TriMED: A Multilingual Terminological Database

F. Vezzani[1], **G. M. Di Nunzio**[2], **G. Henrot**[1]
[1]Dept. of Linguistic and Literary Studies, [2]Dept. of Information Engineering
University of Padua, Italy
federica.vezzani@phd.unipd.it, giorgiomaria.dinunzio@unipd.it, genevieve.henrot@unipd.it

Abstract

Three precise categories of people are confronted with the complexity of medical language: physicians, patients and scientific translators. The purpose of this work is to develop a methodology for the implementation of a terminological tool that contributes to solve problems related to the opacity that characterizes communication in the medical field among its various actors. The main goals are: i) satisfy the peer-to-peer communication, ii) facilitate the comprehension of medical information by patients, and iii) provide a regularly updated resource for scientific translators. We illustrate our methodology and its application through the description of a multilingual terminological-phraseological resource named TriMED. This terminological database will consist of records designed to create a terminological bridge between the various registers (specialist, semi-specialist, non-specialist) as well as across the languages considered. In this initial analysis, we restricted to the field of breast cancer, and the terms to be analyzed will be extracted from a corpus in English, accompanied by all relevant linguistic information and properties, and re-attached to their pragmatic equivalent in Italian and French.

Keywords: medical language, terminology, methodologies and tools for LRs construction and annotation

1. Introduction

Communication in the healthcare domain is characterized by a rigid and closed nomenclature which in many cases produces an opaque lexicon difficult to understand. Medical language often contains inconsistencies of scientific terminology such as semantic ambiguity, incorrect use of suffixes, archaism maintaining, redundancy in the formation of compounds, and etymological inconsistencies (Rouleau, 2003). As a result, patients and in general non-experts in medicine are often exposed to medical terms that can be semantically complex and hardly understandable. Moreover, despite the substantial amount of health-related information available on Internet, little is known about the quality and accessibility of that information. As a consequence, consumers using the Internet may have difficulties finding complete and accurate information on health issues. Deficiencies in information could negatively influence consumer decisions if people are relying on the Internet to make treatment decisions, including whether to seek care or not.[1] Moreover, it is important to state that due to the existence of an international language of communication, that is English, medical vocabulary is full of foreign words that can create problems during the transfer of medical knowledge across different languages.

In this work, we present a methodology for the implementation of a terminological tool that contributes to solve problems related to the opacity that characterizes communication in the medical field among its various actors. The main goals are: i) satisfy the peer-to-peer communication, ii) facilitate the comprehension of medical information by patients, and iii) provide a regularly updated resource for scientific translators. This work aims to provide a multilingual tool, a cross-evaluation study in which the languages considered are English, Italian and French.

The paper is organized as follows: after an overview of works previously developed in this regard, Section 2., we proceed in Section 3. by determining the three categories of people identified as subjects involved and affected by the complexity of medical language: physicians, patients, and technical translators. In Section 4., we go through identifying a working methodology that is the basis of the proposed linguistic resource as well as the description of the linguistic tool. We give our final remarks in Section 5.

2. Related Works

Regarding the complexity of the medical language, numerous studies demonstrate how it appears difficult to understand health information contained in drug package inserts (PATEL et al., 2002), in websites (MCCRAY, 2005; CENTER, 2008), and more generally in patients and medical doctor's communication (MCCRAY, 2005; JUCKS and BROMME, 2007; TRAN et al., 2009).

Patients are often exposed to complex medical terms and numerous research focuses on the concept of understandability related to this subject. The study of (GRABAR et al., 2014) proposes a specific lexicon in order to assess which words are potentially non-understandable and then require further explanations. The implementation of a specific lexicon in which the words are rated according to whether they are understandable or non-understandable for the medical field is considered as a first step towards the simplification of medical texts.

The understanding of words is a complex notion closely linked to Natural Language Processing (NLP) research field. Its purpose is to decide whether given documents are accessible for a given reader. The readability measures are widely used for evaluating complexity of documents (BOUAMOR et al., 2016) and it is possible to distinguish two types of readability measures: classical and computational (FRANÇOIS and FAIRON, 2013). Classical measures are essentially based on the number of characters and/or syllables in words, sentences or documents. Computational measures might involve vector space models and

[1]Health Internet ethics: ethical principles for offering Internet health services to consumers.

a wide range of descriptors and their combinations (ZENG et al., 2005; FRANÇOIS and FAIRON, 2013; LEROY et al., 2008) but, text readability formulas are mostly based on word length and sentence length. However, as McCray states in his study (MCCRAY, 2005), "Zen counting words and syllables and consulting a grade-level word list are most likely not sufficient to determine how readable a text is". This has prompted researchers (KESELMAN et al., 2007) to design more appropriate measures for medical texts which take into account term familiarity and recognition of the lexical form.

The difference between the language used by health care professionals and that used by patients is cited as a source of miscommunication (ELHADAD and SUTARIA, 2007). For example, non-expert people tend to use idiomatic expressions such as "mal de chien" (Fr) [literally, "dog pain"(En)] to refer to "douleur intense" (Fr) [severe pain, En].

The biomedical domain offers many linguistic resources for Natural Language Processing, including terminologies and corpora. However, most of these resources are prominently available for English and the access to terminological resources in languages other than English may not be so simple. Furthermore, there is a large audience of non-English speakers who can benefit from accessing health information in their native language. In this regard, (NEVEOL et al., 2014) review the extent of resource coverage for French and give pointers to access French-language resources.

In this paper, we present a methodology for the evaluation of the understandability and readability of medical language through the implementation of a terminological tool. Our main goal is then provide a linguistic resource available in order to satisfy the need for effective communication between various actors and the transmission of information in a clear and understandable way in three languages (English, French, and Italian). For this reason, our tool aims not only to provide information from a strictly linguistic point of view (like other annotation projects such as Framenet[2], Verbnet[3] or English PropBank[4]) but also to satisfy the needs of the user categories identified in Section 3.

3. Three Categories of Users

In terms of divulgation of scientific knowledge, we have proceeded by identifying three categories of people with problems related to the opacity of the medical vocabulary for different aspects and different levels of communication: physicians, patients and scientific translators.

3.1. Physicians

The international scale release of medical knowledge implies that most of the scientific texts are produced in English. For example, for several years many Italian medical journals have accept contributions only in English and have even anglicized their own denominations: 'Cardiologia' in 'Italian Heart Journal' since 2002, and 'Rivista Italiana di Pediatria' in 'Italian Journal of Pediatrics' since 2001 (SERIANNI, 2005). In terms of spreading new health care protocols and scientific discoveries, language could be a barrier to service transactions among medical specialists speaking different languages because perfect knowledge and mastery of the foreign language is not an expected outcome. At a level of peer-to-peer communication, and then specialist-to-specialist, physicians need to overcome these language barriers and access scientific research information in their mother tongue. In this way, experts could not only import new knowledge on the national territory but also export their scientific discoveries by inserting them into the international circuit. In this way, the direct benefit is raising the awareness of the importance of a proper terminology in the scientific communication between languages, the existence of possible false friends and the availability of tools specifically designed to respond with the utmost precision and reliability.

3.2. Patients

Scientific and technological development has so much influence on medicine and its diagnostic and therapeutic capabilities. This fact has shifted the focus of the physician's attention not on the patient but on the illness itself, and this has led to a crisis in the physician-patient relationship. Patients find a considerable difficulty in understanding information, both oral and written, about their own health with regard to clinical interactions despite laws and policies, emphasizing the real need to document the various health aspects in a more comprehensible way [5]. Physician-patient interaction implies a level of specialist-non-specialist communication, so patients (or more generally the public) would need to understand medical expertise by using their correspondent in the "popular" language or by using an appropriately calibrated language for the communication to be effective. It is also important to consider the increasingly frequent use of the Web as a source of medical and health information. Search engines are commonly used to access information available online but many resources are far from being effective in order to respond adequately to user requests and this may have serious consequences. Furthermore, the fact of exposing people with poor medical knowledge to a complex medical language can lead to self-diagnosis and erroneous self-treatment. In this sense, the Higher Institute of Health in Rome, Italy, has promoted the MEDUSA[6] project (MEDicina Utenti SAlute in rete), which is a citizen portal for the retrieval of qualified and reliable health information on the web. The implementation of this portal is part of the activities of a wider health education project funded by the Ministry of Health, titled *Alfabetizzazione sanitaria ed empowerment del paziente attraverso lo sviluppo di un sistema informativo elettronico nel campo della salute*[7] aiming to raise awareness in health issues and provide access, through a single platform, to information and resources of different types and nature.

[2]https://framenet.icsi.berkeley.edu/fndrupal/

[3]http://verbs.colorado.edu/ mpalmer/projects/verbnet.html

[4]https://propbank.github.io/

[5]https://goo.gl/5Avgrd

[6]https://medusa.iss.it/

[7]Health Literacy and Patient Empowerment through the Development of an Electronic Health Information System https://goo.gl/6dihwr

3.3. Technical-scientific Translators

In the sphere of communication and dissemination of information, translation, as a practical discipline, acquires a fundamental role in the correct transmission of information in different languages. In order that the translated text responds to the deontological principle of the discipline, that is, the fidelity to the source text, the technical-scientific translator must proceed step by step, decoding, deeply understanding and faithfully translating the semantic and informative content of the text. In this case, the level of communication is at a specialist semi-specialist degree. Moreover, the needs of the translation market do not allow time to conduct in-depth terminological research, forcing the professional to skip some key steps for optimal work. Translators need regularly updated terminology resources which can support them in the realization of the final product.

At present, one of the most reliable bilingual resource for terminology and translation which is available on the market is an Italian /English bilingual database of Medical Subject Headings (MeSH). This tool is used in the indexing of articles in biomedical journals of PubMed, by the Higher Institute of Health of Rome. But, as with many other medical dictionaries online, the resource is derived from terminology cards that can complete and meet the requirements outlined above.

4. TriMED

The needs outlined above for the three identified categories of users would find a valid application in a multilingual terminology database. This work aims to provide a methodology for the development of TriMED, that is a multilingual terminological database gathering a set of terminological records for each selected technique terms, at each pragmatic level identified. The resource is named TriMED because the tripartite character is intrinsic in the tool: three are the working languages (English, French, and Italian), three are the identified user categories (physicians, patients, and translators) and, consequently, three are the communication levels which are the object of the analysis of this tool. The type of textual corpus from which technical nomenclature will be extracted concerns oncology and it will collect English-language articles related to care protocols for breast cancer patients. From the corpus so drawn, the medical technical terms will be extracted and analyzed. The term record will report the product of semantic analysis and will provide translations, drawn from parallel corpus in the target languages: either Italian or French. Starting from these, the corresponding technicalities will be selected to meet the multilingual goal of the terminological database.

4.1. Structure

TriMED records will be articulated on three levels of communication:

- specialized communication, providing both scientific definitions to meet the peer communication between experts and the corresponding technical terms translated;
- semi-specialized communication, providing useful information to the technical-scientific translator for the translation of texts for this medical domain, such as semantic analysis of the term, instances of collocations, colligations, hyperonym, iponyms etc.
- non-specialized communication, providing both informative definitions to facilitate proper understanding by patients and the equivalent of the technical term commonly used in the popular language.

TriMED database will provide not only the simple translation of the term in English (in the three registers indicated), but also all the information necessary to make the medical technical term clear and semantically accessible.

4.1.1. Data Collection

At this stage, we proceeded by the selection of the source corpus and parallel corpora and by the extraction of technical terms.

First, we have selected a set of English-language articles representing our source corpus for the analysis of medical terms. In this initial phase of gathering an initial dataset and validating the application, we limit the topic of interest to breast cancer treatment protocols, but in the future, we expect to extend the domain of interest to other medical areas. Documents are selected from specialized online magazine reviews based on the highest impact factor value, such as "Breast Cancer research and treatment"[8], "Archives of Breast Cancer"[9], and "European Journal of Cancer Care"[10]. Afterwards, we have created parallel corpora for the other two languages analyzed: Italian and French. The sources we have drawn for the Italian language are: Fondazione Umberto Veronesi (in particular, protocols belonging to the initiative "Pink is Good"), AIMaC - Associazione Italiana Malati di Cancro, and A.N.D.O.S - Associazione nazionale donne operate al seno. While for the French language: Fondation du Cancer du Sein du Quebec, l'Association francophone pour les soins oncologiques de support (AFSOS) and Cairn.info. Then, we have proceeded through the manual extraction of medical terms for the three corpora. By technical terms we mean all terms that are closely related to the conceptual and practical factors of a given discipline or activity, in this case terms are related to the medical-oncological field.

4.2. Web Application

In this section, we apply the previous outlined methodology by the creation of a Web application that can respond to the principles underlying in this paper. For each technical term, we provided a set of information for the three categories of identified users. Every term has been supplied with:

- Equivalent in the informative and popular language;
- Definition;

[8] http://www.springer.com/medicine/oncology/journal/10549

[9] http://archbreastcancer.com/

[10] http://onlinelibrary.wiley.com/journal/10.1111/(ISSN)1365-2354

- Semantic analysis of the term;
- Formal features.

Upon these information, we have created TRIMed Interface that is presented as a "three-headed" tool: from the homepage you can select the category in which you identify and then access to the related information you need. In Figure 1, we show the main panels of the application: physician (Fig. 1c), translator (Fig. 1a), and patient (Fig. 1b). We have implemented the application with the Shiny R package (CHANG, 2015), the demo is available online to show how the interaction among the three levels of communication works[11].

Consulting the interface in "patient" mode, you can search the technical term (with the suggestion of completing the term) and its equivalent in the popular language and its definition are provided. With the aim of simplifying medical language, this tool allows the patient to "translate" the medical term with a more simplified or commonly used terminology. Furthermore, considering the ternary programming lines (specialists, semi-specialists, non-specialists), we intended to provide the possibility of consulting an "informative" variant for the technical term in question, which is to be considered different from the term in the popular language. This implementation is still being developed and assessed as we have done a distinction between term in popular language, by classifying it as a recurrent word mainly in oral talks between patients (for example, in clinic and hospital waiting rooms), and semi-specialist (or informative) term that has its place in informative articles and which is then extracted from written corpus. There may be perfect correspondence between popular terms and informative terms, as both are intended for the use of a non-specialized user in this field. But it is still interesting to allow for this further linguistic consideration in view of the diastratic evaluation of terminology.

As far as the "translation" mode, the interface is designed for providing to the user the purely linguistic and terminological information that underlie the translating process. After selecting the term, the user can visualize a screen presented in the form of a double-read table: vertical and horizontal. By proceeding with a vertical analysis, the user is able to consult the translation of the term in its scientific, informative and popular language version. Subsequently, the definition of the technical term and its semic analysis will be provided in the decomposition of the meaning of the lexical or morphological unit into atomic units or components of the not further segmentable meaning. Finally, the user can access to the formal features of the term necessary for its lexical framing:

- Gender;
- Pronunciation;
- Derivation and composition of the term.

This information is necessary for a translator in the choice of the translating candidate of the term taken into consideration. Keeping the table with the source term, the translator can set the target language and consult the information of the chosen translating term that will appear in a table adjacent to the source table. In this way, the user can access the same information for the selected translating term and can consult horizontally the information for the two terms.

Finally, the access mode as a "physician" user is an interface that offers in the translation point of view the opportunity for the physician to consult the technical term in his or her mother tongue. This user has the opportunity as well to select the source language and the target language in which he/she wants to examine the word and its related definition. For this user category, a direct link with related MeSH[12] terms has also been provided. By clicking on the term, the physician can access to the various information provided directly by the National Library of Medicine.

5. Conclusions

This work has been developed with a view to the evaluation of medical language in terms of understandability and readability. Through our methodology of language analysis, we are creating a linguistic resource that could answer the initial questions and needs outlined by the three categories of users. At the present time, TriMED consists of a set of 200 technical terms for French and Italian languages. We are working on English version in order to fullfill the multilingual goals.

We are planning to gather enough data to cover medical terminology in the oncology field, in particular breast cancer treatments, by trying to propose a resource that does not only include English but also other target languages requiring documented terminology in that field. Our intent is then provide a linguistic tool consisting of 2500 – 3000 technical terms for this specific domain.

Hence, the structure and idea behind TriMED allow for future implementations. For example, with regard to the enrichment of the "physician" mode, we propose to allow the user to have the direct access to articles from scientific journals related to that precise technical terms. For example, for English terms, a direct link between Mesh terms and PubMed related articles may be helpful. While for Italian terms, the link could be made with the bilingual Mesh database provided by the Roma Healthcare Institute. Similarly, French terminology could be directly linked to the InSerm site[13] which, in co-operation with Inist-CNRS (Istitut de l'Information scientifique et technique du CNRS) contributes to updating the French version of Mesh terms since 2004.

Finally, TriMED is designed to support data and knowledge discovery and integration as well as promote sharing and reuse of data by following the FAIR principles of the research data in Horizon 2020.[14]

6. Bibliographical References

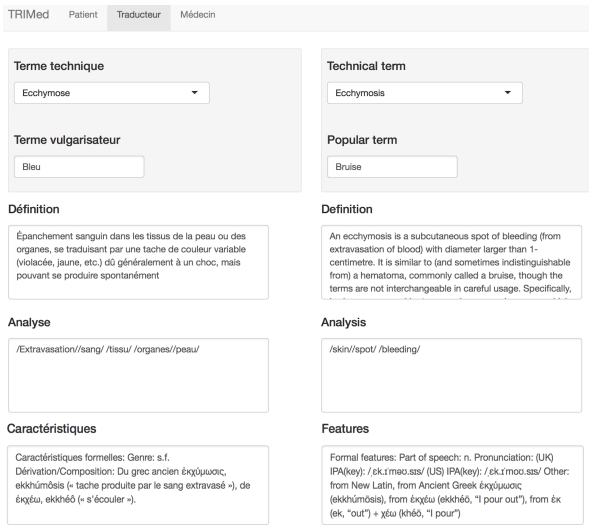
BOUAMOR, D., LLANOS, L. C., LIGOZAT, A.-L., ROSSET, S., and ZWEIGENBAUM, P. (2016). Transfer-based learning-to-rank assessment of medical term tech-

[11] https://gmdn.shinyapps.io/TriMED/

[12] https://www.nlm.nih.gov/mesh/

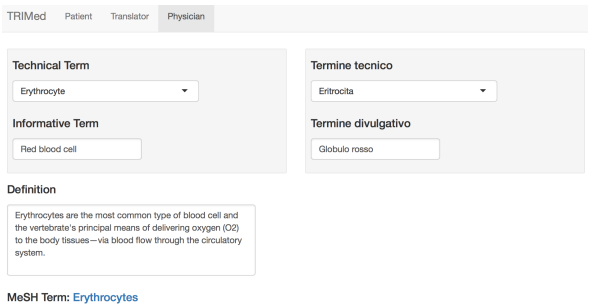
[13] http://mesh.inserm.fr/mesh/

[14] https://goo.gl/WSUd9K



(a) Translator interface (French to English)

(b) Patient interface (Italian)



(c) Physician interface (English to Italian)

Figure 1: Main panels of the TRIMed Web application.

nicality. In *Proceedings of the LREC 2016*, Portorož, Slovenia, May. ELRA.

CENTER, O. P. (2008). Barriers and drivers of health information technology use for the elderly, chronically ill, and underserved. technical report, agency for healthcare research and quality. Oregon Evidence-based Practice Center.

CHANG, W., (2015). *Shiny: Web Application Framework for R*. R package version 0.11.

ELHADAD, N. and SUTARIA, K. (2007). Mining a lexicon of technical terms and lay equivalents. In Proc BioNLP Workshop, pages 49-56. ACL.

FRANÇOIS, T. and FAIRON, C. (2013). Les apports du tal à la lisibilité du français langue étrangère. TAL, 54(1): 171-202.

GRABAR, N., VAN ZYL, I., DE LA HARPE, R., and HAMON, T. (2014). The comprehension of medical words. In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5*, BIOSTEC 2014, pages 334–342, Portugal. SCITEPRESS - Science and Technology Publications, Lda.

JUCKS, R. and BROMME, R. (2007). Choice of words in doctor-patient communication: an analysis of health-related internet sites. Health Commun, 21(3):267-77.

KESELMAN, A., TSE, T., CROWELL, J., BROWNE, A., NGO, L., and ZENG, Q. (2007). Assessing consumer health vocabulary familiarity: an exploratory study. J Med Internet Res, 9(1): e5.

LEROY, G., HELMREICH, S., COWIE, J., MILLER, T., and ZHENG, W. (2008). Evaluating online health information: Beyond readability formulas. In AMIA Annual Symposium Proceedings.

MCCRAY, A. (2005). Promoting health literacy. J of Am Med Infor Ass, 12:152-163.

NEVEOL, A., GROSJEAN, J., DARMONI, S., and ZWEIGENBAUM, P. (2014). Language resources for french in the biomedical domain. In *Proceedings of the LREC 2014*, Reykjavik, Iceland, May. ELRA.

PATEL, V., BRANCH, T., and AROCHA, J. (2002). Errors in interpreting quantities as procedures: The case of pharmaceutical labels. International journal of medical informatics, 65(3):193-211.

SERIANNI, L. (2005). Un treno di sintomi. i medici e le parole: percorsi linguistici nel passato e nel presente. Milano, Garzanti Libri.

TRAN, T., CHEKROUD, H., THIERY, P., and JULIENNE, A. (2009). Internet et soins : un tiers invisible dans la relation médecine/patient. Health Commun, 21(3):267-77.

ZENG, Q., KIM, E., and CROWELL, J.and TSE, T. (2005). A text corpora-based estimation of the familiarity of health terminology. In ISBMDA, pages 184-192, Aveiro, Portugal.