of rules are applied bearing in mind the agglutinative nature of the language. For example, if two dictionary terms are joined with a connective particle it will be considered as a single multi-word term; also, if additional kanji characters are added to the initial or the final part of a dictionary term, the extractor would recognise the whole string of characters as a single term. The terms detected using this rule-based procedure would be classified as medium reliable. The Arabic language is mainly a dictionary-based extractor, which recovers terms from the medical list created for this purpose.

The term extraction has room for improvement in a future stage of the project by including more medical terms, or codes from the International Classification of Diseases version 9 (ICD-9)[4] or the Unified Medical Language System (UMLS) and the Systematized Nomenclature of Medicine--Clinical Terms (SNOMED-CT)[5].

5. The web interface

Users can perform queries in the corpus in two ways: word search (“Search” tab, “Consulta” in the Spanish version) and Medical term search (“Medical Term Search” tab, “Consulta de Términos Médicos” in Spanish). In addition, users can input a free text to detect and extract candidate terms in the domain (“Medical Term Extractor”, “Extractor de Términos Médicos”). This section will first explain how the general word search works (Section 5.1), the medical term search (Section 5.2), and finally, the term extractor system (Section 5.3).

5.1. Word search

Any word in the corpus can be searched according to form, lemma, or Part-of-Speech (PoS). For example, if the user inputs the lemma *cáncer,* the results may be *cáncer* or *cánceres* (respectively, ‘cancer’ or ‘cancers’). The user has the option to look up the collocations of the word (Figure 2) as well as its frequency and log-likelihood value (Dunning, 1993).

---

[4] A Spanish version of the ICD-9 is accessible through the web of the Ministry of Health (http://eciemaps.mspsi.es/ecieMaps/browser/index_9_mc.html)

[5] http://www.ihtsdo.org/snomed-ct/

		Frecuency	Log-likelihood
cáncer de	mama	269	3986
cáncer de	próstata	124	1717
cáncer de	pulmón	128	1592
cáncer de	colon	76	818
cáncer de	ovario	43	540
cáncer de	páncreas	42	467

Figure 2. Most frequent collocations and log-likelihood values

In the search results, frequency values are normalized per million words (hereafter, *pmw*). Counts are also compared to the frequencies in the Corpus de la Real Academia Española (CREA) corpus. This makes it possible to know the *distinctiveness* of the searched word when looked up in a specialized corpus and in relation to a general language corpus. For example, when the word *hepatitis* is searched, the normalised frequency in the MultiMedica corpus is 385.8 pmw, and 6.1 pmw in the CREA corpus. This shows that this token is highly related to this specialised genre. In contrast, if *corazón* ('heart') is searched, the normalised frequency in the MultiMedica corpus drops to 140.8 pmw, which is close to the normalised frequency in the CREA corpus (125.3 pmw). This indicates that *corazón* appears with a similar frequency in a health and a general corpus. Since this is a polysemous word, other senses beyond the anatomical context are used in the general language (e.g. related to feelings, or as a synonym of 'nucleus' or 'core'). Figure 3 shows an example of the search function.

Figure 3. Search results for *corazón* (heart) with normalized frequencies

The search tool for the Spanish corpus also provides information about word distribution (i.e. its frequency in each type of text). This feature makes it possible to compare different text genres (popularization vs. technical documents). For example, if we search for *dolor de espalda* ('upper back pain'), the results show that this term is more frequent in popularization texts (*OCU* subcorpus, the bar on the right side in Figure 4) than in technical texts (*Harrison,* the bar on the left side; note that figures are computed in words per million).

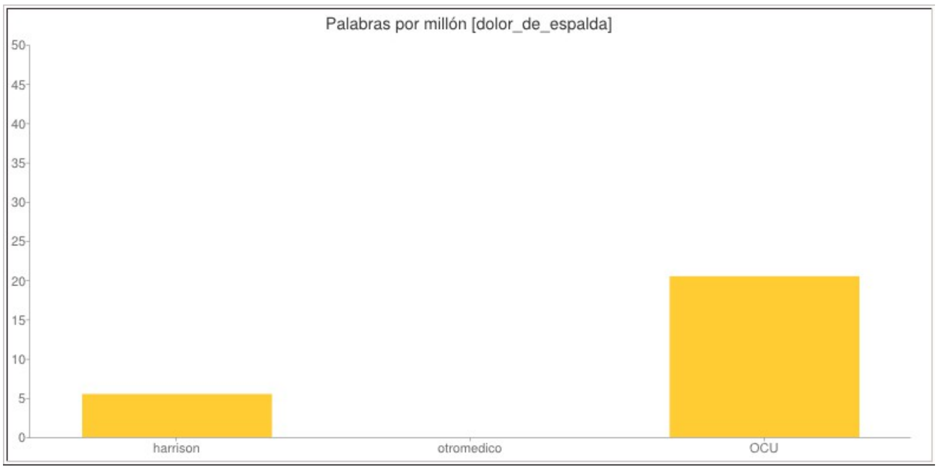

Figure 4. Distribution of *dolor de espalda* in the Spanish text sub-collections

However, when we search for *dorsalgia* (the technical synonym of 'dolor de espalda'), the results reveal that this term is restricted to technical documents (Figure 5).

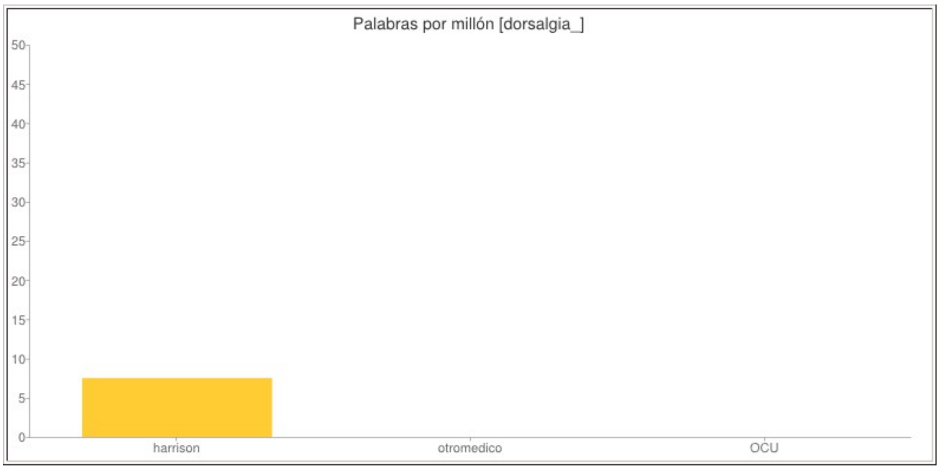

Figure 5. Distribution of *dorsalgia* in the Spanish text sub-collections

With regard to Arabic, Figure 6 is a screen dump of the word search for the Arabic texts. The image shows the results for سرطان الثدي ('breast cancer').

Figure 6. Search results for سرطان الثدي ('breast cancer') in the Arabic corpus

5.2. Medical term search

The medical term search allows users to look up the most frequent medical terms in the corpus. When a user is typing a query, an auto-complete function provides a list of all the possible terms that contain the typed letters. The list is based on the 5000 more frequent terms in the corpus. Figures 8 and 9 show, respectively, examples of the autocomplete function for Spanish and Japanese for the search item hepatitis ('hepatitis') and 乳癌 ('breast cancer').

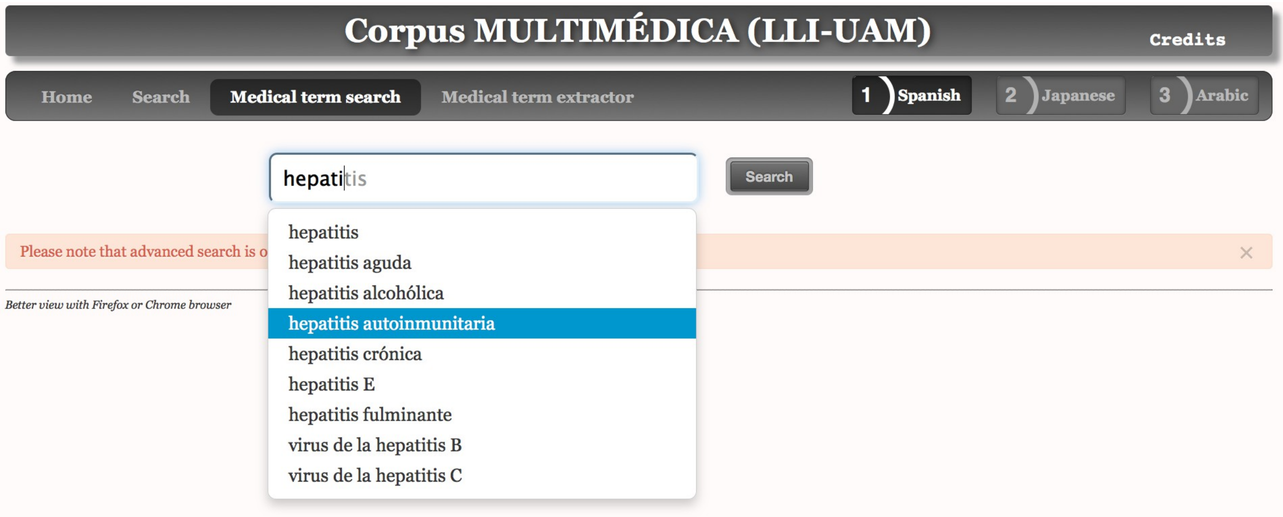

Figure 8. The auto-complete function for the term search in the Spanish corpus

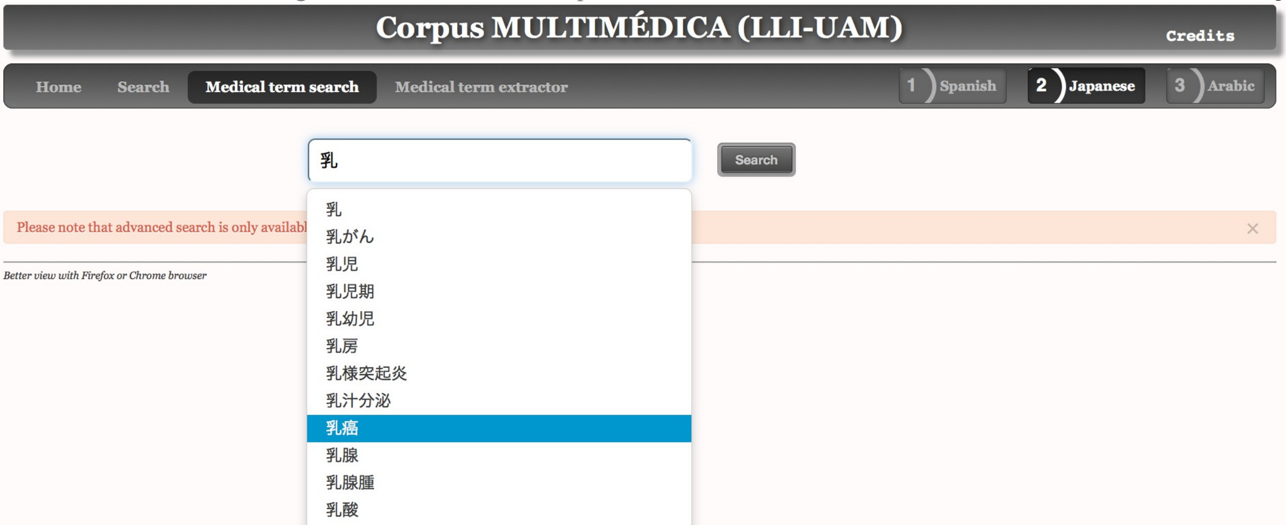

Figure 9. The auto-complete function for the term search in the Japanese corpus

5.3. Medical term extractor

The medical term extractor detects candidate terms from an input text (Figure 10). The tool highlights medical terms according to their level of reliability (medium and high, Figure 11). The user may also download the term list in text format for further use (see the red button in Figure 12). In addition, terms that are found in the BabelNet dictionary (Navigli and Ponzetto, 2012) contain a hyperlink to this resource, which provides their translation in many languages.

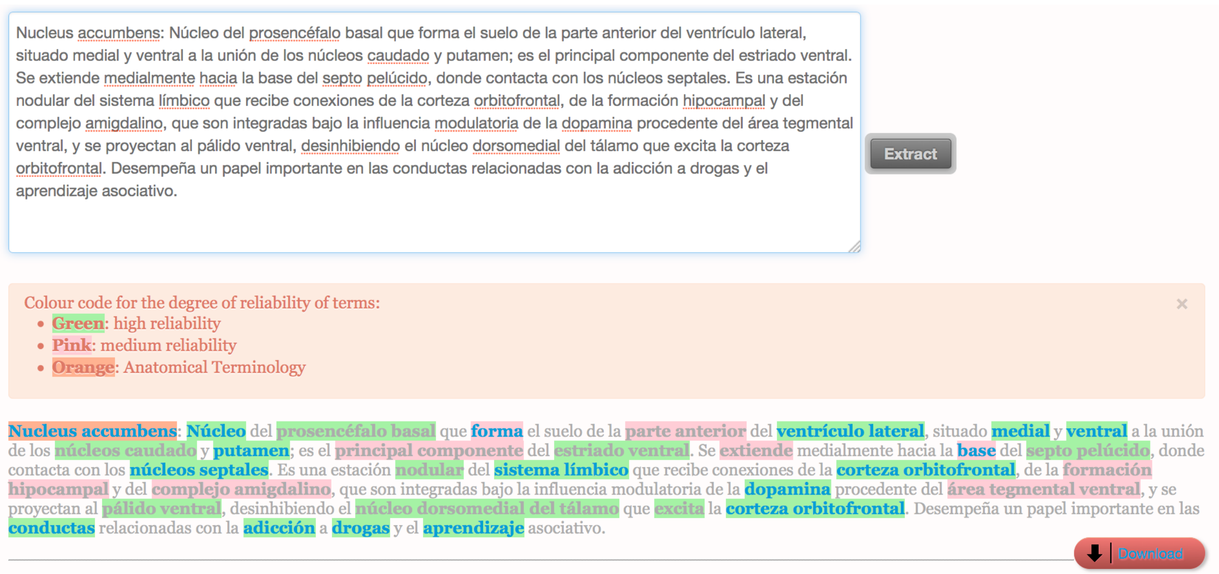

Figure 10. The medical term extractor for Spanish texts

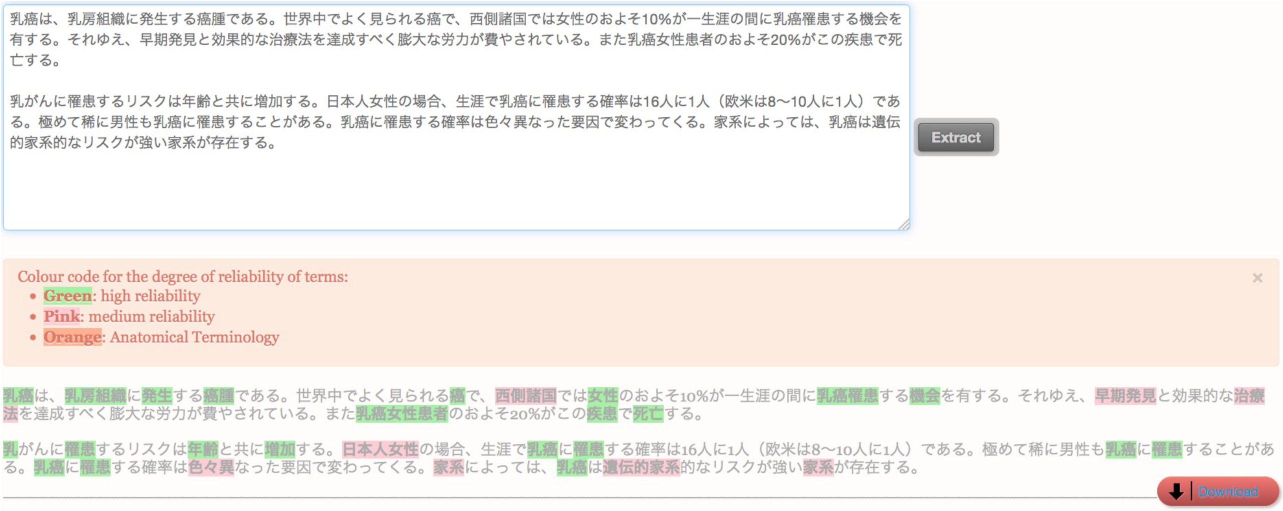

Figure 11. A screenshot of the Japanese term extractor

6. Conclusions

This paper has summarised the methodology followed in the creation of a multilingual corpus of medical texts, their morphological annotation and further indexation, the term list extraction and the development of an online tool for the user to obtain information from it. Since the three languages selected were so different genetically and typologically, we have had to choose specific approaches and tools for each of them. During the three and a half years of the project, we have identified the main problems for the computational treatment of medical terms in these languages. Among them, Arabic is notable for its lack of language resources in medical NLP (from texts to electronic dictionaries). To our knowledge, MultiMedica is a pioneer effort in the topic and for this combination of languages.

The project has also provided an interesting typological insight on how languages behave in relation to the medical domain. Each of our three languages provides different challenges when developing the extractor: the variation in inflection of Spanish terms, variation in the Arabic writing system and the segmentation due to the lack of white spaces between words in Japanese. Even though the initial steps of creating the corpus, tagging, and development of a medical term list was equal in the three languages, the processing of the texts and creation of the extractor had to be adapted to the demands of each language.

We believe that the corpus and online tools may provide the users with a good amount of data for future linguistic research on the biomedical discourse. The term extractor may fulfil terminologists' and translators' needs and help them identify term candidates and find their equivalents in other languages. In addition, health professionals and medical students could make use of this interface to seek and translate biomedical information online.

References

Campillos Llanos, L., A. Moreno Sandoval, and J.M. Guirao. 2013. "An automatic term extractor for biomedical terms in Spanish." In *Proceedings of the 5th International Symposium on Languages in Biology and Medicine (LBM 2013)*. 12th and 13th December 2013. Tokyo, Japan.

Campillos Llanos, L., A. Moreno Sandoval, A., and J.M. Guirao. 2014. "Combined strategies for automatic term recognition in Spanish." (under review).

Dorland. 2005. *Diccionario enciclopédico ilustrado de medicina*. 30th edition (Spanish version). Madrid: Elsevier, D. L.

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

Japanese-English-Chinese dictionary. 1994. Tokio: 朝倉書店 (Asakura Shoten)

Habash, N., O. Rambow, and R. Roth. 2009. "Mada+Tokan: A toolkit for Arabic tokenization, diacritization, morphological disambiguation, POS tagging, stemming and lemmatization." *Proceedings of the 2nd International Conference on Arabic Language Resources and Tools (MEDAR).* Cairo, Egypt, 242-245.

Herrero Zorita, C., L. Campillos Llanos and A. Moreno Sandoval. 2014. "Collecting a POS-tagging a lexical resource of Japanese biomedical terms from a corpus." *Procesamiento del Lenguaje Natural*. 52: 29-36.

Hripcsak, G. and A. S. Rothschild. 2005. "Agreement, the F-measure, and reliability in information retrieval." *Journal of the American Medical Association* 12: 296-298.

Moreno-Sandoval, A., and J.M. Guirao. 2006. "Morpho-syntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation." Y. Kawaguchi, S. Zaima, and T. Takagaki, eds. *Spoken Language Corpus and Linguistic Informatics*. Amsterdam/Philadelphia: John Benjamins. 199–218.

Moreno-Sandoval, A., L. Campillos-Llanos, A. González Martínez, and J. M. Guirao. 2013. "An affix-based method for automatic term recognition from a medical corpus of Spanish." In A. Hardie and R. Love (eds.) *Proceedings of the 7th Corpus Linguistics Conference 2013. Lancaster University (United Kingdom)*, 214-217. 23rd-26th July 2013. Lancaster: UCREL.

Moreno-Sandoval, A., and L. Campillos-Llanos. 2013. "Design and Annotation of MultiMedica – A Multilingual Text Corpus of the Biomedical Domain." *Procedia - Social and Behavioral Sciences*, 95 (25): 33-39.

Navigli, R., and S. Ponzetto. 2012. "BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network." *Artificial Intelligence*, 193: 217-250. http://babelnet.org.

Online Life Science Dictionary (ライフサイエンス辞書オンラインサービス). 2013. Available at: http://lsd.pharm.kyoto-u.ac.jp/ja/service/weblsd/index.html. (accessed 21 May 2014)

Real Academia Nacional de Medicina . 2011. *Diccionario de términos médicos.* Madrid: Editorial Médica Panamericana.

Samy, D., A. Moreno Sandoval, C. Bueno-Díaz, M. Garrote-Salazar and J.M. Guirao. 2012. "Medical Term Extraction in an Arabic Medical Corpus." *Proceedings of the 8th Language Resources and Evaluation Conference,* 640-645. Istanbul: LREC.