Available online at www.sciencedirect.com

ScienceDirect

Procedia - Social and Behavioral Sciences 95 (2013) 33 – 39

Procedia
Social and Behavioral Sciences

5th International Conference on Corpus Linguistics (CILC2013)

Design and Annotation of MultiMedica – A Multilingual Text Corpus of the Biomedical Domain

Antonio Moreno-Sandoval*, Leonardo Campillos-Llanos

Computational Linguistics Laboratory. Department of Linguistics. Faculty of Arts and Humanities, Autonomous University of Madrid, c/Francisco Tomás y Valiente, 1. 28049, Madrid, Spain

Abstract

This article describes the MultiMedica corpus, a multilingual collection of Spanish, Japanese, and Arabic texts from the biomedical domain. This novel combination of languages has been chosen with two purposes: the contrastive study of three languages that are typologically and genetically different, and the creation of a gold standard to develop and evaluate an Automatic Term Recognition (ATR) system. A total of 51,476 documents have been collected from the Web, and the corpus contains over seven and a half million words. Most documents were written by medical doctors and edited by journalists for the general public. Each text has been tagged for Part-of-Speech and indexed in an Information Retrieval system and a concordance interface that is aimed at students of Translation, Medicine, and Medical Humanities.

Keywords: Biomedical discourse; Text corpus; Terminology; Spanish; Arabic; Japanese.

1. Introduction

Publication of scientific papers, articles and news on biomedical research has rocketed in the last few years. This huge amount of data provides a "raw material" easily available to linguists or scholars interested in scientific discourse or technical terminology. Nonetheless, collections of texts from the health and biomedical domains specifically gathered for corpus and linguistic analyses are scarce. Most collected corpora are aimed at Bio Natural Language Processing (BioNLP) tasks (Ananiadou & McNaught, 2006; Bretonnel Cohen, 2010), thus including

* Corresponding author: Tel.: +34-91-497-5250; fax: +34-91-497-4498.
E-mail address: antonio.msandoval@uam.es

1877-0428 © 2013 The Authors. Published by Elsevier Ltd.
Selection and peer-review under responsibility of CILC2013.
doi:10.1016/j.sbspro.2013.10.619

specific annotation for Biomedical Entities (e.g. GENIA, GENETAG, PennBioIE, or DDI corpus) or coreference and anaphora resolution (e.g. MEDCo or MEDSTRACT corpus)[1]. In addition, corpora with documents from these areas in languages other than English are even more rare. For Spanish, a team from the Consejo Superior de Investigaciones Científicas (the Spanish National Research Council) has developed the Iberia corpus (Ahumada *et al*., 2011), which assembles over 8,000 texts and more than 88 million words. However, its design is focused not only on biomedical topics, but also in other subjects from the scientific and technical discourse (e.g. Engineering, Mathematics, Agronomy, or Social Sciences). Moreover, it is not available to the general public.

The aim of this article is to describe a linguistically designed resource, the MultiMedica corpus, and to present its research applications. This corpus has been collected for the homonymous project[2] (Martínez *et al*., 2011) by the Computational Linguistics Laboratory at the Autonomous University of Madrid (LLI-UAM). It is a specialized comparable corpus containing texts about health and biomedical topics written in three typologically and genetically different languages: Spanish, Arabic, and Japanese.

The article is organized as follows. First, we will introduce the general design of the corpus and the text collection procedures (Section 2), and we will describe the resource for each language (Spanish, Arabic, and Japanese) in different subsections. Second, we will address the structure, the coding, and the Part-of-Speech (PoS) tagging of the corpus files (Section 3). Finally, we will outline the current results of the project and the future work (Section 4).

2. Description of the corpora

Before entering into the design of the corpus, we shall make clear what *biomedical domain* means in our project. Biomedicine is a wide area of research ranging from Pharmacology and Biochemistry to Genetics and Microbiology—or even Ethics in Medicine. We apply the general label *biomedicine* to our corpus because, due to the heterogeneous nature of our collection, the type of text and its contents cannot be easily classified into one of the above-mentioned fields. In fact, our texts may also include scientific concepts and terms from other disciplines, such as Statistics, Botany, Zoology, or Environmental Sciences (e.g. in relation to the toxic venom of an animal or a plant, or the consequences of pollution for health).

When designing the corpus, our priority was to obtain comparable texts for the three languages rather than to look for a specific content or biological discipline. Assembling electronic texts from the health and biomedical domains is a relatively easy task for Spanish and Japanese—there are many specialized sources for these languages. However, it is not that easy for Arabic language, since most Arabic speaking doctors write their scientific articles in English or French. Thus, for these domains we focused on divulgation texts published on reliable websites and health sections from general newspapers and magazines.

The corpus was gathered using general text crawling procedures. In our project, we used Wget to capture documents after analyzing the structure of each resource.

The corpus comprises a total of 51,476 documents and over seven and a half million words on the whole. Nonetheless, the representation of each language is not balanced: there are four million words for Spanish, two and a half million for Arabic, and over one million for Japanese (Table 1). Most documents have been written by medical doctors and edited by journalists and therefore the texts are not exceedingly technical. Overall, the corpus is a suitable resource for performing terminological and contrastive linguistic studies.

---

[1] Kim & Tsujii (2006) outline a comprehensive panorama of resources, and the DDI corpus is explained in Herrero Zazo *et al*. (2013).

[2] The MultiMedica project is a coordinated project between the University Carlos III of Madrid (UC3M), the Technical University of Madrid, and the Autonomous University of Madrid. The project aims at performing research on information extraction techniques for natural language processing in the biomedical domain. The results of this project include a generic search tool for looking up information about diseases and drugs (Sánchez Cisneros *et al*., 2012). Further information on the project is available at: http://labda.inf.uc3m.es/multimedica/.

Table 1. Summary of the corpus.

Corpus	Documents	Words or characters
Japanese	3,746	1,131,304
Arabic	43,526	2,559,323
Spanish	4,204	4,031,174

2.1. The Spanish corpus

The Spanish corpus consists of three subcollections, each of them reflecting a different type of medical text. The sources of each sub-collection are:

- *Harrison*: the online edition of a comprehensive textbook for students of Medicine. This subcorpus gathers more than 3,800 professional and scientific texts written by medical doctors. Our team manually revised these texts to avoid repetition of documents.
- *OCU-Salud*: a magazine that publishes journalistic texts written by medical doctors and edited by journalists. This subcollection assembles popularization texts for the general public.
- *Tu otro médico*: a website that includes encyclopaedic articles written by professional doctors for non-specialists. This subcollection mainly gathers divulgation texts.

In total, the Spanish corpus covers 4,200 texts and over 4 million words (Table 2), and reflects a balanced combination of most medical specialities.

Table 2. Description of the Spanish corpus.

Spanish corpus	Documents	Words
Harrison	3,841	3,696,484
OCU-Salud	297	310,894
Tu otro médico	66	23,796
Total	4,204	4,031,174

2.2. The Arabic corpus

The Arabic corpus comprises documents from Altibbi, a Jordanian medical website equivalent to Healthline in the United States. This resource contains medical articles and divulgation news, with a certain degree of control by medical doctors. Altibbi has been the main source of the documents included in the Arabic corpus (43,278 files). Other texts were drawn from the health sections of journals published in three dialect areas in the Arabic world: *Al-Awsat* (Saudi Arabia), *Youm7* (Egypt), and *El Khabar* (Algeria) (Table 3).

Table 3. Description of the Arabic corpus.

Arabic corpus	Documents	Words
Altibbi	43,278	2,460,733
Alawsat	68	58,610
Youm7	83	18,948
ElKhabar	97	21,032

2.3. The Japanese corpus

The Japanese corpus collects abstracts from medical journals on different specialties, from Oriental Medicine to Obstetrics and Gynecology. The total corpus size, which is expressed in Japanese characters (kanji and kana), is 1,131,304 characters (Table 4), as counted by ChaSen (Section 3).

Table 4. Description of the Japanese corpus.

Japanese corpus	Documents	Characters
Kampo Medicine (Oriental Medicine in Japan)	719	214,757
Kansenshogaku Zasshi (Infectious diseases Journal)	858	244,879
Kanzo (Liver diseases Journal)	1,446	432,674
ORLTokyo (Japanese Otolaryngology)	623	203,705
Sanfujinka no shinpo (Advances in Obstetrics)	100	35,289

3. Structure and tagging of the corpus files

Files are coded in XML (W3C, 2012). This markup language is very suitable to easily represent the data needed for our processing tasks. The structure is the same for all files:

- Metadata: they include the following information (see Figure 1):
  - File ID.
  - Source (title of the book/journal, country…).
  - Language and language variety.
  - Production date.
  - Title of the file.
  - Author and information about copyright.
- Text (structured in paragraphs).

```
<?xml version="1.0" encoding="utf-8"?>
<article>
<header>
<file_id>MED-otromedico-17</file_id>
<original_file_path_and_name>www.tuotromedico.com/temas/reflujo.htm.html</original_file_path_and_name>
<source type="online resource">
          <source_name>Tu otro médico</source_name>
          <source_country>Spain</source_country>
          <source_place type=""></source_place>
          <source_place type=""></source_place>
          <source_section></source_section>
</source>
<language>Spanish</language>
<language_variety>Spanish</language_variety>
<production_date>
          <month>01</month>
          <day>01</day>
          <year>2011</year>
          <isodate>20110101</isodate>
</production_date>
<category>Medicine</category>
<title>REFLUJO GASTROESOFÁGICO</title>
<author></author>
<token_number>397</token_number>
<copyright type="privative"></copyright>
<retrieved>
          <retrieved_date>04/04/2011</retrieved_date>
          <retrieved_by>Conchi</retrieved_by>
</retrieved>
</header>
```

Fig. 1. Sample of XML code for the metadata.

The three subcorpora have been automatically annotated using state-of-the-art PoS taggers. For Arabic, MADA+Tokan (Habash, Rambow, & Roth, 2009) was used. For Japanese, we made use of ChaSen, which is provided in The Sketch Engine (Kilgarrif *et al*., 2004). As for Spanish, files were tagged by means of GRAMPAL morphological analyser (Moreno & Guirao, 2006).

We have used a tool developed for the MultiMedica project[3] to search for medical terms in the three corpora (Spanish, Japanese, and Arabic). Figure 2 below shows a screenshot of the MultiMedica tool, and the results obtained for *cardiopatía* ('cardiopathy'). An enhanced web-based interface, which includes a searching option for lemmas and PoS tags, is currently under development at LLI-UAM.

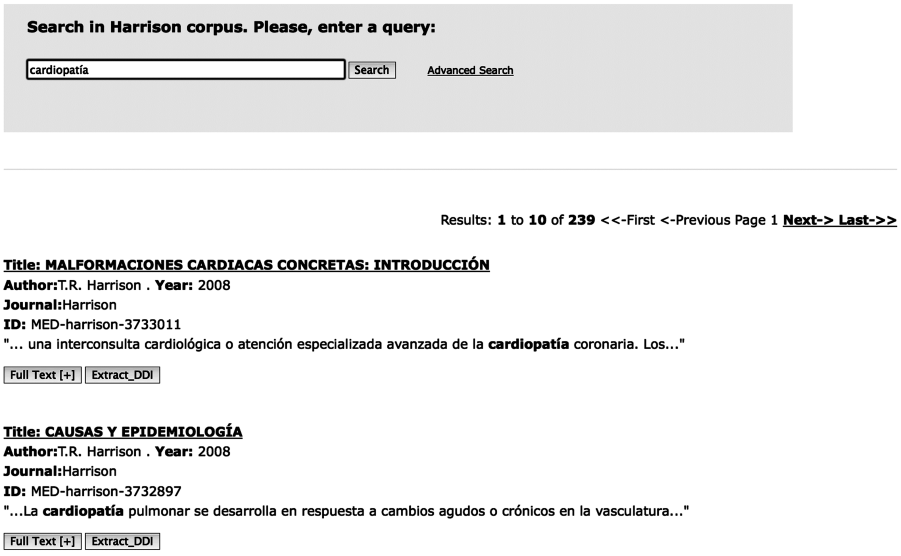

Fig. 2. Screenshot of the MultiMedica tool.

4. Results and on-going research

The goal of the project is twofold: first, to build a tagged collection of comparable texts from the biomedical domain in the three languages; and second, to produce a gold standard of medical terms for developing and evaluating an Automatic Term Recognition system and other tools for Information Retrieval.

So far, the following tasks have been completed:

- Compilation and morphosyntactic tagging of the texts in the three languages, which has been fully accomplished.
- An analysis of biomedical specialized discourse based on this corpus and from a contrastive perspective, comparing the three languages (Samy *et al*., 2012). This research showed that the strategies of biomedical term formation—such as the combination of medical roots and affixes, compounding, or derivation—seem to be quite similar across different languages, regardless of their morphological typology and their writing system (see examples; we use Hans Wehr transliteration for Arabic):

[3] The concordance tool to search the corpora is already available at: http://labda.inf.uc3m.es/multimedica/prototypes.html

Spanish:	***cardio-***	***cardio****patía*	*mio****cardio***
Japanese:	心臓	心疾患	心筋
	shin-zou	**shin**-shikkan	**shin**-kin
Arabic:	القلب	أمراض القلب	عضلة القلب
	al **qalb**	amrāḍ al **qalb**	‘aḍla al **qalb**
	‘heart’	‘heart disease’	‘heart muscle’

- Experimentation on ATR in biomedical texts, which we have started for the development of an ATR tool to teach translation and terminology. We have already carried out an experiment on recognition of candidate terms in Spanish (further details are explained in Moreno-Sandoval *et al*., 2013). This experiment uses a list of biomedical roots and affixes (e.g. *-asa*, ‘-ase’, for an enzyme) together with three different statistical and rule-based procedures[4].
- Semi-automatic extraction of lists of terms from the health and biomedical domain. In a first step, we applied the above-mentioned statistical and rule-based procedures to detect candidate terms. In a second step, we verified the extracted items in specialized medical dictionaries (e.g. the Spanish edition of *Dorland's Illustrated Medical Dictionary,* 2005; or *Diccionario de términos médicos,* by Real Academia Nacional de Medicina, the Spanish Royal Academy of Medicine, 2011). The final outcome is a gold standard of biomedical terms in Spanish. This list contains over 24,000 word forms and will be used in an Automatic Term Recognition system under development.

By the end of 2013, an Information Retrieval (IR) system to search for terms and tagged structures will be available from the LLI-UAM website. This tool is aimed at students of Linguistics, Translation, Medicine, and Medical Humanities, who will have at their disposal thousands of samples of terms in their context of use.

Given that these kinds of language resources are scarce, we hope that this project will contribute to the contrastive and terminological analysis in these languages. As a matter of fact, we are devising future research directions that will take advantage of the results obtained. First, the list of candidate terms extracted can be of interest to terminologists and lexicographers, since new items could be incorporated to medical dictionaries. Secondly, the corpus could be augmented with texts from other languages or even other text genres from the health domain (e.g. abstracts from medical articles). Finally, coding biomedical entities or tagging discourse phenomena such as anaphora and coreference may also be of particular interest in a further stage of the corpus.

Acknowledgements

This research has been supported by the Spanish Government (under the grant TIN2010-20644-C03-03) and by the Madrid Regional Government (grant MA2VICMR).

References

Ahumada, I., Porta Zamorano, J., & Rosal García, E. (2011). Design and development of Iberia: a corpus of scientific Spanish. *Corpora,* 6(2), 145-158. More information about Iberia is available at: http://www.investigacion.cchs.csic.es/elci/node/8

[4] We refer to review articles and manuals (e.g. Kageura & Umino, 1996; Kilgarriff, 1996; Cabré *et al*., 2001; Krauthammera & Nenadic, 2004; Ananiadou & McNaught, 2006) for further information about research on ATR.

Ananiadou, S., & McNaught, J. (2006). Introduction. In S. Ananiadou & J. McNaught (Eds.) *Text Mining for Biology and Biomedicine* (pp. 1–11). Boston, MA: Artech House.

Bretonnel Cohen, K. (2010). BioNLP: Biomedical Text Mining. In N. Indurkhya, & F. J. Damerau (Eds.), *Handbook of Natural Language Processing* (pp. 605–625). 2nd edition. Boca Raton: CRC. Chapman & Hall/CRC Machine Learning & Pattern Recognition series.

Cabré, Mª. T., Estopá, R., & Vivaldi, J. (2001). Automatic term detection: A review of current systems. In D. Bourigault, C. Jacquemin, & M.-C. L'Homme (Eds.), *Recent Advances in Computational Terminology*. *Natural Language Processing* (pp. 53–87), vol. 2. Amsterdam: John Benjamins.

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

Habash, N., Rambow, O., & Roth, R. (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 Zazo, M., Segura-Bedmar, I., & Martínez Fernández, P. (2013). Corpus DDI: Un corpus anotado con fármacos e interacciones farmacológicas. *Proceedings of the 5th International Conference on Corpus Linguistics, Alicante, Spain. March, 2013*.

Kageura, K., & Umino, B. (1996). Methods of automatic term recognition: A review. *Terminology*, 3(2), 259–289.

Kilgarriff, A. (1996). Which words are particularly characteristic of a text? A survey of statistical approaches. *Proc. AISB Workshop on Language Engineering for Document Analysis and Recognition, Sussex University, April 1996*, 33–40.

Kilgarriff, A., Rychly, P., Smrz, P. & Tugwell, D. (2004). The Sketch Engine. *Proceedings of EURALEX 2004, Lorient, France*, 105–116. http://www.sketchengine.co.uk [Accessed: 27/04/2013]

Kim, J-D. & Tsujii, J. (2006). Corpora and Their Annotation. In S. Ananiadou & J. McNaught (Eds.), *Text Mining for Biology and Biomedicine* (pp. 179–211). Boston, MA: Artech House.

Krauthammer, M., & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Biomedical Informatics*, 37, 512–526.

Martínez, P. González Cristobal, J.C., & Moreno-Sandoval, A. (2011). MULTIMEDICA: Extracción de información multilingüe en Sanidad y su aplicación a documentación divulgativa y científica. *Revista Española para el Procesamiento del Lenguaje Natural*, 47, 347–348.

Moreno-Sandoval, A., & Guirao, J. M. (2006). Morphosyntactic Tagging of the Spanish C-ORAL-ROM Corpus: Methodology, Tools and Evaluation. In Y. Kawaguchi, S. Zaima, & T. Takagaki (Eds.), *Spoken Language Corpus and Linguistic Informatics*. (pp. 199–218). Amsterdam/Philadelphia: John Benjamins.

Moreno-Sandoval, A., Campillos-Llanos, L., González-Martínez, A., & Guirao-Miras, J. Mª. (2013). An affix-based method for automatic term recognition from a medical corpus of Spanish. *Proceedings of VII International Corpus Linguistics Conference 2013. Lancaster, United Kingdom, 23rd-26th July 2013.* (Accepted)

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

Samy, D., Moreno-Sandoval, A., Bueno-Díaz, C., Garrote-Salazar, M., & Guirao, J. Mª. (2012). Medical Term Extraction in an Arabic Medical Corpus. *Proceedings of the 8th Language Resources and Evaluation Conference 2012, May 2012. Istanbul, Turkey*. http://www.lrec-conf.org/proceedings/lrec2012/pdf/597_Paper.pdf [Accessed: 27/04/2013]

Sánchez-Cisneros, D., Lana, S., Moreno-Sandoval, A., Campillos-Llanos, L., Martínez-Fernández, P., & Segura-Bedmar, I. (2012). Prototipo buscador de información médica en corpus multilingües y extractor de información sobre fármacos. *Revista Española para el Procesamiento del Lenguaje Natural*, 49, 209–212.

References on the Internet

ChaSen. http://chasen-legacy.sourceforge.jp

Fauci, A. S., Braunwald, E., Kasper, D. L., Hauser, S. L., Longo, D. L., Jameson, J. L., & Loscalzo, J. (eds.) (2008) *Harrison Principios de Medicina Interna.* 17th edition. New York: McGraw-Hill. Online edition: http://www.harrisonmedicina.com/

GRAMPAL. http://www.lllf.uam.es/ING/Grampal.html

OCU-Salud Journal. http://www.ocu.org/ocu-salud/

The Sketch Engine. http://www.sketchengine.co.uk

Tu otro médico Journal. http://www.tuotromedico.com/

Wget. www.gnu.org/software/wget/

World Wide Web Consortium (W3C) (2012). Extensible Markup Language (XML). http://www.w3.org/XML/