DOI: http://doi.org/10.6092/issn.2532-8816/8632

(Not so) Elementary, My Dear Watson! A Different Perspective on Medical Terminology

Federica Vezzani - Giorgio Maria Di Nunzio

Università degli Studi di Padova

federica.vezzani@phd.unipd.it
giorgiomaria.dinunzio@unipd.it

Abstract. Sir Arthur Conan Doyle was an esteemed and highly experienced physician and much of his medical knowledge spreads into his literary works. In this paper, we propose to study the medical terminology in the stories of Sherlock Holmes through a mixed-method of quantitative and qualitative analysis. Our approach is based on 1) the automatic extraction of medical terminology through the *tidytext* R package for text analyses, 2) a terminological analysis by means of the model of terminological record designed for the TriMED database, and 3) the study of collocations through the linguistic tool Sketch Engine. Thanks to this approach, we perform a linguistic analysis in order to evaluate different terminological aspects such as: semantic variation due to temporal and historical factors, a term's different contexts of use, change in meaning based on the reference corpus, variation of use depending on speakers'/writers' register and, finally, the syntactic relationship between terms and their collocations.

Sir Arthur Conan Doyle era un medico stimato e con grande esperienza e gran parte delle sue competenze mediche si riflettono nelle sue opere letterarie. In questo articolo, ci proponiamo di studiare la terminologia medica presente nelle storie di Sherlock Holmes attraverso la combinazione di un metodo misto di analisi quantitativa e qualitativa. L'approccio che proponiamo si basa 1) sull'estrazione automatica della terminologia medica attraverso il package *tidytext* R per l'analisi del testo, 2) su un'analisi terminologica tramite il modello di scheda terminologica progettata per il database TriMED e 3) sullo studio delle collocazioni attraverso lo strumento linguistico Sketch Engine. Grazie a questo approccio, conduciamo un'analisi linguistica al fine di valutare diversi aspetti terminologici come: la variazione semantica dovuta a fattori storici e temporali, la differenza del contesto di utilizzo di un termine, il cambiamento di significato di una parola in base al corpus di riferimento, la sua variazione di utilizzo a seconda del registro del parlante/scrivente e, infine, la relazione tra i termini e loro collocazioni dal punto di vista sintattico.

Introduction

In the last two decades, researchers of different fields have started to pay attention to the benefits deriving from the adoption, both in linguistics and in literature, of mixed approaches, quantitative and qualitative, to the study of texts. In particular, the availability of large quantities of electronic literary texts has increased the attractiveness of quantitative approaches as a way of "reading" literature with benefits in terms of time and accuracy [15]. The classical notion of "close reading" [26] is referred to the act of analyzing a small set of works upon a deep reading and interpretation of local features and aspect of its formal structure and content. This approach is opposed to the recent concept of "distant reading" [22] as the act of understanding literary phenomena through the computational analysis of massive amounts of textual data. The idea is that there are synchronic and diachronic literary features and phenomena that are difficult to detect with traditional reading and local interpretation methods, and which require the close examination of a huge number of texts and documents through a computational approach. Thus, the combination of quantitative and qualitative analyses can provide complementary insights both into the formal and rhetorical-narrative structure and into the topics and formulation of various types of written texts.

Corpus Stylistics Tools

In the last few years, there has been a considerable increase in studies in the new field of research called "corpus stylistics" [28]. Corpus stylistics brings together approaches from corpus linguistics and literary stylistics, and it is concerned with the application of corpus methods to the analysis of literary texts by relating linguistic description with critical interpretation [20]. In order to combine quantitative and qualitative analysis, the central linguistic tool for the analysis of literary corpora is based on the study of concordances as a list of the occurrences of a word extracted from a specific linguistic context [29]. The automatization of this process is provided by software tools such as the Simple Concordance Program (SCP)[1] or the suite WordSmith Tools.[2] Moreover, it is possible to have even more refined results in a KeyWords In Context (KWIC) format by means of domain-specific languages such as the Corpus Query Processor.[3]

The development of software tools and applications which use user-friendly visualizations of text has attracted an increasing number of researchers to corpus stylistics; these visualization tools allow researchers to investigate the meaning of the words in context and co-text [34]. As stated by Tognini-Bonelli [31], in order to find repeated occurrences of words, a concordance is read "vertically" rather than horizontally, as we normally read a text. In this way, the patterns that become visible in a concordance provide information on the meanings of words. Furthermore, it is also possible to obtain a display of non-contiguous sequences, as with ConcGram tool [3]. ConcGram is a software package of corpus linguistics that has been specially designed to find all co-occurrences of words in a text or a corpus regardless of their positional variation generated by the association of two or more words. These are called

[1] http://www.textworld.com/scp/

[2] https://www.lexically.net/wordsmith/

[3] http://fedora.clarin-d.uni-saarland.de/teaching/Corpus_Linguistics/Tutorial_CQP_I.html

"concgrams", and they are defined as "*all of the permutations of constituency variation (increase* in *expenditure, increase* in the share of public *expenditure) and positional variation (expenditure* would inevitably *increase) generated by the association of two or more words*". The software automatically finds the co-occurrences of given words: in other words, the user does not have to set up previous search commands. ConcGram is particularly suitable for discovering the complete phraseological profile of a text or a corpus. This software has a wide range of applications in the fields of corpus linguistics, critical discourse analysis, speech analysis, lexicology, lexicography, pragmatics and semantics.

The display format of a concordance may help the identification of formal schemas associated with functions in the text; at the same time, association models can also be identified without resorting to typical concordance visualizations, as evidenced by the various types of measurements for location extraction or techniques for generating clusters of words, n-grams and so on. For example, the recent GraphColl [2] tool shows that there are other types of collocation displays compared to classical concordance lines. GraphColl is a useful tool for the construction and exploration of links between linguistic collocations. The text in a particular field of discourse is organized into lexical schemas, which can be visualized as networks of words that are placed one with the other. GraphColl is a tool that builds collocation networks starting from corpora, allowing the user to obtain important information on semantic relationships. Finally, one of the most used and useful tools currently on the market is Sketch Engine,[4] that is a corpus manager and text analysis software developed by Lexical Computing Limited in 2003 [18]. This tool was designed for people studying language behavior such as lexicographers, researchers in corpus linguistics, translators or terminologists; it allows one to conduct searches in large text collections according to complex and linguistically motivated queries. One of its key-features is the "word sketch" function, which is a one-page summary of the word's grammatical and collocational behavior. It shows the word's collocates categorized by grammatical relations such as words that serve as an object or subject of the verb and so on. Sketch Engine is therefore a useful instrument for text analysis in order to perform co-occurrence analysis, term extraction or generate frequency lists which can be the basis of qualitative-linguistic studies of the text analyzed.

Related Works

All these tools and quantitative methods of analysis based on concordances highlight the linguistic qualities of a text from lexico-syntactic, phraseological perspective, which can shed light on semantic and pragmatic aspects of language use. For example, concordance tools can generate and collect keywords of texts by comparing the frequencies of the words in the same text with the frequencies of those words in a reference corpus. The work of Scott and Tribble [27] illustrates a study based on keyword analysis by comparing *Romeo and Juliet* with a corpus containing all of Shakespeare's plays. Even in Culpeper [4], key words are used to analyze *Romeo and Juliet*, but the author adopts a different perspective by comparing the speeches of

[4] https://www.sketchengine.eu

individual characters. In this context, the work of Mahleber at al. [21] aims to study Charles Dickens's narrative through the development of a web application, named CLiC, in order to bring together insights from both cognitive poetics and corpus stylistics. This tool supports the analysis of discourse in narrative fiction with search options focused on stretches of text within and outside quotation marks. With this kind of analysis, the authors focused on some characteristics of the narrative of Charles Dickens regarding the use of direct discourse and its suspensions.

Quantitative approaches are used not only to investigate larger interpretative issues like plot, theme, genre, period or modality, but they are also frequently associated with questions of authorship attribution and style. The study of Franzini et al. [10] fits in the context of computational text analysis and it specifically describes an effort to automatically discern the authorship of Jacob and Wilhelm Grimm in a corpus of correspondences. The paper describes the impact of digitization noise on the automatic attribution of a body of letters by comparing three different digitization outputs: (a) manual transcriptions of the original letters, (b) the Optical Character Recognition (OCR) of a 2001 printed critical edition of the Grimm letters, and (c) a Handwritten Text Recognition (HTR) model for the automatic transcription of the original letters.

In the context of the problem of authorship attribution, the concept of stylometry [14], that is the study of measurable features of literary style (such as sentence length, vocabulary richness and various frequencies of words), has found numerous practical applications. These applications are usually based on the hypothesis that there are some elements of personal style that can help to detect the true author of an anonymous text. The study of Eder et al. [6] proposed a method based on the identification of the most frequent word n-grams, using the R programming language, in order to perform multivariate analysis to provide new insights such as: relationships between different books by the same author, between books by different authors or between authors differing in terms of chronology or gender. A recent contribution in this context is the work of Tuzzi and Cortelazzo [32] looking at the case of Elena Ferrante as the presumed pseudonym of an internationally successful Italian novelist. The authors of this work created a reference corpus composed of 150 novels by forty different authors with the main goal of understanding whether, amongst the authors in the corpus, there are any that can be considered candidates for involvement in the writing of the Ferrante's novels. The method is based on the measurement of the degree of similarity between the novels: from these quantitative analysis, Domenico Starnone was identified as the author who has written novels most similar to those of Ferrante. The case of Elena Ferrante is interesting from both the stylistic and the stylometric viewpoints, and has attracted a good number of researchers, especially on the occasion of the Workshop "Drawing Elena Ferrante's Profile" organized by Cortelazzo and Tuzzi at the University of Padua.[5]

Another application of quantitative analysis for the study of literary texts focuses on the vocabulary variation. Quantitative thematic analysis can trace the growth, decay, or development of vocabulary within a thematic domain, or study how authors differ in their expressions of a theme [9]. In Lancashire and Hirst [19], the authors aim to study the

[5] http://www.padovauniversitypress.it/publications/9788869381300

vocabulary changes in Agatha Christie's novels. The innovative part of this work is the use of the linguistic analysis for the detection of indications of dementia in the author's style. Authors analyze the vocabulary of Christie, who, although never diagnosed, was believed to have suffered from dementia in her final years, even as she continued to write. By studying the concordances in the literary works of Agatha Christie, the authors concentrate on measuring the richness in the vocabulary and on syntactic and discourse-level aspects of her texts.

Our proposal

In this paper, we present a combined method of quantitative and qualitative analysis of literary works of Conan Doyle, focusing on the terminological viewpoint. To our knowledge, this is the first attempt which aims to study medical terminology in the stories of Sherlock Holmes through the combination of a computational approach for the extraction of medical terms and their qualitative analysis performed through a new model of terminological record. Our main goal is the study of how a specialized terminology is represented in literary works, focusing on vocabulary changes in terms of diachronic and register variations.

The remainder of this paper is organized as follows: in Section 2, we provide a short background about the life of Conan Doyle in order to show the close connection between medicine and his literary works. Then, we present the studies related to medical topics in Conan Doyle's narrative. In sections 2.1 and 2.2 we define our proposal by outlining our mixed method of analysis based on 1) the extraction of medical terminology through the *tidytext* R package for text analysis,[6] 2) a terminological analysis through a new model of terminological record and 3) the study of collocations through the linguistic tool Sketch Engine. In Section 3, we provide a linguistic analysis of some features resulting from our combined approach and, finally, we provide conclusions and some hints on future work.

Medical Terminology in Conan Doyle's narrative

Numerous studies ([13]; [23]; [25]; [35]; [36]; [37]) reveal the amount of "medical subjects" in Conan Doyle's works: the author was, in fact, an esteemed and highly experienced physician and much of his knowledge spreads into his stories. Conan Doyle received his bachelor's degree in Medicine at the University of Edinburgh in 1881. When he was still a student (1878-1880), Doyle carried out some assistance activities with numerous important physicians, first in the Scottish capital and then in England. Together with this experience, he worked as a "ship's surgeon" for seven months in 1880 on board of an ancient whaling boat named "Hope". In 1885 Conan Doyle left England to go to Vienna, where he began work as an ophthalmologist. He then returned to London in 1891 to open an ophthalmology clinic in the city center. After a short period of activity and a severe summer influenza, in August 1891 Conan Doyle decided to devote himself to writing.

[6] https://cran.r-project.org/web/packages/tidytext/

The vastness of Conan Doyle's medical experience is reflected in the wealth of medical topics covered in the stories of Sherlock Holmes. From a terminological point of view, Key J.D. and Rodin A.E. are the first authors to explicitly mention the number of medical terms in Conan Doyle's works: "[…] *68 diseases, 32 medical terms, 38 doctors*" [17]. Nevertheless, very little has been written regarding the medical terminology in this literary corpus. The most recent and complete contribution is the study by Ernesto Damiani, an expert physician interested in history, who wrote "*Elementare Watson!*" [5]: a dictionary of Italian medical terms used in the translated version of the stories of Sherlock Holmes. The author manually extracted about 98 Italian medical terms with the aim to provide a tool for facilitating the comprehension of such terms while reading Sherlock Holmes detective stories. Damiani offers an overview of the use of the terms and their explanations from a historical and medical point of view.

One of the objectives of our study is to partly reproduce Damiani's study, that is the collection of medical terms and its subsequent analysis. While Damiani performed a manual extraction, we propose an automatic extraction of the medical terminology by means of the *tidytext* R package for text analysis.[7] We expect to verify the benefits deriving from this kind of approach in terms of efficiency and accuracy, since we can automatically process the whole literary production of Conan Doyle in a very short time. We also believe that an automated extraction process may help to find more terms than those identified by Damiani. Moreover, while Damiani focused on a description of the terminology from a medical and historical perspective, our main goal is to combine a qualitative analysis from a purely linguistic and terminological perspective by means of a new model of terminological record, implemented for the linguistic analysis of technical medical terms [33].

To sum up, we propose to analyze medical terminology in literary texts through the combination of a quantitative approach (automatic extraction) and a qualitative analysis (terminological records). Our case study is the entire collection of works on Sherlock Holmes, starting from "A Study in Scarlet" (1887) to "The Casebook of Sherlock Holmes" (1927).

Automatic Term Extraction

The automated extraction of terms starts with the creation of the corpus of all the novels and stories of Sherlock Holmes written by Conan Doyle. We used the "Complete Sherlock Holmes Canon" website,[8] since this website has one of the most complete types of documentation about privacy policy and copyright issues about the re-use of the texts. In particular, we used the ASCII version of the 4 novels and 56 short stories for a total of 60 text files.

We wrote the software to automatically process the files and produce the list of medical terms with the R programming language; we followed the "tidyverse"[9] approach to process the data and documented each step in R Markdown[10] in order to make the entire process completely

[7] https://cran.r-project.org/web/packages/tidytext/

[8] https://sherlock-holm.es

[9] https://www.tidyverse.org

[10] https://rmarkdown.rstudio.com

reproducible by any researcher.[11] We used *tibbles* (modern R dataframe structures) to store the information about each book; we also used the dplyr package to make "pipelined" operations. In particular, before any text pre-processing, we removed from each file the string "Arthur Conan Doyle" at the beginning of the file and the copyright statement at the end of the file:

> ----------
>
> This text is provided to you "as-is" without any warranty. No warranties of any kind, expressed or implied, are made to you as to the text or any medium it may be on, including but not limited to warranties of merchantability or fitness for a particular purpose.
> This text was formatted from various free ASCII and HTML variants. See http://sherlock-holm.es for an electronic form of this text and additional information about it.
> This text comes from the collection's version 3.1.

We used the *tidytext* R package to perform text analyses: we created n-grams (unigrams, bigrams, and trigrams) that are used to match medical terms. The following example shows the source code used to generate unigrams, count the occurrence of each unigram and order the table thus produced by decreasing frequency of occurrences:

```
# create unigrams
doyle_unigrams <- books %>%
    unnest_tokens(unigram, text, token = "ngrams", n = 1)
# count and sort unigrams
unigram <- doyle_unigrams %>%
 count(unigram, sort = TRUE) %>%
 rename(word = unigram, freq = n)
```

At the end of this initial pre-processing phase, we obtained: 19,686 unigrams, 213,656 bigrams, and 481,156 trigrams.

Medical Termbase

In order to match and extract the medical terms from the pre-processed text, we used the Medical Subject Headings (MeSH) database.[12] This database is a controlled vocabulary thesaurus prepared and maintained by the US National Library of Medicine for indexing articles in PubMED.[13] In this context, the R package provided by the Bioconductor[14] website contains a set of annotation maps describing the entire MeSH database (MeSH ORA).[15]

[11] https://github.com/gmdn/SherlockHolmesTidyverse

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

[13] https://www.ncbi.nlm.nih.gov/pubmed/

[14] https://bioconductor.org

[15] https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-015-0453-z

In order to query the database, we needed the RSQLite package.[16] The query selects all the MeSH terms of the database as well as all the synonyms of each term. In fact, many synonyms, near-synonyms, and closely related concepts are included as entry terms in the database to help users find the most relevant MeSH descriptor for the concept they are seeking. Then, we converted all the characters to lower case and remove any dot character from the text in order to match the format of the text files (for example, St. Paul is converted to st paul). The following code shows how mesh_terms and mesh_syns (synonyms) are generated:

```
# get terms and synonyms
select_mesh_syn <- "SELECT DISTINCT MESHTERM, SYNONYM FROM
DATA"
res_mesh_syn <- dbGetQuery(dbconn(MeSH.db), select_mesh_syn)
res_mesh_syn  <- res_mesh_syn %>%
  mutate(SYN_CLEAN = sapply(strsplit(x =
res_mesh_syn$SYNONYM, # split syns
                                     split = "\\|"),
                          FUN = function(x) x[1])) %>% #
get first element
  dplyr::select(MESHTERM, SYN_CLEAN)

# get mesh terms and syns
mesh_terms <- tolower(unique(res_mesh_syn$MESHTERM))
# remove dots
mesh_terms <- gsub(pattern = "\\.", replacement = "", x =
mesh_terms)
mesh_syns <- tolower(res_mesh_syn$SYN_CLEAN)
# remove dots
mesh_syns <- gsub(pattern = "\\.", replacement = "", x =
mesh_syns)
```

At the end of this phase, we extracted 1,347 MeSH unigrams, 116 MeSH bigrams, 8 MeSH trigrams.

Terminological Record

After the automatic extraction of medical terminology, we proceeded to a linguistic and terminological analysis of such terms through the model of the terminological record set out in a multilingual database, named TriMED [33], conceived for the analysis of scientific terms extracted from a specialized corpus.

This tool is designed in order to tackle the problem of the complexity of medical terminology by considering different level of communications. Indeed, there are three categories of people that are mostly affected by the complexity of medical language and who can benefit from the use of this resource: patients, language professionals (translators and interpreters) and physicians.

[16] https://cran.r-project.org/web/packages/RSQLite/

Patients: Patients, or more in general lay people, find a considerable difficulty in understanding information, both oral and written, about their own health [16]. As a consequence, they need to understand medical technical terms by using their equivalents in the popular language: TriMED provides the equivalent of the technical term in the popular language, that is, the most frequently used term (as for example *fever* for *pyrexia*) and provides an informative definition in a non-specialized register.

Translators and Interpreters: TriMED is designed to support "language professionals" by offering terminological records providing the translation into three languages (English, French and Italian) of the technical term and all the linguistically relevant information for the process of decoding and transcoding it in its oral and written forms.

Physicians: 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. In order to overcome these language barriers and to satisfy the peer-to-peer communication, TriMED offers the possibility to consult the translation of the technical term in the specialized linguistic register.

For the qualitative analysis of the medical terms extracted from the literary corpus, we used the model of terminological record designed for this tool. Terminological records are commonly used in terminology and linguistics as a tool for the collection of linguistic data referring to a specific concept [11]. TriMED terminological record is structured around four axes of analysis of the technical term:

- Formal features
- Semantics
- Corpus
- References.

Regarding the formal and lexical framework of the term, we provide information such as: gender, spelling, pronunciation in the International Phonetic Alphabet (IPA) and other information about the etymology, such as derivation and composition of the term. In addition, we propose the spelling variant and the related acronyms which are currently used in medical language. Finally, based on the WordNet resource, [17] the record contains all the nouns, verbs, adjectives, and adverbs deriving from the analyzed term and which fall into the same semantic sphere.

The second section focuses on the semantic features of the term. First, we propose a definition extracted from reliable resources such as Merriam-Webster Medical Dictionary[18] or MediLexicon[19] especially for acronyms and abbreviations. In addition, we provide the semic

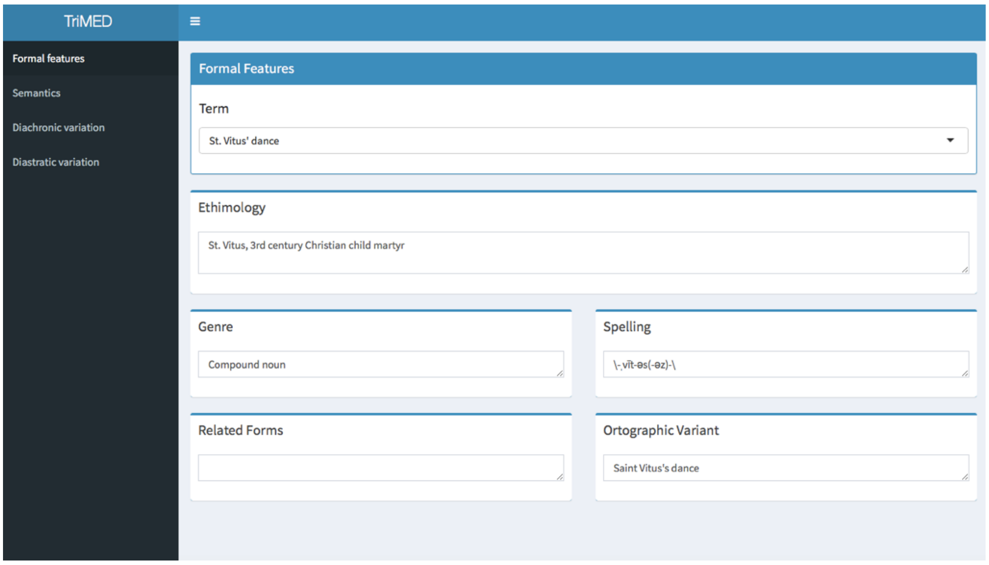
[17] https://wordnet.princeton.edu

[18] https://www.merriam-webster.com

[19] https://www.medilexicon.com

analysis of the term [24], that is, a methodology used in compositional semantics in order to decompose the meaning of technical terms (lexematic or morphological unity) into minimal units of meaning: the semes. Moreover, in order to evaluate the semantic behavior of a term, we collect the phraseology of the term by considering collocations [8] and colligations [30]. Finally, we provide the synonymic variants of the term: in this way, we categorize terms and their semantic relations. In the corpus section, we provide all specialized contexts where technical terms have been extracted and then we proceed through the identification of the domain and the register of the term (popular, slang, familiar, current or standard and specialized). Therefore, the term and its definition take on meaning when they are connected to a specific domain: in our analysis, we identify the domain and subdomains of the text (such as surgery, pathology, pharmacology, etc.). We also provide references to each source, since all of this information has been extracted from different sources.

In our study, we added two field of analysis aiming at the evaluation of the diachronic and diastratic variations of the term. With "diachronic variation", we mean morphological and semantic changes to which a term is subject over time, while "diastratic variation" refers to changes related to the different register used by any given speaker. Figure 1 shows an example of the terminological record of the term "St. Vitus's Dance". Moreover, a demo of the application realized with the Shiny R package is also available.[20]



Term: St. Vitus' dance

Ethimology: St. Vitus, 3rd century Christian child martyr

Genre: Compound noun

Spelling: \-ˌvīt-əs(-əz)-\

Related Forms:

Ortographic Variant: Saint Vitus's dance

Figure 1: Example of Terminological Record.

[20] https://github.com/gmdn/SherlockHolmesTriMED

Linguistic Analysis

In this section, we describe the findings of our analysis. Firstly, the terms extracted belong to the grammatical category of nouns. These nouns can be a single-word term (unigram) or a multi-word term (bigram or trigram), that is a complex lexical unit composed of more than one element. The unigrams extracted, when compared against MeSH terms as the reference corpus, belong to the domain of the human body. Many terms are related to body parts with a different index of frequency in the literary corpus: *hand* (745 occurrences), *eyes* (642 occurrences), *head* (452 occurrences), etc. Other unigrams extracted referred to various type of disease or medical condition such as *wheezing, breathlessness, apoplexy, blindness* etc. At the same time, there are a few bigrams relating to human anatomy and physiology such as: *lower limbs, upper arms, mitral valve, occipital bone, parietal bone, subclavian artery, and carotid artery* etc. Moreover, most of the bigrams extracted refer to different types of disease such as *rheumatic fever, typhoid fever, yellow fever, enteric fever,* or medical conditions as *cataleptic attacks, brain injury or chronic disease*. Regarding trigrams extracted the most two relevant multi-word terms in the medical field are the movement disorder *St. Vitus's Dance* and the surgical procedure *post-mortem examination.*

Thanks to terminological records, we were able to assign each term to a more specific medical subfield. Table 1 shoes some examples of our classification.

Medical subfield	Terminology
Cardiology	Aneurism, Heart Disease, Wheezing, Shock, Rheumatic Fever, Sudden Disease, etc.
Pathology (infectious diseases)	Consumption, Diphtheria, Enteric Fever, Yellow Fever, Tapanuli fever, Leprosy, etc.
Anatomy	Hand, Eye, Head, Heart, Arm, Shoulder, Finger, Brain, Neck, Nose, etc.
Neurology	St. Vitus's Dance, Catalepsy, Convulsion, Headache, Brain Fever, Stroke, Syncope etc.

Table 1: Example of medical subfields identified in Sherlock Holmes.

Moreover, terminological records allowed us to collect a variety of information for the lexico-semantic framing of each technical terms extracted for the medical field. The semantics of the term is useful for the detection of synonymic variants and their relations. For example:

"Post-mortem examination" is in a synonymic relation to *autopsy, obduction, necropsy, or autopsia cadaverum*. The term *autopsy* is more frequent than the term *post-mortem*, when considering the occurrences of both terms on the Web:

autopsy -> 23,200.000 occurrences,

post-mortem examination -> 8,090.000 occurrences.

Moreover, thanks to the semic analysis of the terms we could see that *necropsy* is usually used for animals, while *autopsy* is used for human beings.

“Typhoid fever” and “Enteric fever” are considered by Damiani as synonyms, while according to other sources *enteric fever* is a collective term that refers to both *typhoid* and *paratyphoid fever.*[21]

“Yellow fever”, as an acute febrile illness of tropical regions, is the result of a diachronic variation in terms of linguistic use of three obsolete synonyms: black vomit, yellow jack, Bulam fever.

The section related to the diachronic variation of a term make it possible to compare and contrast the use of the term in the past and its current use. For example:

“Consumption” used to indicate the process of general decay of the organism in place of the current *cachexia* or *wasting syndrome*.

“Brain fever” is an obsolete term used until the first half of the nineteenth century to indicate the association between an irregular set of neurological symptoms and it is now replaced by technical terms such as *encephalitis, meningitis, cerebritis*.

The section related to the diastratic variation of the terms also indicates the level of technicality of a term. In our study, we categorized those terms that are mostly used by lay people such as “popular terms” from the “specialized terms” such the exact terminology used by experts in medicine. We divided popular terms from technical ones such as “St. Vitus's dance” for *Còrea minor*, “Heart Disease” for *Cardiopathy* or “Nosebleed” for *Epistaxis*. In this way, we analyzed changes in the linguistic register by focusing on the diastratic variation resulting from the specialized-popular dualism in order to bridge the gaps between various registers.

For our linguistic analysis we used the previously mentioned software Sketch Engine in order to evaluate the syntactic behavior of the technical terms in their literary context. Our preliminary analysis, carried out through the use of the concordance tool, allowed us to notice the frequent use of some phrasal verbs associated to medical terms (e.g. “to break out” with “fever”). Moreover, with our quantitative approach we extracted the technical term “brainwave”. The term refers both to the medical domain as any rhythmic fluctuations of electric potential between parts of the brain, especially those seen on an electroencephalogram, and to a phraseological expression indicating a sudden idea, understanding, or inspiration. Considering the context in which the term appears with a frequency equal to 1, we excluded the term from our analysis of medical terminology because it was relevant to a phraseological expression:

> "I fear that we have." "Surely you do yourself an injustice. One more coruscation, my dear Watson--yet another brain-wave!”
> (The Valley of Fear, 1914)

[21] https://www.uptodate.com/contents/treatment-and-prevention-of-enteric-typhoid-and-paratyphoid-fever

Finally, another interesting consideration emerged from the study of the collocational behavior of terms is the association of the psychological condition of "depression" to the domain of colors. The examples below show that adjective "black" is frequently associated to depression with a metaphorical connotation suggesting a serious mood disorder.

> "He was bright, eager, and in excellent spirits,--a mood which in his case alternated with fits of the blackest depression."
> (The Sign of Four, 1890)

> "His bright humor marked the reaction from his black depression of the preceding days. Athelney Jones proved to be a sociable soul in his hours of relaxation and face his dinner with the air of a bon vivant."
> (The Sign of Four, 1890)

> "Even the triumphant issue of his labors could not save him from reaction after so terrible an exertion, and at a time when Europe was ringing with his name and when his room was literally ankle-deep with congratulatory telegrams I found him a prey to the blackest depression."
> (The Memoirs of Sherlock Holmes - The Reigate Squire, 1893)

Conclusions and Future Work

In this paper, we presented a preliminary study of medical terminology in the entire collection of works on Sherlock Holmes, starting from "A Study in Scarlet" (1887) to "The Casebook of Sherlock Holmes" (1927). The method proposed is based on the combination of i) a computational approach for the extraction of technical terms through the R programming language and ii) the qualitative analysis of medical terminology performed through the model of terminological record designed for the TriMED database. One of the objectives of our study was to reproduce the analysis proposed by Damiani (consisting in the manual extraction of medical terms) by adopting an automated approach. We were able to extract the medical terminology through the identification of all the MeSH terms in the stories of Sherlock Holmes with efficiency and accuracy. In addition, we extracted a larger number of terms than those identified by Damiani: 1,347 MeSH unigrams, 116 MeSH bigrams, 8 MeSH trigrams. Moreover, in order to favor the reproducibility of our experiments, we give access to the source code and the data used in the analyses of this paper.

Regarding the qualitative analysis performed by means of the TriMED terminological record, we considered some interesting linguistic features from the terminological viewpoint such as: synonymic relation, semantic behavior and terminological variation. In particular, sections related to diachronic and diastratic variation allow to lay the foundations regarding two specific aspects of technical terminology and stylistics. First, we aimed at demonstrating that technical terms designating concepts may change over time. This means that concepts are not defined and crystallized entities, but they can change depending on the historical era and culture.

Second, we were interested in Conan Doyle's stylistics and, in particular, to which extent his profession as a physician affects his writing. Thanks to the diastratic variation section, we will be able to conduct future studies in order to detect if the author tends to use a specialized terminology, while describing medical facts, or if he tends to use a "reader-oriented" terminology, that is a more understandable language adapted to an audience of potentially non-medical experts. In this sense, our future research question is to evaluate how much Conan Doyle the physician affects Conan Doyle the writer.

For the moment, the preliminary results of this analysis showed that the model of terminological record we propose can constitute a valid digital and systematic support, in different contexts and corpora, both in specialized and literary fields. Moreover, we tested the use of the concordance tool Sketch Engine in order to study the collocational behavior of technical terms in literary context. As future work, we intend to take advantage from other functions of the Sketch Engine tool, such as "Sketch Difference" in order to compare and contrast two words by analyzing their collocations and by displaying the collocates divided into categories based on grammatical relations. Finally, as terminological records are also designed for language professionals, we intend to focus on the translation viewpoint in order to conduct a comparative study between English medical terms extracted from the original texts of Conan Doyle and their Italian equivalents in the translated version.

References

[1] Ali, T., Schramm, D., Sokolova, M. and Inkpen, D. 2013. "Can I Hear You? Sentiment Analysis on Medical Forums." In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, Nagoya, Japan, 14-18 October 2013. pp 667–673. https://www.aclweb.org/anthology/I13-1077

[2] Brezina, V., McEnry, T. and Wattam, S. "Collocations in Context: A New Perspective on Collocation Networks." 2015. *International Journal of Corpus Linguistics* 20 (2): 139-73. https://doi.org/doi:10.1075/ijcl.20.2.01bre

[3] Cheng, W., Greaves, C. and M. Warren. 2006. "From N-Gram to Skipgram to Concgram." 2006. *International Journal of Corpus Linguistics* 11 (4): 411-33. https://doi.org/doi:10.1075/ijcl.11.4.04che

[4] Culpeper, J. 2009. "Keyness: Words, Parts-of-Speech and Semantic Categories in the Character-Talk of Shakespeare's Romeo and Juliet." *International Journal of Corpus Linguistics* 14 (1): 29–59. https://doi.org/10.1075/ijcl.14.1.03cul

[5] Damiani, E. 2016. "Elementare, Watson! Dizionarietto Medico Ad Uso Dei Lettori Di Sherlock Holmes". *Scienze Mediche*. CLEUP. ISBN-13:978-8867876440

[6] Eder, M., Rybicki, J. and Kestemont, M. 2016. "Stylometry with R: A Package for Computational Text Analysis." *The R Journal* 8 (1): 107-121.

[7] Elhadad, N., and Sutaria, K. 2007. "Mining a Lexicon of Technical Terms and Lay Equivalents." I*n Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. ACL. pp-49-56.

[8] Firth, J. 1957. "A Synopsis of Linguistic Theory 1930-1955". *Studies in Linguistic Analysis*. The Philological Society. Longman. 1952-59, pp 1-32.

[9] Fortier, P. 2002. "Prototype Effect vs. Rarity Effect in Literary Style." In *Thematics: Interdisciplinary Studies*. Benjamins. Chapter 21. pp 397-405.

[10]Franzini, G., Kestemont, M., Rotari, G., Jander, M., Ochab, J. K., Franzini, E., Byszuk, J. and Rybicki, J. 2018. "Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm." *Frontiers in Digital Humanities* 5: 4. https://doi.org/10.3389/fdigh.2018.00004

[11]Gouadec, D. 1990. "Terminologie - Constitution Des Données". 1990. AFNOR. ISBN:2124848119

[12]Grabar, N., Van Zyl, I., De La Harpe, R. and Thierry Hamon. 2014. "The Comprehension of Medical Words." In *Proceedings of the International Joint Conference on Biomedical Engineering Systems and Technologies - Volume 5*, 334–342. BIOSTEC 2014. ESEO, Angers, Loire Valley, France: SCITEPRESS - Science and Technology Publications, Lda. https://doi.org/10.5220/0004803803340342

[13]Guthrie, D. 1961. "Sherlock Holmes and Medicine." *Canadian Medical Association Journal* 85 (18): 996-1000.

[14]Holmes, D. I. 1998. "The Evolution of Stylometry in Humanities Scholarship." *Literary and Linguistic Computing* 13 (3): 111–17. https://doi.org/10.1093/llc/13.3.111

[15]Hoover, D. L. 2013. "Quantitative Analysis and Literary Studies." In *A Companion to Digital Literary Studies*, 517–33. Wiley-Blackwell. https://doi.org/10.1002/9781405177504.ch28

[16]Keselman, A., Tse, T., Crowell, J., Browne, A., Ngo, L. and Zeng, Q. 2007. "Assessing consumer health vocabulary familiarity: an exploratory study." *Journal of medical Internet research* 9(1): e5. doi:10.2196/jmir.9.1.e5

[17]Key, J. D. and Rodin, A. E. 1987. "Medical Reputation and Literary Creation: An Essay on Arthur Conan Doyle versus Sherlock Holmes 1887-1987." *Adler Museum Bulletin* 13 (2): 2125.

[18]Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P.

and Suchomel, V. 2014. “The Sketch Engine: Ten Years On.” *Lexicography* 1 (1): 7–36. https://doi.org/10.1007/s40607-014-0009-9

[19]Lancashire, I. and Hirst, G. 2009. “Vocabulary Changes in Agatha Christie’s Mysteries as an Indication of Dementia: A Case Study.” In *19th Annual Rotman Research Institute Conference, Cognitive Aging: Research and Practice*, pp 8-10.

[20]Mahlberg, M. 2013. *Corpus Stylistics and Dickens’s Fiction*. Routledge. ISBN:0415800145

[21]Mahlberg, M., Stockwell, P., de Joode, J., Smith, C. and Brook O’Donnell, M. 2016. “CLiC Dickens: Novel Uses of Concordances for the Integration of Corpus Stylistics and Cognitive Poetics.” *Corpora* 11 (3): 433-463. https://doi.org/10.3366/cor.2016.0102

[22]Moretti, F. 2013. “Distant Reading”. Verso Books. ISBN:1781680841

[23]Pearce, D. N. 1995. “The Illness of Dr George Tumavine Budd and Its Influence on the Literary Career of Sir Arthur Conan Doyle.” *Journal of Medical Biography* 3 (4): 236-238. https://doi.org/10.1177/096777209500300411

[24]Rastier, F. 1987. “Sémantique Interprétative. Formes Sémiotiques”. Presses universitaires de France. ISBN:2130740448

[25]Reed, J. 2001. “A Medical Perspective on the Adventures of Sherlock Holmes.” *Medical Humanities* 27 (2): 76–81. https://doi.org/10.1136/mh.27.2.76

[26]Richards, I. A. 2017. “Practical Criticism: A Study of Literary Judgement.” Routledge.
ISBN:1351497316

[27]Scott, M. and Tribble, C. 2006. “Key Words and Corpus Analysis in Language Education.” *Textual Patterns*. Benjamins. https://doi.org/10.1075/scl.22

[28]Semino, E., and Short, M. 2004. “Corpus Stylistics: Speech, Writing and Thought Presentation in a Corpus of English Writing”. Routledge. ISBN:0415286697

[29]Sinclair, J. 1991. “Corpus, Concordance, Collocation”. Oxford University Press. ISBN:0194371441

[30]Sinclair, J. 2003. “Reading Concordances: An Introduction”. Pearson/Longman. ISBN:058229214X

[31]Tognini-Bonelli, E. 2001. “Corpus Linguistics at Work”. *Studies in Corpus Linguistics*. Benjamins. ISBN:1588110613

[32]Tuzzi, A., and Cortelazzo, M. 2018. “What Is Elena Ferrante? A Comparative Analysis of a Secretive Bestselling Italian Writer.” *Digital Scholarship in the Humanities* 33 (3): 685–702. https://doi.org/10.1093/llc/fqx066

[33]Vezzani, F., Di Nunzio, G. M. and Henrot, G. 2018. “TriMED: A Multilingual

Terminological Database." In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018.* http://www.lrec-conf.org/proceedings/lrec2018/pdf/715.pdf

[34]Widdowson, H.G. 2008. "Text, Context, Pretext: Critical Issues in Discourse Analysis". *Language in Society.* Wiley. ISBN:0470758279

[35]Osborn, J. 2002. "Sherlock Holmes, and Evidence Based Medicine". In *Medicina nei Secoli - Arte e Scienza*, 14(2):515-528.

[36]Nordenstrom, J. 2007. "Evidence-based Medicine in Sherlock Holmes' Footsteps". Wiley, 1st edition. ISBN:9781405157131. DOI:10.1002/9780470750957

[37]Rodin, A. E. and Key, J. D. 1984. "Medical Casebook of Doctor Arthur Conan Doyle from practitioner to Sherlock Holmes and beyond". Krieger Publishing Company 1st Edition. ISBN:0898745926

Last URL access: 28/05/2019