CORPUS - BASED ANALYSIS OF TERM EXTRACTION FOR ENGLISH MEDICAL TEXTS

Hoàng Thị Khánh Tâm*

University of Foreign Languages, Hue University

Ngày nhận bài: 16/12/2016; ngày hoàn thiện: 14/1/2017; ngày duyệt đăng: 15/3/2017

Abstract

This is a methodological research that is conducted against the background of a context in which Content and Language Integrated Learning (CLIL for short) has been regarded as an innovative educational philosophy across Europe and it is to be adopted in Vietnam by the year of 2020. It is a corpus-based study that employs the complementary searches with a focus on the search precision and recall values, based on two elements namely specialised occurrences (with prefixes in Stedman's 2011 list) and frequency count (with a threshold at 12 times of appearance) to extract medical terms from 250 English medical texts that are included in the British Academic Written English (BAWE) corpus, which has been authorised to work on for the purpose of academic research. Thanks to the assistance of two free yet powerful statistical soft wares that are entitled AntConc and R (with logged instructions to be executed using the Text Mining package), a statistically workable definition of an English medical term is empirically established during the generation of a sample list of 45 items, with the validation carried out by 10 Vietnamese medical experts, both working in Vietnam and abroad, through an in-depth survey to analyse the key findings, followed by some pedagogical implications.

Key words: CLIL, corpus, extract, medical terms, statistical software

1. Introduction

Content and Language Integrated Learning (CLIL) was founded by Marsh (2002), recognised as an approach or philosophy by Ball (2008), and as an educational paradigm i.e. "fashion" by Van de Craen (2013). Previously a medical school teacher and currently a CLIL instructor in Italy, Ting (2010) stressed the significance of this marriage between science literacy (Medicine for one) and English language proficiency to maximise the learning and teaching motivation.

In the context of Vietnam, English medical texts have recently become all the more widely-used and widely-accessed among the college population. This paper therefore aims to present a workable definition of English medical terms using statistical soft-wares. It then purported to determine the professional validity of how the

* Email: khanhtam1907@gmail.com

terms were extracted, so that a CLIL English course in Medicine and/or an English-Vietnamese medical dictionary could be taken into account.

2. State of the Art

2.1. English medical terms

Bently (2010) introduced four CLIL-based types of word: (1) *content-obligatory vocabulary*, (2) *content-compatible vocabulary*, (3) *high and medium frequency words*, (4) *collocations and phrases*. The first type covered technical terms and jargons used in the subject. The second one referred to the general vocabulary of the subject and sometimes everyday situations; for example, the General Service List (GSL) by West (1953), the University Wordlist by Xue and Nation (1984), the Academic Word List (AWL) by Coxhead (2000), and the Medical Academic Word List (MAWL) by Wang, Liang, and Ge (2008). The third type, also known as functional words, pertained to the most often used vocabulary in general English, and are thus easy to be self-taught by learners themselves. The final type reflected the fixed combinations when it came to curricula content and concepts as in by way of illustration, those studies by Marco (2000) and Jaladi et al. (2015). The present body of research had covered three latter types of CLIL vocabulary, leaving the very first one, in this case English medical terms, untapped by corpus linguists.

2.2. Extraction of medical terms

From a linguistic point of view, Fabozzi (2010) underscored that, "a clinical terminology or controlled medical vocabulary is a structured list of concepts and associated descriptions used to describe diseases, procedures, treatments, medications, etc. and to codify the clinical information captured in an EHR [Electronic Health Records, explanation added] during the course of patient care" (p. 2). From a historical perspective, the advent of Greeks' rational medicine, as opposed to the traditional orthodoxy, observed a few Latin terms creeping into its terminology when Greek medical science migrated to Rome (Banay, 1948). Stedman's (2011) appendix of prefixes, suffixes, and combining forms, among others, would hence be our main source of reference; the list was comprehensive and claimed to be essentially reliable for the study of health professionals with its advanced features and rich content.

Upon extracting a term from a corpus, Fletcher (cited in Hundt et al., 2007) set a goal of maximising two aspects in Information Retrieval, namely precision, which included ***only***, and recall, which covered ***all*** matching database. Meanwhile, in their "establishment of a medical academic word list", Wang et al. (2008, p. 447) adopted and adapted the three principles applied by Coxhead (2000):

(1) *Specialised occurrence:* The word families included had to be outside the first 2,000 most frequently occurring words of English, as represented by West's (1953) GSL. (2) *Range*: A member of a word family had to occur at least 10 times in each of the four

main sections of the corpus and in 15 or more of the 28 subject areas. (3) *Frequency*: Members of a word family had to occur at least 100 times in the Academic Corpus.

2.3. Research questions

The literature summarised the necessity for a CLIL-based medical term list, the question of an English medical term to be characterized in a statistically-friendly way, and finally the analysis of term extraction in the British Academic Written English (BAWE, see Section **3.2**) medical text database that had thus far been left unexplored. Specifically, we were motivated to answer these two research questions (RQs):

RQ1. What is a statistically workable definition of an English medical term?

RQ2. What are the most frequent one-word medical terms in the BAWE medical corpus?

3. Materials and Methods

3.1. Research design

As Hunston (2002) delineated, the term *corpus* had four major characteristics. (1) It involved a **S**trategic collection of linguistics examples, with specific purposes in the designing process. (2) The linguistic examples were supposed to be **A**uthentic, featuring items that occurred naturally in real life. (3) A corpus was synonymous to a **G**igantic collection as compared with that of the few numbers of paper-based. (4) It was **E**lectronic where the means of storage and access were concerned. This English acronym represented *corpus* as **sage** on the stage; the list of written sentences or oral utterances could practically guide the learners to learn how to learn a certain language, and the teachers to practice their own linguistic teaching (Kennedy, 1998).

3.2. Tools of data collection

First of all, the BAWE corpus was downloaded with approval from Oxford Text Archive. The medical texts were then manually picked out from the entire corpus thanks to the FIND functionality running on BAWE Excel Database (Gardner & Nesi, 2012). After the pressing of the "Ctrl + F" cluster keys and simultaneous typing of the respective strings of "medic*" (with the asterisk standing for *medicine*, *medical*, and *medicinal*), "health" (for *health*, *healthy*, and *unhealthy*), "illness" along with "disease", a BAWE medical text database was generated, amounting to 250 .txt files covering 613.526 tokens, or running words, of student written material. The sub-corpus could be downloaded from goo.gl/D56CT3. Next, AntConc was employed as a freeware for corpus analysis in the context of classrooms (Anthony, 2004). Figure 1 showed the tool applied in this paper:

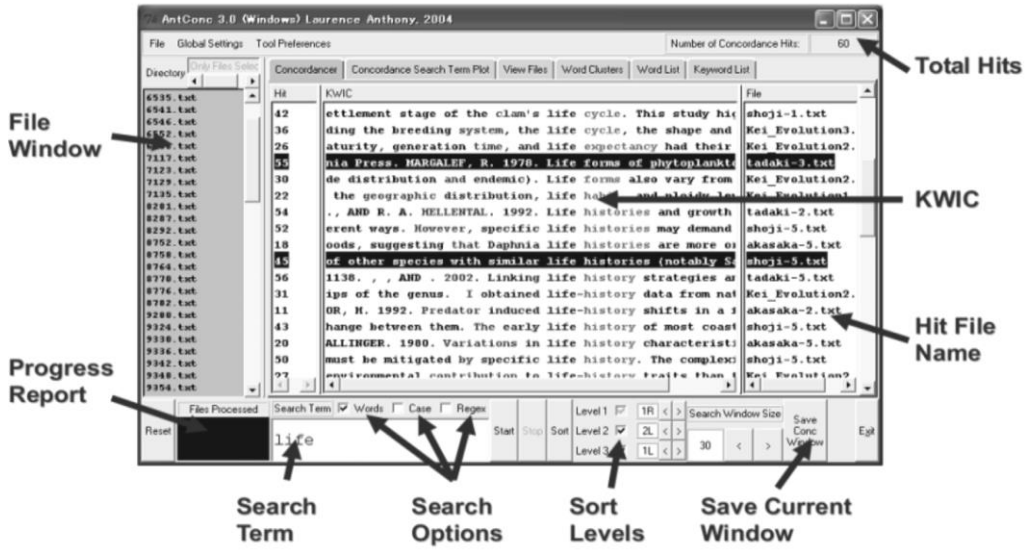

Figure 1. Concordancer Tool in AntConc

One drawback of AntConc was its missing index in use; however, where this freeware could not produce the most accurate results in large-scaled corpora, R fixed it up as a do-it-all software for a corpus linguist (Gries, 2009; Venables et al., 2014).

3.3. Tools of data analysis

With a survey for professional validation, the English medical term list was analysed by ten Vietnamese doctors whose specialties ranged from Cardiology, Dentistry, Family Medicine, Internal Medicine, Osteopathy, to Psychiatry and Public Health. They had been trained both in- and out-side Vietnam (for instance, America, Australia, Belgium, Denmark, France, Japan, Luxembourg, and The Netherlands).

To sum up, this corpus-based research adopted the complementary searches with **precision** and **recall** techniques, based on two elements namely **specialised occurrences** and **frequency count** in **BAWE medical database**, thanks to the combination of free statistical soft-wares **AntConc** and **R** to collect data, and then an **in-depth survey** on **ten Vietnamese medical experts** to analyse the results gathered.

4. Results and Discussion

4.1. Specialised occurrence

After a careful perusal throughout the literature review, we selected Fabozzi's (2010) explanation and just modified the context of the language use - "used by medical students who are writing essays or research articles" instead of "during the course of patient care" (p. 2). It permitted us to represent a wide range of functions that medical professionals are to fulfil and to categorise the complexity of various medical language features; it was

relatively straightforward to apply; and its conceptualization had cognitive, linguistic, and pedagogical values.

Delving into a recent financial lexis that had been built up by the RANGE program and then filtered by the AWL, Neufeld and his colleagues (2011) raised an alarming awareness of how uncritical and "indiscriminate" (to quote their own caution, p. 533) application of such vocabulary profiling tools to academic corpora. Ranging from the long-standing GSL (West, 1953), UWL (Xue & Nation, 1984, AWL (Coxhead, 2000) to the latest and most relevant MAWL (Wang et al., 2008), no corpus ought to be the panacea for every disease. In the present study, for the medical students to fully benefit from the English term list, we decided to take notice of the word stems which reflected more faithfully the academic nature of English medical wording profile.

On top of that, while their latest evidence strongly criticised the entire redundancy of high frequency common words (the third column in CLIL vocabulary in Bentley 2010) due to "the limitations of profiling tools that can lead to anomalies in statistical analysis and consequent misinterpretation of the data" (Neufeld et al., 2011), a more proper treatment and removal of these English stop words, which were also classified as common words to appear in a language like 'and', 'are', or 'of' (Williams, 2014, pp. 12-13), could be provided by a package called tm (a framework for Text Mining that was installed in R, ibid., p. 7). The present study opted for a focus on one-word medical term only, leaving intact medical collocations that were constituted by two or more words in the medical list under construction.

Lindmark, Natt och Dag, and Willners (2007) emphasized how time-consuming term extraction could be and how much manual work it could generally take. For the sake of time and effort, only medical terms starting with letters 'a' or 'b' were extracted for a detailed analysis and the rationale were twofold. On the one hand, the selection would be likely to produce a representative sample word list since 'a' is one of the most popular vowels and 'b' one of the least frequently used consonants in the English alphabet. Norvig (2013) was a dedicated advocate for this representative selection where Google English language corpus was used to update findings on Mark Mayzner's research into the frequency of English words and letters that was published and had been cited in multiple articles since 1965 (cited in ibid.). On the other hand, this paper only aimed to feature "an empirical cycle in which several rounds of data gathering, testing of hypotheses, and interpretation of the results follow each other" (Geeraerts, 2010, p. 73). In the words of Wehrli, Seretan, and Nerima (2010), "the small size of test set is motivated by the fact that the precision is expected to be very high" (p. 32), especially where the terms being extracted were closely scrutinised in the relevant corpus, and exceptions corresponding to the minority of cases would be more easily spotted thanks to the KWiC tool of AntConc.

Once the limitation of the beginning letters was imposed, we followed Lindmark and her associates (2007) in that:

> Whereas a terminologist normally spends a lot of time reading the material and trying to identify what words are typical for the domain, we decided to adopt a more mechanical approach... Most systems use statistics, shallow parsing or alignment of bilingual resources, and most resources are POS-tagged corpora. Since our corpus was not tagged, and since we wanted to use existing tools, we selected commercially available corpus linguistic analysis tool to find the words and phrases which could be considered domain specific terms, and also general language expressions that appeared to be used in a specific way, or were overrepresented. (pp. 369-370)

In effect, with AntConc's Clusters Tool, a basic 'a –' and '– b' wordlist from BAWE medical text corpus was first produced. After this initial automatic collection step, we applied the Concordancer Tool to verify and manually correct the results with an aim to eliminating any possible spurious hits. For example, any typos, numbers, formulas, abbreviated forms and proper names of an author or a publishing house were purposefully removed.

At first, we took advantage of every single item (prefix, suffix, and combining form) that was enumerated in the list proposed by Stedman (2011), with prefixes being among the most frequently used elements in the formation of words; suffixes being the terminal letters of syllables added to the stem to modify or amplify its meaning; and compound words being defined as terms which have a second stem as a component part (Banay, 1948). However, as the term extraction procedure progressed, it was noticed that the suffixes more often than not only made the inflections ('–ing' or '–ed' forms of a verb; adjectival or adverbial forms of a noun; or singular and plural forms of an accountable one); we stopped examining every affix and focused only on prefixes later on. There a new hypothesis occurred as to a medical term was one that started with 'ab' (e.g. 'abduct', 'absent') or 'bio' (like 'biology', 'biography'). In order to leverage the available list, the rest of the word forms were grouped under one item with an asterisk, for instance, 'abdomen*' ('abdomen', 'abdominal', 'abdominis', 'abdomino-perineal'), following the terminology integration principle of organising knowledge by concept in the Unified Medical Language System by McCray and Nelson (1995).

4.2. Frequency count

In the resulting list, which featured 45 word families extracted from 28,415 types and 613,808 tokens, the most frequently used item was cited as 'admission' standing at 284 occurrences and the least popular item was 'albumin' accounting for the threshold (minimum frequency) of 12 times (Hyland, 2008). The medical term list was then tabulated based on the frequency and not alphabetical order like the AWL and UWL. Flowerdew (2008) made a good point in striving for a list "without hierarchical relation between the

terms" (p. 625) because it would be easier for any searches without the aid of FIND functionality on the computer. For the sake of educational purposes, nevertheless, this frequency-based order would naturally expose the language users to the language itself without any manipulation; it turned out to work more effectively from the largest to the smallest numbers of counts.

Among the three aforementioned word lists, we focused on MAWL (Medical Academic Word List by Wang et al., 2008), and not UWL (University Word List by Xue & Nation, 1984) or AWL (Academic Word List by Coxhead, 2000). Firstly, the UWL had been compiled across any fields but Medicine, which proposed to our medical term list a niche in the relevant literature. Secondly, the AWL deemed for the same purpose as ours, which was inspection of essays in the British context and not necessarily by British native speakers who were undergraduates, the former mainly covered New Zealand English and American English only. Finally, the assumption that the basic items of English lexis "should be familiar to most students entering universities" turned out to be not realistic in Vietnam where this study was based. As far as the AWL was concerned, Neufeld and his team (2011, p. 535) illuminated that

These top 30 'general words' would appear as 'academic' as the ones in the first column from the so-called AWL, which really brings us to consider whether the AWL can usefully serve as a generic list of academic lexis, especially as it was constructed as 'an artefact of the GSL' (Cobb, 2010).

With reference to MAWL, there were 10 out of 43 lemmas to repeat in our current medical term list namely 'abdomen' (or 'abdominal'), 'absorb', 'acid', 'acute', 'adverse', 'algorithm', 'antibiotics', 'antigen', 'bacteria' (or 'bacterium'), and 'biopsy'. The low percentage was probably because of the density of purely terminological jargons in our list of English medical terms.

4.3. Professional validation

The frequency of the medical terms, which had been proposed as "***stranger***", "***acquaintance***", "***friend***", "***best friend***", "***sweetheart***", or "***family member***" in the survey, corresponded with the increasing size of the number of hits yielded in AntConc out of the BAWE medical text documentation, to very few exceptions, and were consequently retained without major changes in the ordered term list. As for the next question, in spite of the theoretical suggestion by Neufeld et al. (2011), the inclusion of various forms under one asterisked item (with a linguistic concentration on word stems) received a four-fifths agreement thus stayed the same; nevertheless, there were several changes like in Case 6 'bronchi' becomes the head word because Respondent 1 thought that 'bronchial' was too long and not major enough. 'Adenosine' and 'adenosylmethionine' were deleted in Case 3 because both referred to an acid amine and were totally irrelevant to 'adeno' meaning glands, as was advised by this medical doctor. One might also argue that the suggestive list should be

otherwise simpler without the compilation of stemming items; nonetheless, "in most cases learning the derived form requires very little extra work once the base form is known (Xue & Nation, 1984, p. 216).

In a similar vein, 'thelarche', 'attosecond', 'acuhaler', 'absent', 'biotin', 'aqueous', 'adrenalin', 'menarche', and 'aura' were suggested to be removed from the current list by other respondents as these words did not appear regularly in the medical content. Judging by the number of hits using Concordancer Tool in AntConc, all but the last three items were reserved. Especially, 'biotin' and 'adrenalin' were said to be chemical substances (Respondent 3), but "chemical compound words are formed very irregularly. They are hybrid (using Greek and Latin stems combined in one word)." (Banay, 1948, p. 17) whereby very much deserved their due position for medical reference. With regards to the term 'absent', Flowerdew (2008, p. 43) keenly observed that,

Goodman and Payne's (1981) definition of technical terms having congruity among scientists (unlike the term 'cell', for example, which has a different meaning in biology to that in general English). Here, we have an example of determinologization which refers to a process whereby specialist terms such as those relating to computers make their way into general language through the mass media or direct impact (Bowker & Pearson 2002).

She further reminded that collocations should be classified as a set of technical words because they were terms specialised to the relevant specific domain, even though each separate word in the combination was likely to occur in general English. Our term list did not include any terms with more than one word, as a medical term traditionally was, but this observation should be heeded when we were working with another term list in the future.

Specifically, Respondent 9 advised us to employ the software Medic 2.7 for more information. This applied to the method of external cross–references that was once put forward by Bodenreider (2004). The *Medical Terminology for Health Professions* (7th edition) by Ehrlich and Schroeder (2013) and *Medical Terminology: A self-teaching guide* (4^{th} edition) by Steiner (2003) also proved to be practical for cross references to other terminologies or database, which should be feasible as medical lexis had been abounding thus far. This was also why the specific domain of Medicine was selected in the first place (Wermter, 2009).

Apart from this, Respondent 4 urged for a medical term list presented with images, videos, or animation (if possible); Respondent 8 drew attention to how the terms should be pronounced in a correct way; Respondent 10 suggested that, "the list should have been categorized into specialized majors". Respondent 7 shared complete agreement with this suggestion in that he recommended each term attached with the corresponding individual field for faster information seeking. These were precious features that might well boost the pedagogic value of the existing medical term list.

5. Conclusion and Implications

5.1. R log for a corpus-based frequency count of English medical term list

The R log to conduct the frequent count of the English medical term list in this research was delineated as follows:

First of all, we have to go to the folder that contains the text documents and then load a sample collection within a folder named "BAWEMedicalCorpus"

```
setwd("D:/KTthesis")
getwd()
cname = file.path (".", "BAWEMedicalCorpus", "txt")
```

After loading the tm (Feinerer & Hornik, 2014) package into the R library we are ready to load the files from the directory as the source of the files making up the corpus, using DirSource( ). The source object is passed on to Corpus ( ) which loads the documents. We save the resulting collection of documents in memory, which should then be stored in a variable called *medicalterms*.

```
library(tm)
medicalterms = Corpus(DirSource(cname))
medicalterms
```

Generally, the text data should be pre-processed to get ready for the text analysis. The basic transforms are all available within the package tm (which accounts for Text Mining). We will apply each of the transformations, one-by-one, to remove unwanted characters from the text.

```
library(tm)
medicalterms = tm_map(medicalterms, removeNumbers)
medicalterms = tm_map(medicalterms, removePunctuation)
medicalterms = tm_map(medicalterms, removeWhitespace)
medicalterms = tm_map(medicalterms, content_transformer(tolower))
inspect(medicalterms[13])
```

Next, we create a document term matrix, which is simply defined to be "a matrix with documents as the rows and terms as the columns and a count of the frequency of words as the cells of the matrix" (Williams, 2014, p. 17).

```
dtm = DocumentTermMatrix(medicalterms)
dtm
dtm = sample(1:10, 100, replace=T)
x = sort(table(dtm), decreasing=T)
write.csv(x, "mytable.csv", quote=F)
```

5.2. Conclusion

By and large, after a process of trials and errors, the final definition of an English medical term that we managed to come up with was as follows:

An English medical term is one that describes in English a disease, procedures, treatments, medications, etc. and codifies the clinical information used by medical students who are writing essays or research articles under the university context. It is made up of any one from the list prefixes or combining forms in Stedman's (2011), excluding inflections, capitalizations, and abbreviations. More importantly, it has to be scrutinised by medical experts for professional validation.

Table 1. Sample medical term list extracted from BAWE corpus

Nº	**Medical term**	**Hits**
01	admission(s)	284
02	abdomen* (abdomen, abdominal, abdominis, abdomino-perineal)	237
03	acute(ly)	219
04	angina* (angina, angioplasty, angiographic, angiogenesis, angiogram, angiography, anginaNO, angiotensin)	167
05	arthritis* (arthritis, hemoarthrosis, arthroplasty, athroconidia, athropathy, arthroscopy, arthroscopic, osteoarthritis)	167
06	artery* (artery, arteries, arterial, arteriogram, arteriosus, arteritis, arterioles)	158
07	acid* (acid, acidaemia, acidic, acidosis, acidotic, acids)	131
08	Anaemia	118
09	abnormal* (abnormal, abnormally, abnormality, abnormalities)	115
10	absorb * (absorb, absorbs, absorbed, absorbing, absorbance, absorption)	114
11	abuse* (abuse, abused, abuser, abusers, abusive)	109
12	abort* (abort, aborted, aborting, abortion, abortions)	108
13	adeno* (adenocarcinoma, adenocarcinomas, adenocarinoma, adenolymphoma, adenoma, adenomas, adenomatous, adenoviruses)	095

5.3. Implications

The key findings above have illuminated their pedagogical implications for the Vietnamese learners in a CLIL context, both linguistically and medically. Once the learner is aware of his/her own knowledge about the English medical terms to master in advance, it is recommended that he/she try to use the learning strategies that make the most of their language drill and competence. Also, once the learner's potential and resourcefulness are tapped, more self-confidence is supposed to ensue, and the students should no longer rely on the teacher to have everything ready for them to achieve a certain level of another language that is required from the CEFR exams. Instead they can create their own activities at home and with their peers to build up their linguistic command accompanied by the wealth of medical information written in a foreign language while catering for enhancement of the learner autonomy. In this way, the self-study approach should be given a special significance.

Not only do the language learners benefit from the research results but there are also pedagogical implications for the content and language teachers. The word families in the list are worthy of consideration while a certain English for Medical Purpose course in Vietnam is designed and a course book along with relevant handouts are prepared, where a CLIL medical class is about to come into operation by 2020 (The Government of Vietnam, 2008). The list can also make a helpful reference for a Medical English lexis curriculum where Xue and Nation guided that,

The high frequency words deserve individual attention. The best approach to dealing with the low frequency words is to teach ways of dealing in context rather than "teaching" the words themselves. (1984, p. 215)

For the sake of further research, an informal version of English medical term list needs to be prepared so that not only in university essays and research articles will the students be able to improve their English competence but also throughout their communication with the patients, preferably using less "medicalese" and more daily conversation in English. The sub-corpus will additionally become more valid and reliable thanks to a rechecked replication in larger corpora.

References

Anthony, L. (2004). AntConc: A learner and classroom friendly, multi-platform corpus analysis toolkit. *Proceedings of IWLeL 2004: An Interactive Workshop on Language e-Learning*, pp. 7-13. Tokyo: Waseda University.

Ball, P. (2008). *What is CLIL?*. Retrieved on July 8, 2013, from http://www.onestopenglish.com.

Banay, G. L. (1948). An introduction to medical terminology I. Greek and Latin derivations. *Bulletin of the Medical Library Association, 36*(1), 1.

Bentley, K. (2010). *The TKT course CLIL module*. Cambridge: Cambridge University Press.

Bodenreider, O. (2004). The unified medical language system (UMLS): Integrating biomedical

terminology. *Nucleic Acids Research, 32*, 267-270.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Ehrlich, A., & Schroeder, C. L. (2013). *Medical terminology for health professions* (7th ed.). Clifton Park, NY: Delmar, Cengage Learning.

Fabozzi, N. (2010). Kaiser's donation of its convergent medical terminology dictionary puts the spotlight on the role of clinical terminology services in driving meaningful use of EHRs. *Healthcare and Life Sciences, Frost and Sullivan.*

Feinerer, I., & Hornik, K. (2016). *tm: Text mining package*. R package version 0.6. Retrieved from http://CRAN.R-project.org/package=tm.

Fletcher, W. H. (2007). Concordancing the web: promise and problems, tools and techniques. In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus linguistics and the web* (pp. 25-46). Amsterdam: Rodopi.

Flowerdew, L. (2008). *Corpus-based analyses of the problem-solution pattern*. Amsterdam: John Benjamins.

Gardner, S., & Nesi, H. (2012). A classification of genre families in university student writing. *Applied Linguistics, 34*(1), 1-29.

Geeraerts, D. (2010). The doctor and the Semantician. In D. Glynn & K. Fischer (Eds.), *Quantitative methods in cognitive semantics: Corpus-driven approaches* (pp. 63-78). Berlin: De Gruyter Mouton.

Gries, S. T. (2009). *Quantitative corpus linguistics with R: A practical introduction*. The United Kingdom: Taylor & Francis.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*, 4-21.

Jalali, Z.S., Moini, M.R., & Arani, M.A. (2015). Structural and functional analysis of lexical bundles in medical research articles: A corpus-based study. *International Journal of Information Science and Management, 13*(1), 51-69.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.

Lindmark, K., Natt och Dag, J., & Willners, C. (2007). Lexical semantics for software requirements engineering – a corpus-based approach. In R. Facchinetti (Ed.), *Corpus Linguistics 25 years on* (pp. 365-385). Amsterdam: Rodopi.

Marco, L. (2000). Collocational frameworks in medical research papers. *English for Specific Purposes, 19*, 63-86.

Marsh, D. (2002). *CLIL/EMILE - The European dimension: Actions, trends and foresight potential*. Brussels, Belgium: The European Union.

McCray, A. T., & Nelson, S. J. (1995). The representation of meaning in the UMLS. *Methods of information in Medicine, 34*, 193-201.

Neufeld, S., Hancioğlu, N., & Eldridge, J. (2011). Beware the range in RANGE, and the academic in AWL. *System, 39*, 533-538.

Norvig, P. (2013). *English letter frequency counts: Mayzner revisited or ETAOIN SRHLDCU*. Retrieved on June 1, 2014 from http://norvig.com/mayzner.html.

Stedman, T. L. (2011). *Stedman's medical dictionary – illustrated in colour* (28th ed.). Philadelphia, PA: Lippincott Williams & Wilkins.

Steiner, S. S. (2003). *Quick medical terminology: A self-teaching guide* (4th ed.). Hoboken, NJ: John Wiley & Sons.

Ting, Y-L. T. (2010). CLIL appeals to how the brain likes its information: Examples from CLIL-(Neuro)Science. *International CLIL Research Journal, 1*(3), 13-73.

Van de Craen, P. (2013). The emergence of a new paradigm. *Approaches to language teaching and learning for multilingual education (December 18, 2013)*. Lecture conducted from Vrije Universiteit Brussel, Brussels, Belgium.

Venables, W. N., Smith, D. M., et al. (2016). *An introduction to R - Notes on R: Programming environment for data analysis and graphics*. Retrieved February 16, 2016 from CRAN.R-project.org.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes, 27*, 442-458.

Wehrli, E., Seretan, V., & Nerima, L. (2010). *Sentence analysis and collocation identification*. Beijing: COLING Workshop on Multiword Expressions (MWE 2010).

Wermter, J. (2009). *Collocation and term extraction using linguistically enhanced statistical methods*. Thuringia, Germany: Friedrich Schiller University of Jena.

West, M. (1953). *A general service list of English words with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London: Longmans.

Williams, G. (2016). *Hands – on data science with R: Text mining*. Retrieved 16 February 2016 from Graham@togaware.com.

Xue, G., & Nation, I.S.P. (1984). A university word list. *Language Learning and Communication, 3*(2), 215-229.

PHÂN TÍCH DỰA TRÊN KHỐI NGỮ LIỆU ĐỂ TRÍCH XUẤT THUẬT NGỮ TỪ CÁC VĂN BẢN Y HỌC TIẾNG ANH

Tóm tắt. CLIL (Content and Language Integrated Learning) là triết lý giáo dục có tính cải tiến ở Châu Âu có thể sẽ áp dụng ở Việt Nam trước thềm 2020. Hai kĩ thuật tìm kiếm bổ trợ nhau là "precision" (tính chính xác) và "recall" (tính toàn diện); hai yếu tố được phân tích nhằm trích xuất thuật ngữ y học bằng tiếng Anh từ tài liệu văn bản y học của BAWE (British Academic Written English) là "specialized occurrence" (sự xuất hiện của từ chuyên môn, dùng AntConc) và "frequency count" (mức độ xuất hiện thường xuyên của thuật ngữ đó, dùng phần mềm thống kê R). Định nghĩa thuật ngữ y học tiếng Anh dựa theo thống kê đã được thiết lập qua quá trình thực nghiệm xây dựng bộ thuật ngữ đơn cử với 45 mục từ kiểm định bởi 10 chuyên gia y tế Việt Nam trong và ngoài nước.

Từ khoá: CLIL, khối ngữ liệu, trích xuất, thuật ngữ y học, phần mềm thống kê