Using semi-automatic compiled corpora for medical terminology and vocabulary building in the healthcare domain

Rut Gutiérrez-Florido [rgut@uma.es]
Gloria Corpas-Pastor [gcorpas@uma.es]
Miriam Seghiri [seghiri@uma.es]

Abstract

English, Spanish and German are amongst the most spoken languages in Europe. Thus it is likely that patients from one EU member state seeking medical treatment in another will speak or understand one of these. However, there is a lack of resources to teach efficient communication between patients and medics. To combat this, the TELL-ME project will provide a fully targeted package. This includes learning materials for Medical English, Spanish and German aimed at medical staff already in the other countries or undertaking cross-border mobility. The learning process will be supported by computer-aided tools based on corpora. For this reason, in this workshop we present the semi-automatic compilation of the TELL-ME corpus, whose function is to support the e-learning platform of the TELL-ME project, together with its self-assessment exercises emphasising the importance of specialised terminology in the acquisition of communicative and language skills.

1 Introduction

When a medical emergency takes place, the initial moments are crucial. Being aware of this problem, the TELL-ME project (ref. no. 517937-LLP-2011-UK-LEONARDO-LMP) has developed an e-learning platform in order to help medical professionals acquire a specific language knowledge, which allows them to understand patients who do not share their mother tongue in first contact situations. Consequently medical professionals can determine signs and symptoms of different illnesses quickly and effectively.

The TELL-ME e-learning platform is based on Natural Language Processing and corpus linguistics which combined establish the foundations of the resources that will help medical professionals to better understand their foreign patients.

In this workshop, we shall present a methodology for the semi-automatic compilation of a medical multilingual corpus within the TELL-ME project —named the TELL-ME corpus— using a tool called *BootCat*. This corpus will be used to support the generation of self-assessment exercises, medical dictionary and the intelligent dictionary assistant; and is also expected to highlight the importance of terminology within specialised medical communication and doctor-patient interaction.

2 Compilation protocol for the TELL-ME corpus

Before outlining the specific protocol for the semi-automatic compilation of the TELL-ME corpus, it is necessary to describe the design criteria.

2.1. Design criteria

According to the aims[1] of the TELL-ME project, UMA is going to create a virtual, i.e. based on resources available on-line, comparable and trilingual corpus (English, German and Spanish). It will made up of three different types of articles: specialised, semi-specialised and informative, in order to have the three levels of communication for future exercises; and 11 medical specialties:[2] first contact, emergency, internal medicine, ophthalmology, general surgery, cardiac/vascular

[1] For further information, see <http://tellme-project.eu/about-tell-me>.

[2] The three medical partners of the Consortium (Health Education West Midlands from the UK, Hospital F.A.C. Dr. Pascual from Spain and Universitätsmedizin Mannheim from Germany) provided these 11 specialties as the more common in first contact situations.

surgery, plastic surgery, traumatology, otorrinolaryngology, urology, and gynaecology.

This corpus is called virtual because it is a collection of texts that will be obtained solely by downloading data from the Internet, due to the restriction of the semi-automatic compilation. In the same way, it is a comparable corpus, i.e. the documents are originally written by native speakers, due to its importance to represent the actual language rather than fictitious situations.

Once the design criteria are established, we will proceed to describe the compilation process.

2.2. Semi-automatic compilation with *BootCat*

BootCat[3] is a semi-automatic compilation tool which makes use of on-line information to build corpora. Additionally, it provides access to large amounts of data in a few minutes, reducing the time of manual or non-automatic compilation significantly. *BootCat* performs automatic searches on the net by means of two different items that should be included in the tool: keywords and URLs. This is the reason why this tool is not based on automatic but semi-automatic search.

In order to obtain a representative corpus of this field of specialisation, i.e., medicine, we are going to follow the compilation protocol —divided in four stages— established by Corpas Pastor (2008) (see also Corpas Pastor and Seghiri, 2009; and Seghiri, 2011), with the peculiarity that some of the stages of the compilation protocol are going to be performed automatically.

The solution adopted in this study is to use different sub-files according to the language, genre and medical specialties. In this manner, three comparable subcorpora have been extracted from the multilingual TELL-ME corpus, e.g. the English, the German and the Spanish subcorpus. Furthermore each subcorpus is composed of four components appertaining to the genre, i.e., specialised articles from specialised journals (like *BMJ*,[4] *SEMES*[5] and *Ärzte Zeitung*[6]), informative articles from divulgative journals (like *Medline Plus*,[7] *Consumer*[8] and *Onmeda.de*[9]), informative articles from Wikipedia, and patient information leaflets from *eMC*,[10] *Vademecum*[11] and *Diagnosia.de*.[12] Finally, each of these four components contains 11 sub-components relevant to the medical specialties which were chosen during the selection of the design criteria (first contact, emergency, internal medicine, ophthalmology, general surgery, cardiac/vascular surgery, plastic surgery, traumatology, otorrinolaryngology, urology, and gynaecology).

So, the first stage of the corpus compilation consists of locating and accessing information available on the Internet. Regarding the location of this information, it is performed manually since we have to search the net to obtain the ideal resources. When accessing information, the process is automatic as the tool is responsible for dealing with this particular task.

Once the documents have been located and accessed, the second stage consists of downloading the data, as well as the acquisition of information. The downloading is performed automatically by *BootCat*.

The next stage encompasses text normalisation, i.e., converting the collected texts to plain text format or ASCII format in order to allow the future exploitation of the corpus. In this case, *BootCat* also performs this task. Therefore, this stage is executed by the tool itself.

Finally, the fourth stage consists of the storage of the data. In this last stage, it is essential to identify the documents correctly and arrange them, which makes this task also manual.

Once the number of subcorpora, components and sub-components relative to the TELL-ME corpus is known, a proper codification may be established in order to arrange the documents in each sub-file.

Thus the first text provided by *BootCat* within a specific domain will start with the code 01. Then, this code is followed by the information regarding the language in which the text is written —ES for Spanish, EN for English, DE for German. After determining the language, the genre (see Table 1) and secondly, the medical specialty (see Table 2) are indicated. The list of codes for the last two identifiers is included in the following tables.

[3] <http://bootcat.sslmit.unibo.it/>.
[4] <http://www.bmj.com/>.
[5] <http://www.semes.org/revista_EMERGENCIAS/>.
[6] <http://www.aerztezeitung.de/>.
[7] <http://www.nlm.nih.gov/medlineplus/>.
[8] <http://www.consumer.es/>.
[9] <http://www.onmeda.de/>.
[10] <http://www.medicines.org.uk/emc/>.
[11] <http://www.vademecum.es/>.
[12] <http://www.diagnosia.com/de/>.

Genres		
EN	ES	GE
specialised journal (SJ)	revista especializada (RE)	wissenschaftliche Zeitschrift (WZ)
patient information leaflet (PIL)	prospecto (PR)	Packungsbeilage (PAC)
divulgative journal (DJ)	revista divulgativa (RD)	populärwissenschaftliche Zeitschrift (PZ)
Wikipedia (WI)	Wikipedia (WI)	Wikipedia (WI)

Table 1: Multilingual list of the genres included in the TELL-ME corpus and their corresponding codes.

Medical specialties		
EN	ES	GE
primer contacto (PC)	first contact (FC)	Erstkontakt (ER)
urgencias (UG)	emergency (EM)	Notdienst (NO)
medicina interna (MI)	internal medicine (IM)	Innere Medizin (IM)
oftalmología (OF)	oftalmology (OF)	Ophthalmologie (OP)
cirugía general (CG)	general surgery (GS)	Allgemeine Chirurgie (AG)
cirugía vascular (CV)	cardiac/ vascular surgery (CS)	Gefäßchirurgie (GE)
cirugía plástica (CP)	plastic surgery (PS)	Plastische Chirurgie (PC)
traumatología (TR)	traumatology (TR)	Traumatologie (TR)
otorrinolaringología (OT)	otorrinolaryngology (OT)	Otorhinolaryngologie OT
urología (UR)	urology (UR)	Urologie (UR)
ginecología (GI)	gynecology (GY)	Frauenheilkunde (FR)

Table 2: Multilingual list of the medical specialties included in the TELL-ME corpus and their corresponding codes.

In order to illustrate the codification already outlined, two examples have been given. The codification for a text which happens to be the first compiled (01) in English (EN) within the genre of specialised journals (SJ) and which belongs to the medical domain of emergency (EM) should be codified as 01ENSJEM. On the contrary, if the tenth text (10) compiled in Spanish (ES) within the genre of patient information leaflet —*prospecto* (PR), in Spanish— and whose medical specialty is urology, —*urología* (UR)— is encoded, the code given to this document would be 10ESPRUR.

Nevertheless, this codification should be recorded along with the keywords and the URLs which were used during the compilation in order to avoid loss of information, in case a new research is necessary. For this purpose, a *table of references* has been designed. This table includes the code of the plain text document, the keywords used to search the documents, and the URL provided to track documents.

In spite of the multiple advantages of the semi-automatic compilation using *BootCat,* there are a few limitations presented by the tool which restrict the regular development of the process of compilation. Its most important limitations can be summarised as follows:

1. The tool uses only the Boolean operator "AND" when performing the searches, which causes a less accurate search than if Boolean operators such as NOT and NOR were used.
2. The keywords have to be closely related to each other in order to obtain results.
3. Sometimes the tool freezes during the search, which may be due i) to a poor selection of keywords, ii) to a poor choice of the website, which is restricted or does not provide any text, iii) to a depletion of the search credit —it allows 5000 free searches of word chains each month— iv) or simply to an unusual technical problem of the tool.
4. It forces the user to include a diatopic delimitation at the first stage of the search, which has been proven to be completely unnecessary.
5. The impossibility to perform a new compilation without closing and opening the tool again impedes the work.
6. As a final disadvantage, it should be pointed out that the tool lacks technical support. In addition, the FAQs section does not contain any of the errors that we have highlighted.

Having explained the procedure and difficulties of the semi-automatic compilation, the resulting corpus should be conveyed as follows: TELL-ME corpus is a multilingual comparable corpus with 149,559 types and 10,823,893 tokens in English; 121,244 types and 4,711,886 tokens in Spanish; and 172,200 types and 8,163,489 tokens in German.

The representativeness of the TELL-ME corpus has been proven in terms of quality thanks to the design criteria (determined by the goals of the project) and the protocol of compilation in four steps. In terms of quantity, the application ReCor[13] guarantees the representativeness of the corpora, subcorpora, components and subcomponents.

Once the compiled corpus is proven to be representative of the field of specialisation of medicine according to the goals of the TELL-ME project and from the point of view of both the quality and the quantity, the TELL-ME corpus is become the tool in which the exercises of the e-learning platform, i.e. vocabulary exercises, and its own language courses are based.

3 Exercises based on the TELL-ME corpus

As stated, the TELL-ME e-learning platform is based on NLP and corpus linguistics which both establish the foundations of the resources. The exercises on vocabulary have been created using both technologies and extracting the information from the texts compiled by the corpus. That means that the TELL-ME project has set up its own language exercises based on the TELL-ME corpus.[14]

Within the exercise tab, different types of exercises can be found, i.e., listening comprehension, vocabulary and pictures. In this case, the exercise on vocabulary is presented as a quiz, where several multiple choice exercises are offered to the user, who must answer each of them to complete the quiz.

Some of the exercises that have been developed using the TELL-ME corpus can be described as follows:

- Insert the term in the right position: a sentence with some blanks is provided and the same number of terms which match the blanks are also given.
- Select the correct term for a blank: each blank has a drop-down list with four terms, of which only one of them is the right one.
- Fill in the blank where the definition is provided: the user has to pick a given term which fits both with the sentence and the actual definition amongst some others.
- Fill in the blank where neither definition nor options are provided: the user has to read the sentence and provide the correct term.
- Definition match: a sentence where a determined term is highlighted is provided, as well as five different definitions. The user has to link the term with its correct definition.

A user's feedback report has been designed within the TELL-ME project to evaluate the quality of the exercises based on Moodle.

It is important to point out that, by means of the user's feedback report, the exercises described above have been selected by users as the preferred activities within the platform, which increase the value of the TELL-ME corpus.

4 Conclusions

As we have seen, it is possible to semi-automatically compile a virtual, comparable and multilingual corpus representative of the field of medical specialties using *BootCat.* At the same time it allows us to considerably reduce the time invested.

It is also possible to use the compiled corpus as a base for an e-learning platform, mainly to obtain different types of exercises using the advantages that Natural Language Processing offers.

We have also had the opportunity to analyse the opinions of different users —medical professionals from UK, Germany and Spain— when testing the platform. After analysing the data, self-assessment exercises on vocabulary, which are based on the TELL-ME corpus, have been proven to be the most accepted activities

[13] ReCor is a software application that enables accurate evaluation of corpus representativeness. For further information, see Corpas Pastor and Seghiri, 2009; and Seghiri, 2011.

[14] These exercises within these courses are based on Moodle and its implementation was carried out by the University of Wolverhampton and the University of Saarland.

(84.95%). This means that the corpus has a great acceptance between the testers. Besides, 88.17% of the testers have highlighted the importance and utility of the TELL-ME corpus within the platform and 95.70% of the users have emphasised the rational judgement of organising the course content by medical specialties, which corresponds to the structure of the corpus.

All the data presented above goes to prove the importance that users afford to the field of terminology for the acquisition of medical communication skills is deserved and can only strengthen communicative and terminological needs of medical professionals.

Acknowledgments

The TELL-ME project has been funded with support from the European Commission. This publication reflects the views only of the authors, and the Commission is not responsible for any use of the information contained therein.

This paper is supported by Andalucía Tech's programme "Atracción de investigadores de reconocido prestigio".

TELL-ME corpus will be extended to its use in a national related project, INTELITERM (ref. no. FFI2012-38881), also using semi-automatic compilation.

References

Françoise Salager-Meyer. 1994. Hedges and textual communicative function in medical English written discourse. English for Specific Purposes, 13(2): 149-170.

Gloria Corpas Pastor and Míriam Seghiri. 2009. Virtual Corpora as Documentation Resources: Translating Travel Insurance Documents (English-Spanish). Corpus Use and Translating. Amsterdam/Philadelphia: John Benjamins, 75-107.

Gloria Corpas Pastor and Ruslan Mitkov. 2013. Nuevos entornos formativos para la comunicación médico-paciente en un contexto transnacional. Uciencia. Málaga.

Gloria Corpas Pastor. 2008. Investigar con corpus en traducción: los retos de un nuevo paradigma. (StudienzurromanischenSprachwissenschaft und interkulturellenKommunikation, 49). Frankfurt, Berlin and New York: Peter Lang.

John M. Sinclair. 2004. Corpus and Text: Basic principles. Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxford University.

Maurizio Gotti and Françoise Salager-Meyer. 2006. Advances in medical discourse analysis: Oral and written contexts. Amsterdam: Peter Lang.

Míriam Seghiri. 2011. Metodología protocolizada de compilación de un corpus de seguros de viajes: aspectos de diseño y representatividad. Revista de lingüística teórica y aplicada (RLA), 49 (2): 13-30.

Randolph Quirk. 1992. On Corpus Principles and Design. Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm, 4-8 August 1991. Berlin/NewYork: Mouton de Gruyter, 457- 469.

William H. Fletcher. 2004. Facilitating the Compilation and Dissemination of Ad-Hoc Web Corpora. Fifth International Conference on Teaching and Language Corpora. Amsterdam: Benjamins, 1-18.