UNIVERSITÀ DEGLI STUDI DI MILANO

Dipartimento di Lingue e Letterature Straniere

Dottorato di Ricerca in Lingue, Letterature e Culture Straniere

XXVI Ciclo

ANALYSING MEDICAL ENGLISH LEXIS: A CORPUS-BASED RESEARCH IMPLICATIONS FOR LEARNING AND TEACHING

L-LIN/12– Lingua e Traduzione - Lingua Inglese

Tesi di Dottorato di:
Francesca Ripamonti
R09349

Tutor:
Chiar.ma Prof.ssa Luciana Pedrazzini

Co-tutor:
Chiar.mo Prof. Giovanni Iamartino

Coordinatore del Dottorato:
Chiar.mo Prof. Giovanni Iamartino

Anno Accademico 2013/2014

To my father

Acknowledgments

I would like to express my gratitude to my supervisor, Prof. Luciana Pedrazzini for her invaluable comments on earlier drafts of this dissertation and for her inspiring and focused insights. I have learnt a lot from her throughout my PhD course.

My sincere thanks also go to Prof. Giovanni Iamartino, for his challenging and constructive suggestions, for his guidance and encouragement and for enabling me to gain insight into the fascinating world of Corpus Linguistics studies.

Special thanks go to the several people who have contributed with insights and professional assistance at different stages of this research.

I am also indebted to my family, my husband and son, for their patience and their immeasurable support.

Abstract

In the last few decades medical specialized communication has become progressively dynamic and prolific with an ever growing number of researchers employing English as a lingua franca. Medical specialized communication unquestionably constitutes a challenging problem with non-native medical students and health care practitioners who are increasingly faced with the need to have an active command of Medical English, that vast set of standardized and non-standardized terms used to describe and represent the changes and the results accomplished in the medical field.

My direct involvement in a language trial test aiming at investigating the productive vocabulary knowledge of a group of Italian medical undergraduates was the background that provided the incentive for this dissertation to investigate the language of medicine and devise strategies and materials that were specific enough to help non-native speakers from different medical fields acquire the English language skills they needed step by step.

Trying to find the convergence between non-native medical doctors' vocabulary needs and pedagogy, my research has developed along two lines, purposefully called 'diagnosis' and 'remedy'. Firstly, I endeavoured to 'diagnose' which key-words are homogeneously distributed across mainstream professional medical writings opportunely collected in the Medical English Corpus (MedEnCor), a specialized corpus extensively representative of current healthcare domains and biomedical topics. Secondly, I attempted to seek a 'remedy' to the non-native learners' lexical 'impairments' categorizing all the extracted key-words into semantic wordlists, suitably catalogued into the Medical English Corpus Lexical database (MedEnCor-Lex), a web-based monolingual glossary (www.medencor.com) meant to provide non-native users not only with denotative information on medical key-words, but also with appropriate instances of their collocational and phraseological context and use.

Although this writing tool is currently being completed, my goal is equally educational and professional because by compiling a specialized lexical database I do not only mean to make the English used in medicine accessible to the healthcare community, but, first and foremost, to make non-native recipients familiarize themselves with the terms and expressions relevant to the scientific register. An essential skill for their career.

Contents

List of abbreviations

AWL	Academic Word List
BNC	British National Corpus
CEFR	Common European Framework of Reference
CLIL	Content and Language Integrated Learning
DDL	Data-Driven Learning
EAP	English for Academic Purposes
EEP	English for Educational Purposes
EFL	English as a Foreign Language
EGAP	English for General Academic Purposes
EMP	English for Medical Purposes
EOP	English for Occupational Purposes
ESP	English for Specific Purposes
L1	First language
L2	Second language
MAWL	Medical Academic Word List
RA	Research Article
TTR	Type/Token Ratio

List of figures

Appendix 1

List of tables

Appendix 1

Chapter 1

Introduction

1.1 Introduction

> *"It ought not to be difficult to know what a word is [...] Yet when we look a little more closely, a word turns out to be far from the simple and obvious matter we imagine it to be"*
>
> M. A. K. Halliday (2005, 1)

Any given language is constituted by all the lexical elements that become part of it. They all enjoy equal status because each individual word contributes to the construction of meanings. Yet, only few words are *key-words*[1], *i.e.* words that play a pivotal role in identifying important elements in a given area or text. They are vectors of communication other than knowledge. This applies to poetry where the greater o lesser intensity poured by the poet in the choice of words results in lesser or greater elation shared by the reader and, at a less

[1] Given the prominent role played by *key-words* in this research, emphasis has been intentionally added.

aesthetic level but with equal relevance, to specialized languages where the right choice and use of words have fundamental implications for the successful circulation of ideas in each target discourse community. Medical English, as a specialized sublanguage of English, is no exception.

1.2 Overview of the research

This dissertation is divided into seven chapters and one case-study (Appendix one). The first chapter introduces the concept of medical specialized communication and tries to explain the reasons why it is important for junior clinicians, biomedical researchers and senior physicians to learn Medical English, today. The second chapter provides the theoretical framework to the study with a detailed description of the various steps of the research: starting from a real needs analysis evaluating the vocabulary 'necessities' of a target group of medical undergraduates (the full account of the pilot study is given in Appendix 1) up to the definition of specific aims and research questions aimed at addressing these linguistic 'impairments'. As the present study draws heavily on earlier findings in EMP corpus-based research related to written language, a detailed discussion of previous studies, theoretical influences, approaches and subjects involved is given in the literature review of Chapter 3 with also final considerations on gaps in previous

research. Chapter 4 focuses on methodology and corpus linguistics: it begins by framing the essentials of corpus-building, it presents the *MedEnCor*, a specialized medical corpus and the methodology followed for its compilation and key words extraction (a full list of the top 500 keywords used in the study is included in Appendix 2). Chapter 5 discusses the results of the data analysis and the categorizations of semantic groups of keywords defined as a sample of the *core lexis* of medicine, thus deserving particular attention by non-native medical learners. Chapter 6 presents the pedagogical implications of the study describing the *MedEnCor-Lex*: a web-based medical glossary which returns medical keywords in their collocational and phraseological context if properly queried. Finally, the main conclusions are drawn in Chapter 7, outlining the contributions of the present research to corpus linguistics, to medical register analysis and to EMP pedagogy. Suggestions for further research are also discussed.

1.3 Background of the study

In the last few decades medical specialized communication has become increasingly dynamic and prolific with an ever growing number of researchers engaged in international mobility[2] or involved in academic

[2] Academic mobility in Europe is favoured by the European Union's efforts to enhance students or staff exchanges and academic internationalization through the

projects employing English as a *lingua franca* (Birch-Bécaas 1994, Salager-Meyer 1997; Crystal 2003; Mauranen 2009, 2011; Seidlhofer 2011; Björkman 2013; Gotti 2014). English medical language has also established its global position with about 2 million health-related articles published annually worldwide[3] (Maher 1986a; Cooter 2000; Règent 2000; Gotti 2006; Baethge 2008) and with 80% of online specialized information consulted by more than two-thirds of the world's scientists, practitioners and scholars (Crystal 1995; Flowerdew 2000; 2001; Graddol 2000; 2008, Berghammer 2008; Molhim 2011). Proportionally, a variety of EMP (English for Medical Purposes) courses have evolved into curriculum components in most university faculties, while Medical English proficiency has become a mandatory standard required to

---

Bologna Process: a set of accords signed by forty-nine European Ministers of Education. The collective goal is to create a 'European higher education area' by presenting academic degree standards more comparable across Europe thus making European higher education and research more attractive to non-Europeans. (*The official Bologna Process website: 2007–2010).*

[3] According to the *2012 English-language STM* (Scientific, Technical and Medical) *publishing report* there were about 28,100 active scientific and scholarly peer-reviewed journals in mid 2012, collectively publishing about 1.8–1.9 million articles a year. The USA dominates the global output of research papers with a share of about 21 percent, China has moved into second position with 10 percent of global output followed by the United Kingdom (7%), Japan (6%), Germany (6%) and France (4%). (An overview of scientific and scholarly journal publishing. *The STM report*. Nov. 2012). As for Italy, Giannoni (2008) reports that 99 percent of Italian-authored biomedical research publications are now in English, with the national language reserved for less research-intensive local publications.

The anglicization of medical science is also confirmed by the *Index Medicus/Medline* American journal catalogue where the percentage of English-language journals has risen from from 35% to 89% in the last 130 years (Baethge, 1)

doctors on registration to professional boards or on bestowal of licenses to practice. (Allum, Wright & McCullagh, 2013).

1.4 Statement of the problem

The predominant use of the English language in the world of research and scholarship together with the accelerating progress in biomedical studies have brought to the fore the rising communicative needs of the scientific community where new achievements must be appropriately expressed and rapidly circulated. Specialized communication unquestionably constitutes a challenging problem with non-native medical students and health care professionals who are increasingly faced with the need to have not only a passive but also an active command of Medical English, an essential skill for their career. They need to acquire, understand and develop a high degree of proficiency in academic terminology, that "vast set of standardized and non-standardized terms used to describe and represent the changes and the results accomplished in the medical field" (Maglie, 2009, 15).

The aim of this study is primarily intended to fill this gap. Trying to find the convergence between non-native medical doctors' vocabulary needs and pedagogy, my research developed along two lines, purposefully called 'diagnosis' and 'remedy'. First I endeavored to 'diagnose' which *key-words* were homogeneously distributed across

mainstream academic and professional medical writings opportunely collected in the *Medical English Corpus* (MedEnCor), an *ad hoc* specialized corpus containing documents about health and biomedical topics, specifically designed to analyze the scientific discourse and vocabulary. Second, I attempted to seek a lexical 'remedy' categorizing all the extracted *key-words* into semantic wordlists, subsequently catalogued into the *Medical English Corpus Lexical* database *(MedEnCor-Lex)*, a web-based monolingual glossary (www.medencor.com) meant to provide non-native users not only with denotative information on medical *key-words,* but also with appropriate instances of their collocational and phraseological context and use.

My goal was equally educational and professional because by compiling a specialized lexical database I did not only mean to make the English used in medicine accessible to the healthcare community, but, first and foremost, to make the non-native recipients familiarize themselves with the terms and expressions relevant to the scientific register.

1.5 Specialized communication and English for Medical Purposes (EMP)

In applied linguistics the terms 'specialized communication', 'domain-specific languages'[4] or 'language for specific/special purposes (LSP)'[5] have been coined to designate that kind of language use associated with 'communication among specialists' (Fuertes Olivera 2005, p. 41) characterized by 'peculiar linguistic signs which makes it less accessible for those who do not have adequate background knowledge in the field' (Garzone, 2006, 9). Given its enormous range of domains, the concept of specialized communication is neither monolithic nor uniform, but is defined with reference either to the professional, disciplinary or technical field to which it pertains or to the users' specifiable working purposes and communicative needs, *i.e*. English for Medical Purposes (Widdowson, 1983; Dudley-Evans & St Johns, 1998; Johns & Price Machado, 2001). This conforms to both the general definition of ESP (English for Specific Purposes) "whose main concerns have always been with needs analysis and effective communication in the tasks prescribed by the study or work situations" (Dudley-Evans and St John 1998, I) and

[4] Also called "special languages, microlanguages or technolets" (Berruto 1980: 29 in Garzone 2006, 9).

[5] It should be noted that the broad term *LSP (Language for Specific Purposes)* indicates that the specific purpose approach can be applied to any language. Since this study focuses primarily on the teaching and learning of specialized communication in English, the term *ESP (English for Specific Purposes)* will be used throughout.

also to its *absolute characteristics*[6] whose contents should meet the learners' specific needs and whose lexis should be appropriate to their particular occupations or disciplines (Strevens 1988, 3).

The above explanations make it clear that language use and language needs are the two complementary forces responsible for the diversification of specialized communication. Accordingly, I will try to cast light on the leading role that both forces have played in the categorization of the language of medicine as a special language, which, relying on its own lexicon, syntax and rhetorical organization and being shared with different degrees of adaptation by the whole scientific community is better qualified as "English for Medical Purposes" (EMP).

As for language use, nowhere more than in medicine, has specialized vocabulary and effective communication been widely recognized as central to clinical outcomes and their dissemination. Conversely, when it comes to analyze language needs, it is important to differentiate whose needs EMP refers to; whether medical students', practicing doctors' or consultants' in hospitals. Each of these groups needs different degrees of language specificity, appropriate to each educational institution or workplace. Medical students, for example, are expected to read textbooks and articles as well as write essays and short clinical reports. Practicing doctors and consultants, on the other hand, have different requirements which may include publications of research articles in

[6] The *variable characteristics* of ESP involve the specific skills to be learned and the teaching methodology to be adopted (Strevens 1988 in Dudley-Evans and St John 1998, 3).

international journals, presentations of papers for conferences or interactions with patients. Hence, EMP qualifies as an umbrella term encompassing distinct specialisms basically merging into two macro areas: English for Academic Purposes (EAP) and English for Occupational Purposes (EOP) (Dudley-Evans and St John, 1998, 7), depending on whether the focus is on learning specialized English for academic study or for work and professional training (Figure 1.1).

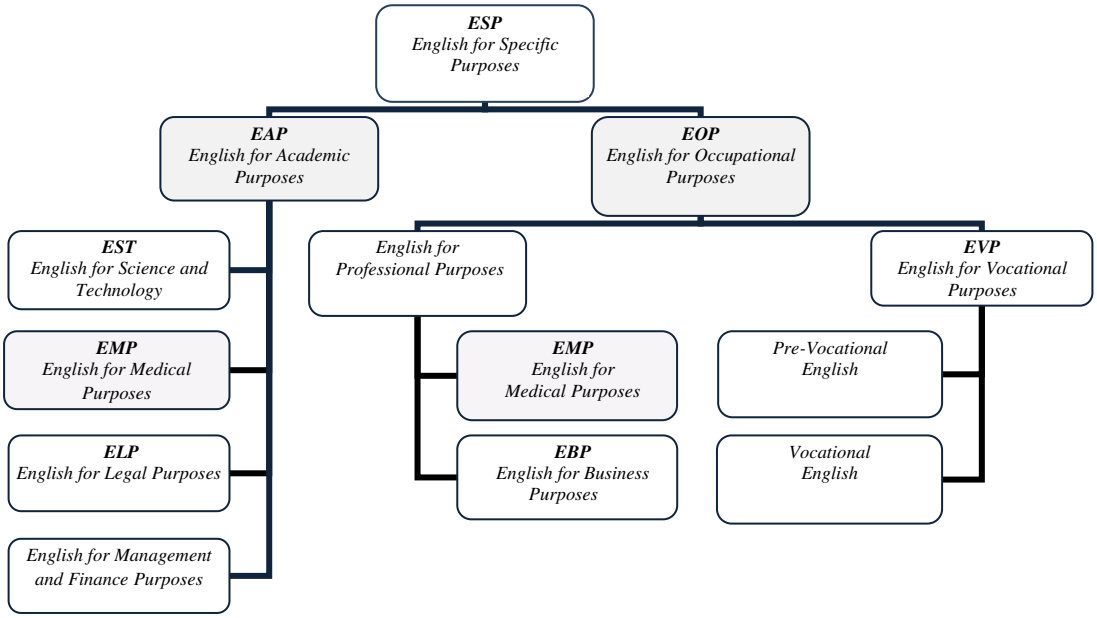


Figure. 1**.1.** *ESP classification by academic and occupational areas*

This classification beside visualizing learners' specific needs it also highlights two other prominent paradigms in specialized communication: the 'disciplinary specificity'. (Swales, 1985, 181) and the 'delicacy of context' (Richards 1989, 215), thus confirming that language-using

communities are at the heart of each subject discipline in that they provide the context within which the constituents learn to communicate and to interpret each other's jargon, gradually acquiring the recognized discourse conventions to participate as members (Hyland 2011, 11-12).

1.5.1 English for Medical Purposes

The term "English for Medical Purposes" (EMP) in Maher's pioneering definition (1986b) referred to "the teaching and learning of English for doctors, nurses and other personnel in the medical professions […] for a utilitarian purpose and identifiable goal –typically, the successful performance of work or the optimum effectiveness of medical training" (112). Today, with the emergence of English as the foremost international language of science and medicine[7] and the constantly generating offshoots and hybrid permutations of EMP courses[8] around the world varying:

[7] According to Wulff (2004, 188) we have entered the era of medical English, which resembles the era of medical Latin in that, once again, medical doctors have chosen a single language for worldwide communication. Today, the English language is also responsible for the anglicization of new and old medical terms (partly or wholly composed of English borrowings -*bypass, screening, scanning*-) which, in turn, are imported and naturalized in national medical languages by doctors from non-English-speaking countries (e.g. the French, who do not favour anglicisms, translate *screening* with *pontage*).

[8] For more extensive lists of EMP varieties and EMP courses involving non-Anglophone doctors, see Belcher (2009, 1-2) and Allwright and Allwright (1977, 58).

- in duration (i.e. intensive vs longer courses),
- in target audience (clinicians vs researchers; pre-medical students vs post graduate students in the clinical phase of their training),
- in medical specialty (e.g. oncologists, cardiologists, urologists, etc.),
- in medical skills and situations (e.g. doctor-patient consultation),
- in medical genres (e.g. English for report/journal article writing),

Maher's words still sound constructive being grounded on the three basic assumptions common to any teaching and learning for medical specific purposes: (a) EMP is designed to meet the specific English language needs of the medical learners (both novices and colleagues in academic, professional and vocational fields); (b) EMP focuses on themes and topics specific to the medical context; (c) EMP focuses on a restricted range of skills which may be required by the medical learner.

1.5.2 EMP: purposeful learning and teaching

EMP involves two major implications: (1) a new approach to English by learners whose needs become the foci of the didactic intervention; (2) a redirection by teachers of their pedagogical skills on circumstances of use which become of paramount importance in any EMP teaching and learning environment. Circumstances of use are not only a fundamental pre-requisite even to the selection of the particular linguistic forms or structures that ought to be taught (Schutz & Derwing, 1981, 31) but they

also entail a constant channeling into the specific EMP course content, the syllabus design and the choice of the teaching materials, thus to enable appropriate needs-responsive instruction while concurrently develop extensive expertise and know-how. Therefore, a typical EMP course mainly catered to non-native doctors and students wishing to develop their medical English proficiency, should necessarily prioritize these three key design principles: (a) the use of authentic texts and tasks; (b) a variety of activity mode and type; (c) rehearsal as closely as possible of the target assignments (e.g. writing abstracts or poster presentations) (Maclean cited in Ferguson, 2013).

1.5.3 EMP and medical language

By prioritizing circumstances of use, EMP courses manifestly rely on medical language and communication. However, since nurses, medical students, theatre operatives, hospital clinicians and conference-hopping doctors have pedagogically incompatible requirements, the generic category EMP needs further separation. Not, this time, at the EAP/EOP 'macro' level (academic or occupational) but –for the purpose of this study– at the wider oral/written 'discourse' level. Accordingly, within the large field of medical communication Ferguson (2013, 243-48), Skelton (2012, 1) and Fleischman (2001, 471-2) distinguish two partially overlapping categories. The first is the pedagogic language-related EMP

research aiming at improving the English language skills of non-Anglophone junior or senior health learners and focusing on written medical genre studies, studies of specific grammatical features, and vocabulary studies, with obvious interconnections between them. The second wider category is the literature on communication in health settings, especially nurse-patient interactions or doctor-patient encounters. The latter enjoys by far the lion's share in the output on medical language both for number of studies and publications in dedicated specialized journals (Candlin, 1976; Sarangi and Roberts, 1999; Ainsworth-Vaughn 2001; Ferguson 2001; Bosher 20013), however, its full review lies beyond the scope of this research whose prominent aim is to advance my understanding of the language problems encountered by non-native medical users dealing with the written language of medicine in English.

Chapter 2

Theoretical framework to the research

2.1 Rationale for the study

Conceived as an evidence-based thesis in Applied Linguistics focusing on the language of medicine, right from the start this research has been purposely meant to find a coherent integration between (a) language theories and language use -*i.e*. the 'Linguistics' and 'Applied' components of the discipline and of this study[1]-; (b) medical lexis, the specialised subcategory of EMP object of my investigation; (c) corpus linguistics, which, in its role as a catalyst for the qualitative and quantitative analysis of the selected data -than otherwise possible-, has yielded fresh insights into medical language with practical and useful implications for healthcare professional education and (d) pedagogy, *de facto*, the fourth complementary and completing factor of this 'specialized', 'specific' and unquestionably 'special' synergism.

[1] Illuminating for my understanding of an empirically based research in Applied Linguistics have been the ideas and contributions offered by these leading academics closely related to the discipline, respectively: Cristal D.(1980); Brumfit C. (1995); Hudson D. (1999); Schmitt N. (2002; 2010); Cook V. (2004); Davies A. (2007); Zoltán D. (2007); Hunston S. (2010) and Burns A. (2012).

Trying to strike an overall balance between the four parts to create a comprehensive framework has been a demanding experience requiring careful study, repeated analyses and critical reconsiderations of my goals. For convenience's sake the entire process has been divided into six phases precisely following the development of my research and to which I turn to.

2.2 Theoretical framework

2.2.1 Starting the research: evaluating needs

The prelude to my research has been a pilot study aiming at investigating the productive vocabulary knowledge of a group of one hundred EFL medical undergraduates involved in an English language test anonymously administered at the University of Palermo (see Appendix n.1, case-study).

Consistent with the EAP (English for Academic Purposes) target intended for this trial test, my language analysis was framed within four cornerstone definitions ensuring and regulating the success of the data accession and preparing the ground for the subsequent lexical evaluation: (1) 'vocabulary knowledge' involving 'the knowledge of the spoken and written form, its morphology, meaning, collocational and grammatical patterns, connotative and associational features and the knowledge of social or other constraints to be observed in the use of a word' (Richards,

1976; Nation, 1990; Laufer, 1997, Read, 2000); (2) 'productive vocabulary' implying 'the mastery level of word knowledge reflected in the learners' ability to successfully use the vocabulary' (Meara, 1997; Schmitt, 2010); (3) 'academic vocabulary' relating to that 'set of lexical items that are not core words but, unlike technical terms, are frequent in academic texts, regardless of the discipline […] sometimes used as a synonym for subtechnical vocabulary or discourse organizing vocabulary" (Paquot, 2010, 9) and (4) 'academic writing' referring to the ability of organizing writing 'to convey major ideas […] demonstrating command of standard written English including grammar, phrasing, sentence structure, spelling, punctuation and a range of vocabulary appropriate for the topic' (Hinkel, 2004, 18-19).

In line with the EAP (English for Academic Purposes) objectives mentioned above, I only examined the data from the two timed writing assignments included in the trial test, each related to contexts easily encountered in their experience in academia. Later, the data was collected in two learner corpora specifically compiled to measure how the 1st and 2nd year students' academic language proficiency[2] might affect the efficacy of their writings.

[2] Academic language refers to the written, oral and auditory language proficiency required to learn effectively in colleges and academic programs—i.e., it is the language used in classroom lessons, books, tests and assignments and it is the language that students are expected to learn and achieve fluency in. Frequently contrasted with general language, academic language includes a variety of formal-language skills —such as vocabulary, grammar, punctuation, syntax, discipline-specific terminology or rhetorical conventions— that allow students to acquire knowledge and academic skills while also successfully navigating school policies,

The preliminary results, despite pointing to inappropriate lexical choices resulting from a limited size and range of both general and academic vocabulary, have been of paramount importance for the future development of my research in that they served to anchor the 'what' and 'how' of my investigation.

The importance of carrying out a needs analysis for developing EAP tests has been emphasized by McDonough (1984); Carrol (1980, cited in Fulcher, 1999) and Fulcher (1999). Also, collecting data as language needs has a vital role in any language course, whether it be English for Specific Purposes (ESP) or general English course, and its centrality has been acknowledged by several scholars and authors (Munby, 1978; Hutchinson and Waters, 1987; Berwick, 1989; Brindley, 1989; Tarone and Yule, 1989; Robinson, 1991; Johns, 1991; West, 1994; Allison et al. 1994; Seedhouse, 1995; Jordan, 1997; Dudley-Evans and St. John, 1998; Iwai et al. 1999; Hamp-Lyons, 2001; Finney, 2002; Shongori, 2008; Kaewpet, 2009; Fatah-ELrahman Dafa-Allah .A.M., 2012). In my case, the rigorous and systematic processing of the information I had gathered from the trial language test served (1) first, to cast light on what non-native medical learners were required to do with the foreign language in their educational studies (2) second, to understand what the learners currently knew and still needed to know to successfully function in the

---

assignments, expectations, and cultural norms. Even though students may be highly intelligent and capable they may still struggle with academic language if they have not yet mastered certain terms and concepts, or learned how to express themselves and their ideas in expected ways (Cummins, 1979; 1984; Krashen & Lee, 2004; Krashen, 2007).

target language; (3) third, to make me pause on how non-native learners might best master and maximize the language during the time of learning in academia.

2.2.2 Tailoring the research: defining subjects

In my attempt to find a solution to some of the previously discussed challenges I was faced with new quandaries concerning, this time, the "who" and the "what" of my inquiry, namely the subjects and the specific language I was to investigate. Embedded in my mind was Hyland's (2009) definition of "needs" which:

> embraces both a consideration of the *present situation,* "starting where the students are", looking at what they can do now and what they want to do, and of the *target situation*, considering their future roles and the linguistic skills and knowledge they need to perform competently in those roles (204).

Therefore, in my plan, the diagnostic needs analysis conducted on the sample of non-native medical novices struggling with their academic writing proficiency was supposed to be just the nucleus of a larger frame also including healthcare postgraduates and practitioners. Following Knight, Lomperis, van Naerssen & Westerfield (2010, 7) I wanted to

bring together both (*a*) medical language learners who were in the process of developing expertise in their fields and needed English communication skills as tools in their academic training and (*b*) language learners who were already experts in the medical fields and needed English communication skills as tools to succeed in their career.

Ideally, my orientation was to move the focus of the research from EAP to EMP, on the basis of two epistemological assumptions. The first, more general, led to the indications of the Council of Europe (2001) setting high competences in specialized languages as a priority for academics and professionals across the member states of the Union and beyond[3]; the second, more specific, presumed that the remedial measures I meant to devise to address the freshmen's identifiable vocabulary deficiencies would also benefit their non-native senior fellows to whom achieving native-like writing proficiency is vital to get findings published[4]. Ultimately, medical research is all about sharing findings

[3] The guidelines set forth by the Council of Europe through the *Common European Framework of Reference* for languages (CEFR) define "what knowledge and skills learners have to develop so as to be able to act effectively" (CEFR, 2001, 1) and also emphasize that educators need to equip European learners with those life-long skills needed to handle the communicative tasks in "the *personal, public, occupational* and *educational* domains" (CEFR, 2001, 54). The CEFR is also intended to meet the real needs of the learners in order to overcome the barriers to communication among professionals from different educational systems in Europe (*Idem*, 1).

[4] "Publish or perish" has now become a cliché in medical circles, not only because researchers need to gain recognition or get extra edge in the professional sphere, but mostly because both legislation and industry sponsors' policies require reporting of clinical trial results and publication of even early phase studies on the basis that sharing results with others enhances research itself (Leighton C., Leslie C., Juli C., Frank S. D., Robert E., Michelle E., et al., 2010, 1967).

with others and the way to reach the widest audience is to publish findings in internationally reputable journals. Needless to say that this constitutes a disadvantage[5] for non-Anglophone researchers[6] who, compared to their Anglophone counterparts, are more likely to have their works rejected on the basis of poor writing and language quality criteria[7] (Benfield and Howard, in Swales, 2004, 46-47; Benfield and Feak, 2006, 1728-29, Ferguson, Perez-Llantada, Plo, 2011).

2.2.3 Limiting the research: selecting language

After defining the participants of the research and the domains within which they should operate with their written communicative acts, my next priority was to identify which language difficulties are commonly

[5] Swales speaks in terms of "*the Anglophone grip*"(1990, 97) and "*Tyrannosaurus Rex*" (1997, 347) referring to the harmful phenomenon of the dominance of English in published research which has led native speakers of English to enjoy a preferred status and the "loss of specialised registers in other healthy languages" (*Idem*, 376).

[6] John R. Benfield (2000, 648; 2006, 1730) Austrian Professor of surgery and ex-editor of considerable experience, laments the "added burden" English language imposes upon non-native doctors coping with the world of publications in that high ranking journals accept only a very small proportion of their submissions "as is". Therefore, he urges the Anglophone more privileged colleagues to take more responsibility for assisting their non-native peers, possibly through co-publication.

[7] *The British Medical Journal* currently boasts a rejection rate around 93% with fewer than half of the received articles sent for external peer review. From the homepage, *The BNJ* editorial staff bluntly appreciates "that authors do not want to waste time by sending their research articles" without considering "the journal's editorial requirements, submission processes, publication ethics, peer review, and effective communication, much of which has traditionally been seen as mysterious to authors". (*The BNJ*, http://www.bmj.com/about-bmj/resources-authors/article-submission/article-requirements).

experienced by non-native academics -both junior and senior learners- while writing in English, and, possibly, to mitigate these challenges.

Once again, the starting point for my investigation was the information gathered from the pilot study conducted with the sample group of medical undergraduates, whose language *lacks* and *necessities*[8] mostly lied in inappropriate lexical choices and stumbling collocation uses[9]. The right use of words and set-phrases play a great role in increasing the lexical density of a text; transmitting the intended meaning and economising on the use of words. These are crucial requirements in high-proficiency writing, either academic or specialized (Llach, 2011, 49-51). Correspondingly, when the amount of these linguistic items decreases, the quality and the accuracy of the writings also diminish. Studies on vocabulary size (Laufer, 1996, 2001; Nation 1993; Meara 1996; Morris & Cobb, 2004; Qian, 1999, 2002) observed how learners with bigger vocabularies are more proficient in a wide range of language skills. Written production also benefits from large vocabulary breadth (Laufer & Nation, 1995; Lee, 2003; Meara *et all,*

[8] J. D. Brown (2009) identifies as *discrepancy needs* any differences between learners' expected language performances and what they can really do. Inside this framework he further spots three other types of needs, defined as (1) *necessities:* what learners need to know to successfully function in the target L2; (2) *lacks*: differences between target L2 proficiency and what learners currently know; (3) *wants:* what and how the learners would like to learn (271-72).

[9] These deficiencies were also confirmed by the post-test questionnaire collecting the candidates' feedbacks: 20% of the students rated the writing tasks as "difficult", with a peak of 8% who rated them as "very difficult". Among the reported difficulties: 18% lamented "problems with EAP and EMP word choice" or felt handicapped by "a less rich vocabulary" and "less facility in expression" (see Appendix 1, case study, Feedback questionnaire).

2000; Meara & Bell, 2001; Morris & Cobb, 2004; Nation, 2001), because "a rich and varied vocabulary and an adequate knowledge of words are a prerequisite for effective language use" (Read, 2000, 83). Nation and Waring (1997, 2) calculated that a native university graduate has a vocabulary of around 20,000 word families. This figure drastically diminishes to 5,000 word families for ordinary adult learners of English, with, instead, significant native-like rates for educated second language learners who have studied English for several years, although they are not the norm.

The good news for my research was (a) to discover that not all words are equally useful in that only a small number occur very frequently, and (b) most of these are content words (Meara and Jones, 1990; Milton and Meara, 1995; Nation & Webb, 2011; Schmitt, 2000). I consequently decided to frame my investigation around these enlightening assumptions, in the belief that knowing a very large proportion of the most frequent content words of educational and scientific texts would correspondingly allow a good degree of written proficiency.

2.2.4 Narrowing the research: identifying gaps in previous studies

Providing EMP learners with the language necessary for their studies or professions to a large extent means giving them the specialized words they need, which, specifically for EMP, still remains an open question,

particularly as regards the nature of specialized vocabulary itself: ‘what is meant by ‘specialized language’?. The existing literature has not been very supportive in providing adequately satisfactory answers to this question. Not in terms of research or studies undertaken in the field —which are instead numerous— but rather for the controversy arisen over a shared definition of specialized vocabulary, *i.e.* those words that should deserve more attention than others in language learning for special purposes. Excluding Nation’s (2001, 187) vocabulary division into four levels *–high frequency words or basic vocabulary, sub-technical vocabulary, technical vocabulary and low frequency–* which offers valuable insights into the acquisition of vocabulary in different stages of advanced learning, the issue around a common classification of specialized terms remains a debated topic. Researchers have generated word lists comprising the most important words for specific fields; accordingly, for learners with academic goals, the 570 word family *Academic Word List* (Coxhead 2000) is like a specialized extension of the 2,000 high frequency words selected by West (1953) in his *General Service List*. Coxhead’s vocabulary has been variably called ‘academic vocabulary’ (Martin 1976), ‘sub-technical vocabulary’ (Cowan 1974), ‘semi-technical vocabulary’ (Farrell, 1990), ‘specialised non-technical lexis’ (Coehn, Glasman, Rosenbaum-Cohen, Ferrara and Fine, 1988), “common words that occur with special meanings in specific and technical fields” (Trimble 1985), ‘laytechnical’ or ‘cryptotechnical’ (Fraser, 2006, 68). However, despite the high occurrence across twenty

eight subject areas and academic fields, Coxhead's inventory is neither truly technical nor truly academic, in that it is not typically associated with just one field (Chung and Nation, 2003) nor can it evenly cover the vocabulary of academic discourse irrespective of the specific field of study, since all disciplines shape words for their own uses (Hyland and Tse 2007).

Uncertainty persists even with the definition of technical vocabulary. Essentially, technical words are recognizably specific to a particular field or discipline; frequently occur in a specialized text making up about 5% of the text's running words and are low frequency in other fields (Nation 2001, 198); they are context-bound and topic-dependent (Salager 1985, 6). Nevertheless, the lack of information about how technical vocabulary relates to other types of vocabulary (Chung and Nation, 2003) and the necessity of systematic approaches able to determine which words are technical enough to be categorized under this heading (Lowe 2010, 1-4) make this category elusive and difficult to classify.

2.2.5 Focusing the research: establishing goals

Right from the onset of the project, my major goal has been to provide both medical students and professionals with the English vocabulary they really need in their target contexts. Given the notoriously complex nature of the medical lexis and -as seen above- the impossibility, as a

specialized language, of systematizing all its terms unequivocally, the focus of my research was therefore directed to the *core lexis* of medicine *i.e*., those lexical items which are frequently and homogeneously distributed across the medical spectrum, whatever the medical specialty. Not necessarily, though, technical terms which, by virtue of their semantic univocity (Maglie, 2009, 24) are automatically learnt studying the discipline, but *keywords,* which, carrying lexical meaning allow experts to communicate more rapidly.

Corpus linguistics would help me to 'diagnose' those *keywords* that really matter in the healthcare register, because besides providing a variety of domain-specific materials for language teaching and learning (McEnery and Wilson, 1996, 119-120), corpus linguistics also delivers statistical information (quantitative analysis) and allows direct observations of how vocabulary is used in context (qualitative analysis) (McCarten 2007, 3).

Defined the diagnosis, my second aim was to find a 'treatment' to the users' lexical 'disorders'. I meant to provide non native speakers with a writing aid that included the necessary information on how to properly use the medical *keywords* in order to produce well-constructed scientific texts. Hence the idea of a web-based lexical monolingual glossary, the *MedEnCor-Lex* (www.medencor.com), showing medical *keywords* in context, listed in semantic wordlists and screened with all their possible collocational and phraseological combinations.

2.2.6 Directing the research: framing research questions

Confronted with these issues, two were the research questions I posed which also served to guide and direct my study:

(1) exactly, what kind of words make up the medical lexis that medical undergraduates and practitioners need?;
(2) once identified, how should such vocabulary be learned and taught?

2.3 Significance of the study

By isolating *keywords* and uncovering patterns of real language use my purpose was fourfold: I intended (1) to 'heal' the language 'impairments' presented by non-native medical trainees and doctors; (2) to ease the effective use of *keywords* in the specialized register; (3) to trigger the rapid growth of the learners' disciplinary vocabulary and (4) to help healthcare users conform to the written conventions of the scientific discourse.

My ultimate goal is to transform students from language learners to language users. As such the *MedEnCor* database also aims to raise the students' awareness towards the importance of medical technical documents written in English contemporarily making them proficient in using resources readily available online. Namely, to help them learn

English and use the language for professional purposes. The latter aim is not limited to making students autonomous in their learning. Rather, it aims to foster attitude towards language use, especially after graduation when there are fewer opportunities to receive language training at the workplace. In point of fact the *MedEnCor* is not only an instrument likely to satisfy the diverse language needs of its users either for distance education or self-learning but also an e-learning system intended to maintain, improve and broaden the medical linguistic knowledge and skills as well as to develop a positive orientation towards continuing specialized development. All essential qualities and skills required for its users' future career and professional lives as doctors and researchers.

Chapter 3

Literature Review

3.1 Language–related research in EMP

There is a considerable body of EMP research related to written language, most focusing on intra–professional and interdisciplinary team[1] communication and genres, though, as Roberts and Sarangi (2003) noted, much of this research has been conducted by health care 'outsiders' rather than 'insiders', thus indicating a need for more dialogue between 'research and researched' (339).

Since most of the existing EMP language research entirely relies on corpora of various sizes, in this review I have conveniently organized the

[1] The NHS (*National Health Service*, launched in 1948 by the UK Department of Health and grown to become the largest publicly funded healthcare system in the world) defines 'intra-professional and interdisciplinary team' two key-words in medical communication referring to a group of experts bound by a common purpose who meet regularly to share, collaborate and consolidate knowledge from which plans are made, actions determined and future decisions influenced. Specifically, the NHS further subdivides the team-collaboration as follows: (a) *interdisciplinary team:* a group of health care professionals from diverse fields who work in a coordinated fashion toward a common goal for the patient; (b) *intraprofessional team*: a team of professionals who are all from the same profession and collaborate on the same case; (c) *multidisciplinary team*: a team of professionals including representatives of different disciplines who coordinate the contributions of each profession in order to improve patient care; (d) *transdisciplinary team*: a team composed of members of a number of different professions cooperating across disciplines to improve patient care through practice or research. (The NHS, available at http://medical-dictionary.thefreedictionary.com/intraprofessional+team)

related studies under the umbrella term of EMP corpus–based studies, then I further divided and synthesized this generic group into five more comprehensive sub-categories: EMP word lists; EMP genre studies; EMP grammatical studies; EMP vocabulary studies and EMP vocabulary studies across disciplines and languages ; with obvious links between all levels

3.2 EMP corpus based studies

For some time linguists have recognized the value of compiling large corpora of language and subjecting these to computerized analysis to discern patterns of language use across a broad range of human social practice. Although corpus linguistic research methods are a relatively new application in relation to medical texts; the benefits of corpus work have already been widely accepted and documented by medical researchers and professionals and, as a result, a growing number of publicly available (bio)medical corpora and data sets have come to light during the last years. The existing healthcare corpora differ considerably in size, quality, coverage, encoding and depth of linguistic and structural characteristics. The vast majority are monolingual English corpora, and cover different medical domains, time spans, registers and genres. One of the principal reasons for this expansion is that corpus linguistic research can be flexibly applied to healthcare data and can be used to address a whole variety of questions, topics and ideas.

In this chapter, for the sake of clarity and convenience, I will take into consideration only two diachronically divergent healthcare corpus-based studies, purposefully representative of the earliest and the latest –though less specialized– written medical communication, leaving more detailed information on the subject to the following subsections of this review.

Pride of place in this list of medical corpus-based studies is given to the *Corpus of Early English Medical Writing* (CEEM, 1375–1800), consisting of three diachronically divided sub-corpora: *Middle English Medical Texts 1375–1500* (MEMT), *Early Modern Medical Texts 1500–1700* (EMEMT) and *Late Modern English Medical Texts 1700–1800* (LMEMT). It is a register-specific corpus of English vernacular medical writing compiled at the University of Helsinki (2005) covering the entire history of medical writing in English from the earliest manuscripts to the beginning of modern clinical medicine. The texts (about two million running words) contain a representative sampling of medical prose divided into four categories: surgical texts, specialized texts, remedies and *materia medica*, aimed at different target audiences[2]. Combining both quantitative and qualitative analysis, the corpus has allowed the identification of new contagious diseases (*plague, smallpox, dysentery, consumption, typhus, syphilis, scurvy*); new cures and reliefs (*tobacco,*

[2] The texts in the corpus range from highest learning (originally circulating among the most highly-educated medical professionals) to practical health guides and other instructions written for the general public. The corpus compilation required the interdisciplinary collaboration of linguists, manuscript scholars and medical historians in order to ensure representativeness and philological accuracy of corpus data and scientific writing (CEEM, *Corpus of Early English Medical Writing*. Available at http://www.helsinki.fi/varieng/CoRD/corpora/CEEM/Emod context. html).

coca and sassafras, quinine, myrrh, petroselinum, rhubarb, laudanum) and new theories (*Galenic medicine, alchemical Paracelsianism, experimental science*).

Equally useful, though diachronically opposed is the *Teenage Health Freak Corpus* (THF) comprising emails sent via the 'Ask Doctor Ann' facility on the Teenage Health Freak website (http://www.teenage healthfreak.org) over a more recent time frame: 2004–2009. The THF corpus (about one million words) was released by the University of Nottingham (2007) and it explores the adolescents' use of language in relation to their health. The analysis of the vocabulary occurrences and the comparison between the THF corpus with the CANCODE (*Cambridge & Nottingham Corpus of Discourse in English*) revealed the predominant frequency of key-terms related to sexual and reproductive health (*penis, pregnant, period, gay*) with a particular concern for *normalcy.* A finer–grained analysis of concordance lines showed that *normal* did not simply mean statistically average but also effectively *desiderable* (Corbett & Lu, 2010, 66–68). Despite the apparent irrelevancy of this data to the scientific language of written EMP research under investigation, the THF findings are of special interest to medical professionals commonly unaccustomed to considering adolescents' lay beliefs about sexual health (Harvey, 2013, 197–98).

3.2.1 EMP word lists

Coxhead's (2000) interdisciplinary *Academic Word List* (AWL) claiming to cover about 10% of the total words of general academic texts (from four different areas: law, art, science and business) has long been established as the lexical academic threshold to be reached in ESP and EAP courses. Its general lexical coverage, however, has been recently questioned by Hyland and Tse (2007) who argued that each academic subject has its own register which varies across disciplines.

Investigating a corpus of 50 medical RAs Chen and Ge (2007) confirmed the high lexical coverage of the AWL in medical research articles but, since such coverage was far from complete in representing the academic words frequently occurring in medical texts, they proposed the compilation of a new medical academic word list. Encouraged by the findings of Chen and Ge's research, Wang, Liang and Ge (2008) established the *Medical Academic Word List* (MAWL) of 623 word families frequently used across 32 subfields of medicine. These word families were selected from a one–million–word corpus of medical research articles on the basis of predefined criteria that included: specialized occurrence, range and frequency.

Following Wang and his colleagues, Mungra and Canziani (2013) compiled a corpus of 200 medical case histories, which after being lexically profiled by Nation's *RANGE* freeware, produced the *Medical Academic Word List for Clinical Cases* (MWLCC), a list of 241 families to be considered as typical of clinical cases.

Similarly, Hsu (2013) created a more restricted *Medical Word List* (MLW) bridging the gap between non–technical and technical vocabulary. Compiling a corpus of 155 medical textbooks across 31 medical subject areas (totaling about 15 million running words) and examining the frequency and range of words outside the most frequent 3,000–word families from the *British National Corpus* (BNC), the researcher formed the MWL which accounted for 10.72% of the tokens in the medical textbooks under study.

Breaking down the divisions between general, academic and technical vocabulary, Fraser (2009) created a *Pharmacology Word List* (PWL), a single discipline-based word list containing the 601 most frequently used words families in the *Pharmacology RA Corpus* (2,570 word families). This list, regardless of the overlapping with the GSL2000 and the AWL, was mostly conceived as a way of implementing the words that learners needed to know.

More recently, Ng Yu Jin (2013) analyzing a nursing corpus made up of essential core textbooks, highlighted the 2,000 most frequently used nursing words, useful to reduce students' reading and writing deficiencies and guide educators to the teaching of nursing vocabulary.

3.2.2 EMP genre studies

Written medical genres include research articles, abstracts, case reports, review articles, peer reviews, letters to the editor, book reviews

and letters of referral. Pride of place among these, however, must go to the research article, the 'gargantuan genre' given its 'central communicative mechanism' in scientific discourse settings (Swales, 1990, 95).

Drawing inspiration from Swale's seminal work on genre, Nwogu (1997) examined the constituent elements of contemporary research articles (RAs) in a corpus of 30 medical texts from five leading journals (*Lancet, British Medical Journal, New England Journal of Medicine, Journal of Clinical Investigation, and Journal of the American Medical Association*) and observed the recurrence of the traditional IMRD structure (Introduction, Methods, Results, Discussion). Using the Swalesian move–and–step schema, Nwogu, first described the specific functions and lexico–grammatical characteristics contained in each RA move and, then he also provided a schematic eleven–move–map of the RA intended to assist the novice RA authors. Skelton (1994), on the other hand, had identified fifteen moves in his analysis of 50 medical papers from the *British Journal of General Practice*, emphasizing the general 'optionality' rather than 'obligatoriness' of their usage.

Since genres are not static constructs, it is no surprise to see that Nwogu's model was updated by Li and Ge (2009) who made moves 1 (presenting background information) and 6 (describing data analysis procedures) obligatory at the expense of move 9 (highlighting overall research outcome), now considered an optional move.

Most recently, also Fryer has analyzed medical research articles using a systemic functional and structural move analysis approach (2012). In

the analysis of 16 experimental research articles published between 2004 and 2006 in five medical journals, Fryer used a modified version of Swales' structural move analysis thus creating a novel and combined methodology which described medical RAs in terms of their function and lexicogrammar. The potential pedagogical and methodological applications of the study are discussed in relation to previous research.

Next to importance to RAs is the medical review article, a synoptic genre usually composed by an authoritative figure who synthesizes findings from a variety of sources to present a more comprehensive picture of a particular disease or treatment. Measuring a corpus of 158 review articles against 10 established quality criteria, McAlister et al (1999) reported that only two met the highest methodological standards.

Also case reports, with their function of recording the pathology of a single patient's disease (from diagnosis through treatment and outcome) are crucial in written medical communication. In order to study the evolutionary path of register in case reports, Atkinson (1992) diachronically analyzed the rhetorical and linguistic changes of this genre in the *Edinburgh Medical Journal* between 1735 and 1985, observing a gradual shift from a narrative mode toward the more conventionalized IMRD structure typical of RAs.

Taavitsainen and Pahta (2000), drawing on a corpus of case reports from the *British Medical Journal* and *The Lancet* covering the period 1850–1995, observed substantial changes in the genre which they attributed both to the new trends of medical research being oriented to

large volumes of clinical data, and to the increasingly rapid growth of the medical discourse community.

Closely related to the above mentioned medical genres is the abstract whose importance has considerably increased in the medical literature with the emergence of online healthcare databases providing free access to abstracts but not to related articles. In such circumstances, explicit information and clear organization have become a priority in abstract writing[3]. There is a substantial body of ESP research on abstracts, but few of these publications have a specific focus on medicine (Ferguson, 2013, 250). Exceptions, dated back to the 1990s with two papers by Salager–Meyer (1990, 1992): in the first, after analyzing abstracts from three genres (RAs, case reports, review articles) she concluded that only half (about 52%) of the samples conformed to journals' guidelines; the second paper, instead, contained a detailed analysis of the distribution of verb tenses and modals across the main moves of the genre.

In the written medical scenario there are also studies related to less prominent genres, which I just mention. These include letters to the editor (Magnet and Carnet, 2006); consensus statements (Mungra, 2007); Book reviews (Salager–Meyer, Ariza and Pabòn, 2007); journal editorials (Giannoni, 2008) and peer reviews (Mungra and Webber, 2010).

[3] *The Lancet*, in its homepage tutorial for abstract writing titled 'How to get published? What distinguishes a good manuscript from a bad one', explicitly requires these headings: *background, methods, results, clinical implications.* (Available at http://www. elsevier.com/data/assets/pdf_file/0011/239294/Get-Published-Quick-Guide.pdf)

3.2.3 EMP grammatical studies

Many EMP studies have focused exclusively on the grammatical features of mainstream medical genres. Examples go back to Adams Smith (1984) who studied a variety of verbal and non verbal modals across a selection of editorials and clinical cases from the *British Medical Journal*.

Salager–Meyer et al (1989) examined 17 grammatical variables (verb tenses, voice, and forms) within the communicative functions of 51 medical English scholarly papers across three genres: case reports, editorials and research papers. Three axes of variable distribution were revealed that helped categorize and distinguish the analyzed text types as such. Pedagogical guidelines for preparation and use of teaching materials were provided.

Salager–Meyer (1994) also identified the distribution of modulation devices (hedges) across the different rhetorical sections of research papers (RP) and case reports (CR) in a corpus of 15 articles drawn from five leading medical journals. She concluded that there was tendency for specific verbs to be used for hedging in the 'discussion' and 'comment' sections of these genres. The author also pointed out the importance of her findings for EAP and EMP teachings.

A similar contextual analysis was carried out by Hyland (1999) with a corpus of 26 research articles, whose processing confirmed the value of

hedges[4] in scientific research writing as a resource for academics to present claims with caution and anticipating peers' possible rejection of their propositions. Kindred study was undertaken by Vartalla (1999) with a corpus of 15 texts from the *New English Journal of Medicine* (NEJM) whose results, according to the author, deserved careful consideration in EMP books and teachings where hedges were largely unaccounted if not totally neglected.

Biber and Finegan (1994) investigated the linguistic variations in the IMRD moves of a corpus of 20 medical research articles and noted several linguistic peculiarities, *e.g*. predominance of present tenses in 'introductions' and 'discussions'; past tense use in 'results' and 'methods' where agent–less passives were also dominant.

Another grammatical feature that has been widely discussed in the literature on scientific discourse is the use of 'if–conditional'. Ferguson (2001) compared two written genres (RAs and journal editorials) and one spoken (doctor–patient consultations) and enumerated a total of 177 such conditionals with differences in use between the two media: more operative in the former and more polite in the latter.

---

[4]Ken Hyland defines hedges as

> 'expression of tentativeness, indirectness and possibility central to medical writing where they play a critical role in gaining ratification for claims from a powerful peer group by allowing writers to present statements with appropriate accuracy, caution, and humility, expressing possibility rather than certainty and prudence rather than overconfidence. In a context where the accreditation of knowledge depends on the consensus of the research community and the need to evaluate evidence, to comment on its reliability, and to avoid potentially hostile responses, expressions such as *might, perhaps,* and *possible* can contribute to gaining the acceptance of research claims' (1999, 33).

Carter-Thomas and Rowley-Jolivet, E. (2008), analyzing the usages of 'if–conditionals' in a corpus of research articles, conference presentations, and editorials, proved that the use of these syntactic patterns in medical discourse contrasted with the theory on conditionals that learners were likely to meet, thus advocating a more genre sensitive approach to the teaching of syntax in the EMP classrooms.

3.2.4 EMP vocabulary studies

Medicine is well known for its open corpus of technical terms: mostly borrowed from Greek and Latin, some are English loans, others are anglicized hybrid forms with little morphological adaptation or newly coined ones. Given the ever–expanding number of medical terms and the salience of technical words in medical texts[5] it is no surprise that the pedagogical treatment of this category of vocabulary is largely debated in EMP circles, with opposing theories and views, largely depending on what perspective the issue is being tackled: EMP learner's first language (L1), level of EMP background or which form of word knowledge is at stake. Whatever the angle of vision, medical lexis has received increased

[5] Chung and Nation (2003) estimated, for example, that technical words, defined as 'ones with a narrow range of occurrence and largely unknown in general use' accounted for as much as 37,6 percent of all word types in an anatomy text as against 16.3 percent of types in an applied linguistics text (105-108).

research attention recently, with the common objective to consistently inform and assist vocabulary teaching and learning.

According to Nattinger (1980), the study of specialized vocabulary cannot be separated from its 'compositional nature' which consists in 'stitching together preassembled lexical clusters and phrases into discourse', since learning such collocations and phrases 'leads to fluency in writing and enables students not to violate incongruities of register' (76–77). An early attempt in this direction was made by Gledhill (2000) and colleagues who compiled the *Pharmaceutical Sciences Corpus* (PSC) with 150 cancer research articles in order to examine the high frequency formulaic patterns and phraseological units in pharmaceutical RAs 'introductions'. Similarly, processing the use of collocational frameworks in a corpus of medical research papers compiled at the University of Zaragoza, Luzòn Marco (2000) demonstrated the usefulness of corpus–based analysis to discover some restricted sets of lexical items favoured by a specific genre. In line with the new trends in corpus and phraseological studies there is also *SciE–Lex*, a lexical database that provides morphological, semantic, syntactic and collocational information of specialized (bio)medical terms used in scientific research articles. By presenting the phraseological conventions of the genre, Laso and Salazar (2013) intended to help Spanish scientists, especially those in the medical community, to write native-like scientific articles in English.

3.2.5 EMP vocabulary studies across disciplines and languages

EMP vocabulary studies have not only highlighted the lexical components of target discourse but have also drawn attention to comparative analyses of medical genres with those of other disciplines or languages. For example, the KIAP (*Cultural Identity in Academic Prose*) project, a corpus of 450 research articles covering three disciplines (Economics, Linguistics and Medicine) and three languages (English, French and Norwegian) located at the University of Bergen has been investigated by Fløttum and colleagues (2013) in order to establish whether cultural or language identities may be identified in academic prose, and, if so, whether these identities are language or discipline–specific in nature. On this issue, also the CADIS (*Corpus of Academic English*) compiled by Gotti and colleagues (2007) at the CERLIS research centre (University of Bergamo), is an aid in the identification of textual variants arising from the use of English as a first language, second language, or lingua franca of the scientific community. Giannoni for example, one of the CERLIS members (in Gotti, 2011, 40–42), investigating the CADIS corpus (12million tokens, comprising four different text types –RAs, abstracts, book reviews and editorials– from four different disciplinary areas –Medicine, Law, Economics and Applied Linguistics– in two alternative languages Italian and English–) noticed that hyperbolic, ironic or emotive language was almost inexistent in hard disciplines (Medicine) while it was very frequent in soft disciplines (Linguistics and Economics).

Another aspect investigated by the CERLIS members is the relationship between writing practice and linguistic background. Indeed, a scholar's direct participation in a predominantly English–language cultural environment is considered a privileged condition for academic acceptance and advancement. In a cross–cultural analysis of medical RAs written in English by both native speakers (Ns) and Italian non–native speakers (NNs), Maci (in Gotti, *ibidem*, 43) demonstrated that both versions conformed to the codified IMRAD patterns; differences, instead, emerged in the argumentative strategies, where the Italian authors prioritized a more direct approach rarely adopting hedging devices.

Years before, also Salager–Meyer (Alcaraz Ariza & Zambrano, 2003) had already examined cross–cultural differences and historical changes in the medical discourse of English, French and Spanish medical texts. Using a corpus of historical texts (1930–1995), the authors found more 'passion' in the early Mediterranean writings compared to those of the Anglo–Saxon peers. However, they noted that during the last decade of the twentieth century the more neutral hedging style had prevailed over the 'national' divergences, owing to a more competitive professional market which had obvious implications on scientific discourse.

3.3 Limitations of existing corpus-based studies

The review of the existing corpus-based research related to EMP language has identified a number of key studies in the medical field which have complemented and substantially modified the current approach to the written scientific communication in English. Despite the remarkable added value and the potential pedagogical and methodological implications of each contribution, rarely, have such studies been explicitly combined with detailed and practical explanations for higher education writing; and -with just a few exceptions- have careful informative and didactic instructions been made publicly available by the authors on how to apply, implement and put the findings into practice in learning contexts.

To address some of the limitations mentioned above, to add to existing studies, and to complement my own work with the current medical research on EMP vocabulary and teaching, I have worked towards the compilation of the *MedEnCor*, an open corpus of medical texts and its parallel *MedEnCor-Lex* database, a web-based glossary showing medical terms in context, which I will fully describe in Chapter 4 and Chapter 6 of this thesis.

Chapter 4
Methodology

4.1 Methodology of the research

As demonstrated in the literature review in chapter 3, interest in the use of language corpora and computer tools for language analysis has grown tremendously in the past decades. Understandably, linguists have emphasized the usefulness of corpora in language research because they allow the study of language variety based on authentic texts; display language patterns in context; deliver statistical information and frequency data on smaller or larger scale; provide empirical evidence about language use and, implicitly, make users more active and independent analyzers of lexical items (Johns, 1986, Taylor, 1991; Leech, 1992; Stevens, 1995; Cobb, 1997; Biber *et al*, 1998). Put it in simple terms, language corpora and computer tools are best described as favoring "the study of language based on examples of *real life* language use" (McEnery and Wilson 1996, 1).

As the present study is also aimed at investigating real lexis with an empirical stance, the approach and methods offered by language corpora were found appropriate, notably for the *Medical English Corpus* (MedEnCor), the specialized corpus containing documents about health

and biomedical topics, specifically designed and compiled to analyze the scientific discourse and vocabulary, *i.e* the *core lexis* of medicine, focus of my research.

In this chapter firstly I present some theoretical guidelines for the design and compilation of a specialized corpus in compliance with Corpus Linguistics standards; secondly I provide a comprehensive description of the stages followed in the creation of the *Medical English Corpus* and its characteristics; and finally, I report the preliminary results obtained from the basic analysis of the corpus: frequency lists, statistical information, lexical profiling of keywords and their semantic categorization.

4.2 Specialized corpora

Bowker and Pearson (2002) define a specialized or special purpose corpus as "a collection of texts that focus on a particular aspect of a language […] a particular subject field, text type, or language variety used by members of a certain demographic group" (12). Because of its specialized nature, a special purpose corpus cannot be used to make observations about language in general, its focus being "patterns of *langue*, the shared area of meaning-creation in a speech community" (Tognini-Bonelli, 2010, 22). As such a specialized corpus must be

compiled with a collection of texts relevant to a target domain or with specific genres appropriate for the task.

In assessing the specialization of the documents fit for corpus inclusion Kὺbler and Aston (2010, 507-8) suggest some useful criteria to take into account: (1) the extent to which certain texts 'parallel' those of the target source; (2) the extent to which they can be considered authoritative texts (it would be unwise to treat texts written by non-experts or by non-native authors as reliable sources of vocabulary or terminology) and, (3) the intended readers' presumed expertise in the field which entails the categorization of the corpus in divulgative, academic, didactic, explanatory and authoritative texts.

Counter arguments are presented by Tuber and Cermàkovà *(*2007, 67-69) who assert that there is not a standard recipe for the composition of special corpora, all what the compilers have to do is to draw up a set of hypotheses that will guide a carefully constructed composition deliberately representing what the discourse community agrees upon.

Bearing all these definitions in mind and assuming that "specialized corpora do not grow on tree" but they "have to be compiled appropriately for the task" (Kὺbler and Aston, 2010, 507), my first consideration was to check whether there were any operating corpora already available for my research purpose. There existed plenty of specialized corpora related to English for Medical Purposes and academic medicine but as already stated in chapter three they did not suit my needs in several respects. Some of them collected samples from

one or few specialized domains; others included only one or few genres or covered multiple languages. Medical English *keywords* across several domains and text-types were underrepresented and so were the pedagogical guidelines to revise the related EMP lexical repertoire. Admittedly, this constituted the embryonic planning stage prior to the *MedEnCor* actual design and compilation.

4.3 Corpus design and compilation: the essentials

In describing the complexity of compiling a corpus, Leech (1998, xvii) remarked that "a great deal of spadework has to be done before the research results [of a corpus analysis] can be harvested" Creating a corpus, he commented, "always takes twice as much time, and sometimes ten times as much effort" because of all the work that is involved in designing a corpus, selecting genres, collecting texts and computerize data. This has much to do with the meaning of 'corpus' which in modern linguistics has moved away from the original Latin definition of 'body of writings addressing a certain topic"[1] taking on more specialized meanings as (a) "a finite-sized body of machine-readable texts, sampled in order to be maximally representative of the language variety under consideration" (McEnery and Wilson, 1996, 32); (b) "a collection of pieces of language texts in electronic form, selected

[1] Available at http://latindictionary.wikidot.com/noun:corpus

according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (Sinclair, 2004); (c) "a large and principled collection of natural texts […] especially designed to address specific research questions" (Biber *et al*, 1998, 12) and (d) "a corpus must represent something and its merits will often be judged on how representative it is" (O'keefe, A., McCarthey, M., and Carter, R., 2007).

Considering these definitions of corpus, there are some important common features to highlight: 'representativeness'; 'size'; 'authenticity', 'variety'; 'balance', 'sampling' and 'chronology'. These characteristics are what make corpora different from other types of text collections and I will examine each of them in turn.

4.3.1 Representativeness

Representativeness is an essential feature of a corpus and distinguishes it from an archive or a random collection of texts. A corpus is designed to represent a particular language or language variety (*i.e. population)*, yet the task of collecting authentic and principled texts typifying the language intended to capture is enormous since it is virtually impossible to sample every extant utterance or sentence of a given language (Reppen, 2010, 3). The aim of representing the general usage of a language variety through a set of linguistic samples is a

controversial issue. The discussion stems from the complexity found in defining representativiness itself and achieving that some fractions of a language in a corpus can be considered as characteristic of the whole. Biber (1993) defines representativiness as

> the extent to which a sample includes the full range of linguistic variability in a population; *i.e*. different linguistic features are differently distributed (within texts, across texts, across text types), and a representative corpus must enable analysis of these various distributions […] If a corpus does not represent the range of text types in a population, it will not represent the range of linguistic distributions (243).

Confirming that representativeness heavily depends on sampling from a broad range of genres, Sinclair (2004) also suggests important steps towards achieving as representative a corpus as possible: (1) draw up a comprehensive inventory of text types using external criteria only; (2) put the text types in a priority order taking into account all the factors that you think might increase or decrease the importance of a text type; (3) estimate the number of text types, a target size for each text type and the practicality of gathering quantities of it (*e.g*. copyright issues),

4.3.2 Size

In addition to being a principled collection of naturally occurring texts, another defining characteristic of a corpus is that it is a *large* collection of texts. However, *large* is a relative term, especially in view of the increasing growth of corpus size favoured by the current advancement of technology. How big is *large*? In the 1960s, when some of the first electronic corpora were built, one million words were considered large for a general corpus. Now, just over 50 years later, corpora reach millions of words. Although the notion of size is rather fluid, it is important to realize that size is a reflection of the purpose of the corpus. Basically, general corpora are often larger than specialized corpora in that the latter represent a smaller selection of language, yet Kennedy (1998, 22) asserts that big corpora do not represent a register or a language better than smaller ones since "we simple do not know how big a corpus needs to be for general or particular purposes".

Studies have shown that one million words are sufficient to obtain reliable and generalizable results, especially to address linguistic patterns of use and grammatical co-occurrences (Biber, 1993). Counter-arguments are found in Wales (1996, 197) "bigger corpora mean better judgments made on the basis of better evidence" and similarly in O'Grady and colleagues (2005, 609) "the larger the collection of texts comprising the corpus, the more useful it becomes, since the chances of its covering language as it is actually used increase". According to

Reppen (2010, 4) these principles should be applied to lexical investigation where larger corpora are needed "to ensure that all the senses of a word are represented". This had been anticipated by Sinclair (2004) who claimed that "the more you can gather, the clearer and more accurate will be the picture that you get of the language" because "several hundred instances of the simplest objects […] help to penetrate below the surface variation and isolate generalities".

4.3.3 Balance and variety

Corpus-based analyses have proved that there are considerable differences in the use of lexis, grammar and discourse features among language varieties. For this reason, it is essential to introduce language samples from varieties of topics, authors, registers and sources. A key procedure to succeed in representing linguistic varieties is to introduce a balanced range of samples, texts types or genres shared by a language community. As with representativeness and size, there is not a reliable and scientific measure of corpus balance. Rather, a wide range of text is said to be balanced and representative of a group if it is "typically and proportionally sampled" so as to ensure "a manageably small scale model of the linguistic material which corpus builders wish to study" (Atkins *et al*, 1992, 6).

4.3.4 Sampling

If balance and representativeness are considered a *sine qua non* of corpus design (especially of specialized corpora), sampling is just as important and closely related to both. Since natural language cannot be exhaustively described, compilers need to sample it according to pre-determined decisions so that the resulting corpus is a *sample* of a much larger *population*. In order to obtain a representative sample, the first concern to be addressed is to define the *sampling unit* and the boundaries of the *population*, or *sampling frame*. For written texts, for example, the *sample unit* may be a whole book or a fixed number of words; while a population can be defined using demographic distribution (e.g. sex, age, social class) or language as a product identifiable in the text categories or genres representing the subjects'discourse.

4.3.5 Chronology

Time criterion defines the span of time when the samples were produced, *i.e*. the period of time that the corpus covers. In terms of time, there are synchronic and diachronic corpora. A synchronic corpus is a static collection of texts, aiming at representing the language within a particular time-frame; whereas a diachronic corpus is dynamic and systematically embraces longer intervals of time in order to study

language changes and development (Rizzo, 2010, 3-4). With respect to a specific corpus compiled for terminological studies, it is advisable to gather samples covering the last 10 years prior to the date of compilation (Pearson, 1998, 51).

4.3.6 Gradual approximation

Balance, representativeness, sampling and variety are all principles that corpus builders strive for but rarely, if ever, fully attain. In truth, they are matters of degree and remain largely heuristic notions (McEnery and Hardie, 2012, 10). As Leech (2007) notes, the debate around balance, representativeness and variation might lead researchers to reject these concepts as problematic and unattainable, however,

> even if we cannot achieve them 100 per cent, we should not abandon the attempt to define and achieve them. We should aim at a *gradual approximation*[2] to these goals, as crucial desiderata of corpus design. It is best to recognise that these goals are not an all-or-nothing: there is a scale of representativity, of balancedness, of comparability. We

[2] Emphasis has been intentionally added.

should seek to define realistically attainable positions on these scales, rather than abandon them altogether (143-4)

There is little doubt that, as the corpus approach to language develops, the above concepts will undergo further critical scrutiny which, in turn, will lead to incrementally better definitions.

4.4 The *MedEnCor* corpus: design and compilation

From the onset of the research, the Medical English Corpus (*MedEnCor*, hereafter) has been planned to serve specific research purposes within the written register of the medical domain, therefore its design might be considered reasonably representative of the language of the scientific community it is addressed to. In pursuance of this aim, I have attempted to adhere as far as possible to the criteria listed above which have been completed with the recommendations from the literature on specialized corpora[3].

[3] Expert Advisory Group on Language Engineering Standards (EAGLES) is an initiative of the European Commission, within DG XIII *Linguistic Research and Engineering programme*, which aims to accelerate the provision of standards for:
(a) very large-scale language resources (such as text corpora, computational lexicons and speech corpora); (b) means of manipulating such knowledge, via computational linguistic formalisms, mark up languages and various software tools; (c) means of assessing and evaluating resources, tools and products (http://www.ilc.cnr.it/EAGLES/intro.html).

The *MedEnCor* is a written biomedical[4] corpus currently totalling 3,099,260 running words equally distributed across a balanced variety of 720 full-length documents[5] representative of 30 medical specialties. The corpus is synchronic, balanced and representative. Synchronicity is ensured by the limited time frame within which I have confined the selection of the data (from the year 2007 to present). As for balance and representativeness my analysis was based on a horizontal and vertical classification of medical communication in an attempt to embrace as many areas as possible, which, being in constant evolution are difficult to grasp. Thus, the horizontal division was made on the basis of different medical domains, the newly born included (*e.g*. Complementary Alternative Medicine). For the vertical division, instead, I distinguished according to the degree of specialization among partners and the text and genres involved in their communication[6] (Löning,1981, 83; Zweigenbaum *et al*, 2001, 248) (Table 4.1):

[4] Before entering into the design of the *MedEnCor* it is important to specify the meaning of *biomedical* in this project. Biomedicine is a wide area of research ranging from Biochemistry to Genetics, Pharmacology and Microbiology or even Environmental Sciences. Given the heterogeneous nature of the texts included in the collection I have applied the label *biomedical* to the *MedEnCor,* although the more generic term *medical,* will often replace it, hereafter.

[5] With the exception of samples from handbooks and encyclopaedias whose length never exceeded 2,000 words.

[6] Deliberately, communication involving non-professional partners (e.g. patients) has been excluded because it moves away from the criteria defined in the research aims and research questions.

Table 4.1. *Degree of specialization in medical communication and relative text-types/genres*

Communication partners: professionals-professionals (doctor-doctor)	
Aim:	dissemination of current specialized knowledge
Style:	scientific
Text types/genres:	abstracts; research articles; case studies; clinical trials; study reports
Communication partners: professionals-semi-professionals (doctor-medical students/health personnel)	
Aim:	transfer of basic knowledge.
Style:	educational; instructive
Text types/genres:	handbooks, textbooks, training manuals; dissertations and theses; conference reports
Communication partners: professionals-professionals-semi-professional (doctor-doctor-medical students)	
Aim:	reference knowledge
Style:	scientific; informative
Text types/genres:	encyclopedias; reviews; editorials; protocols and guidelines; *doc-to-doc* forums/blogs; PIL (Patient Information Leaflets) and healthcare information materials; study reports;

Therefore balance and representativeness are guaranteed on three levels: (1) by the range of texts types evenly drawn from 12 different medical genres characterizing written medical communication and the kind of literature likely to be consulted by both medical students and health care practitioners (two for each genre) (Table 4.2);

Table 4.2 *Variety and balance of texts/genres represented in the MedEnCor*

text-type/genre	***number***	***date***	***author***
Abstracts	2	Jan 2007-Feb 2015	Professional
Case studies;	2	Jan 2007-Feb 2015	Professional
Clinical trials .	2	Jan 2007-Feb 2015	Professional
Dissertations and theses	2	Jan 2007-Feb 2015	Lay
doc-to-doc forums/blogs;	2	Jan 2007-Feb 2015	Professional
Editorials	2	Jan 2007-Feb 2015	Professional/Lay
Encyclopedias	2	Jan 2007-Feb 2015	Professional
Handbooks and manuals	2	Jan 2007-Feb 2015	Professional
PIL/information materials	2	Jan 2007-Feb 2015	Professional/Lay
Protocols and guidelines	2	Jan 2007-Feb 2015	Professional
Research articles	2	Jan 2007-Feb 2015	Professional
Reviews	2	Jan 2007-Feb 2015	Professional/Lay

(2) by the comprehensive spectrum covered by the 30 medical specialties I decided to include (Table 4.3); (3) by the variety covered by the texts which encompass the main issues of each medical domain: *i.e. anatomy* (where the disease is located); *etiology* (what its cause); *pathology* (what

goes wrong in the body); *diagnosis* (how the disease is found) and *treatment* (how the disease is cured) [7] (Figure 4.1).

Table 4.3. *30 medical subject areas included in the MedEnCor*

01 Anaesthesia	16 Nutrition and Dietetics
02 Anatomy and tissue structure	17 Oncology and Cancer Metastatic Effects
03 Biomedical science Biotechnology Bioengineering	18 Orthopedics and Physiotherapy
04 Cardiology .	19 Orthotics, Ophthalmology and Optometry
05 Cardiovascular Physiology	20 Pain Management Medicine and Complementary Alternative Med .
06 Dentistry .	21 Paediatrics and Child Health
07 Dermatology	22 Pharmacology, Pharmaceutical Med, Antivirals and Antibiotics
08 Endocrinology-Diabetes	23 Primary Healthcare and General Practice
09 Gastroenterology and Hepatology	24 Psychiatry , Mental Health , Rehabilitation
10 Genetics and Metabolic Disorders	25 Pulmonology and Respiratory Medicine
11 Gerontology and Geriatric Medicine	26 Radiology, Medical Imaging and Radiotherapy
12 Gynaecology and Obstetrics	27 Rheumatology
13 Haematology	28 Sports Medicine
14 Infection, Immunology, Microbiology	29 Surgery, Surgical Specialties and Transplantation
15 Nephrology	30 Urology

[7]This scrupulous query was favoured by two courses I attended at the Faculty of Medicine of the University of Milan in order to learn how to search for topic related documents in bibliographic databases and to wring the best out of *PubMed* (the largest digital biomedical archive) using the right *MeSH* term (MEdical Subject Heading thesaurus): (1)"*PubMed*:La principale banca dati biomedica disponibile gratuitamente: come interrogarla al meglio delle sue grandi possibilità; (2) "*Google* in medicina. Informazioni sulle potenzialità del più noto motore di ricerca generalista per la ricerca bibliografica in campo medico e la condivisione scientifica tra ricercatori, medici, operatori della sanità".

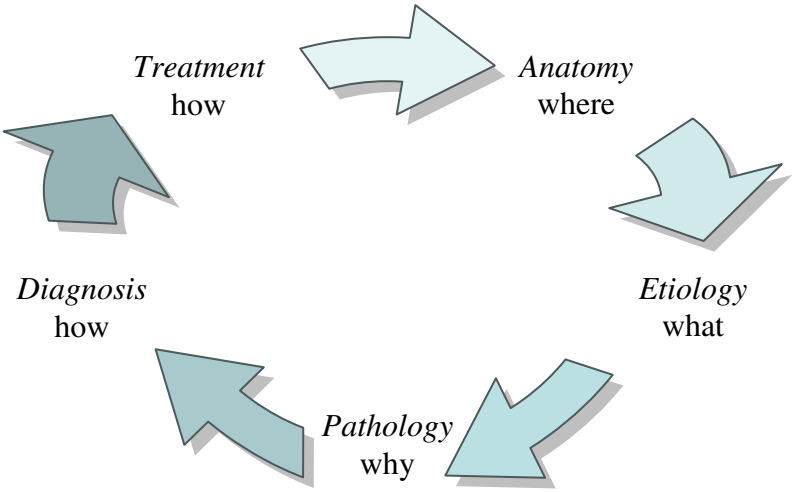


Figure. 4.1. *Guiding principles for the text selection*

Additionally, in order to assure authenticity, all the texts included in the *MedEnCor* had to be written either by native English speakers or by professionals affiliated with an institution where English is spoken as first language so as to make the corpus linguistically more homogeneous and to minimize the non-native writer's influence

4.5 Data collection and processing

The addition of documents to the corpus comprises several steps. The documents must first be obtained and this raises problems of property permissions. Consequently, to overcome copyright issues, most texts were downloaded from *PubMed* the most important and widespread free digital archive of biomedical and scientific literature (http://www.ncbi.nlm.nih.gov/pmc/). PubMed includes more than 24

million citations for biomedical literature from MEDLINE(the database of biomedical literature of the United States National Library of Medicine), life science journals, and online books. Citations may include links to full-text content from PubMed Central and publisher web sites. Other useful websites used to retrieve scholarly and well-documented materials were:

- *DART-Europe E-theses Portal* (http://www.dart-europe.eu/basic-search.php): an up-to-date source for medical dissertations and theses;
- *Drugs.com* (http://www.drugs.com/): the most comprehensive site of drug information online, providing free and peer-reviewed data on more than 24,000 prescription drugs and natural products;
- *MedicalStudent.com* (http://www.medicalstudent.com/#Learn): a digital library of authoritative medical education information for the medical students
- *doc2doc* (http://doc2doc.bmj.com/): a professional networking community for healthcare professionals worldwide provided by the *British Medical Journal*. It is an independent and secure medium of communication aiming to improve the working lives of doctors and other healthcare professionals who can meet and talk about their clinical and non-clinical interests on discussion forums, blogs, or through direct contact with other members.

Texts were all kept at their original length[8]. Writers' names, footnotes, endnotes, acknowledgments, bibliographies, figures, captions and tables were omitted when the texts were *txt*-edited into the corpus so as to eliminate the factors unrelated to the lexical analysis and to ensure that the texts stored in the corpus were readable by the software.

A minimal structural encoding was introduced to facilitate the storage of the samples and guarantee an accurate file classification, hence domains and genres were abbreviated using the *Acronym Generator* freeware https://www.cs.uoregon.edu/research/paracomp/anym/ (Table 4.4 and Table 4.5)

Table.4.4. *Text-types and genres: acronyms*

ABS	BLG	REV	CST	EDT	ENC	GUI	HAN	PIL	RAR	THS	TRL

Table.4.5 *Domains: acronyms*

ANAT	ANES	BIOT	CARD	CARV	DENT
DERM	DIET	ENDO	GAST	GENE	GERO
GYNO	HEMA	IMMU	NEPH	PULM	ONCO
ORTP	OPHT	PEDC	PHAR	PHEA	PSYC
PULM	RADI	RHEU	SPOR	SURG	UROL

[8] With the exception of samples from encyclopaedias and handbooks (about 2,000 words).

So that, for example, file n.1 corresponding to a research article on oncology was stored as ONCO-RAR-1.

The data was processed and analysed with the software suite *WordSmith Tools 6.0* (Scott, 2012) and its three main functions: *WordList; Concord* and *KeyWord* and with Cobb's (2005) *Compleat Lexical Tutor* a web-based suite for lexical analysis (downloadable from http://lextutor.ca/) that arranges words of texts into either first and second thousand levels, academic words, and the remainder offlist[9].
All these lists encompass both the base and the derived forms of the words.

4.6 Data analysis

I started my analysis by generating a simple list of single frequent words using the *WordSmith Wordlist* . Frequency lists are useful because they both give us some idea of what texts are about and rank words in descending order of frequency (Bowker 2002, 145). Before investigating data I set a general criterion for word selection, *i.e.* what I counted as a word. My choice fell on lexical items and their related inflected and

[9] The Vocabulary Profiler is based on Laufer and Nation's Lexical Frequency Profile so that a text is broken into four word lists: the first (BASEWORD.1) includes the most frequent 1000 words of English; the second (BASEWORD.2) includes the second 1000 most frequent words; the third (BASEWORD.3) comprises words frequent in university texts from a wide range of subjects; and the fourth list (NOT IN THE LISTS.4) contains those words that are not included in the previous three lists (highly technical words).

derived forms. This decision was guided by three main principles: (1) medical lexis was the core of the overall design of my project; (2) it was also the specific objective that my research questions were trying to achieve and, (3) it was supported by studies which asserted that comprehending word families or controlling word building does not require much effort by learners if they know the base root (Bauer and Nation 1993, 235).

4.6.1 Word-list and basic statistical information

WordSmith WordList counted 3,099,260 *tokens* (*i.e.* sequences of characters divided by blank spaces or punctuation marks, often repetitions of same words) and 60,916 *types* or word-forms (*i.e*. number of different words, including each form derived from a main lemma or headword). Basically, the set of *types* constitutes the vocabulary of the corpus.

The relationship existing between the total number of types and tokens is given by the type/token ratio (TTR): a helpful measure of vocabulary variation within a written text which can be calculated as follows: (Table 4.6)

Table 4.6 *Type/token ratio (TTR)*

*type-token ratio = (number of types/number of tokens) * 100*
= (60,916/3,099,260)* 100=37,52

The more types there are in the corpus in comparison to the number of tokens, then the more varied is the vocabulary, *i.e.* there is greater lexical variety (on the contrary, a lower ratio means a lower lexical burden due to the repetitions of the same forms). Other interesting statistical information retrieved from the *MedEnCor* is given in Table 4.7:

Table 4.7. *The MedEnCor Corpus: basic statistical information*

No. of tokens	3,099,260	
No. of types	60,916	
Standarised Type/Token ratio	37,52	
Mean word lenght	05,25	
Hapax legomena =1	21,797	<36%
Dis legomena =2	9,148	>15%
Tris legomena=3	4,444	> 7%

From the 3 million-word sample, 21,797 words occur only once, which correspond to the 36% of the types; whereas 35,389forms occur from one to three times, totalling about the 58%.

Table 4.8 displays the top-35 items in the frequency-sorted word list: all the items in the list (except for *patients*, n. 21) are function words and

neither tell us much about what the texts in the corpus are about, nor fall within my criteria for word selection.

Table 4.8. *Top 35 most frequently occurring items from the MedEnCor*

Rank	***Word***	***Frequency***
1	*#*	214664
2	*the*	146603
3	*of*	113724
4	*and*	86593
5	*in*	66650
6	*to*	55521
7	*a*	46888
8	*is*	37160
9	*with*	29749
10	*for*	27939
11	*or*	22214
12	*be*	19459
13	*are*	16961
14	*as*	16638
15	*by*	15700
16	*that*	14886
17	*on*	12680
18	*was*	12606
19	*this*	12027
20	*it*	11655
21	***patients***	**11015**
22	*may*	10770
23	*at*	10592
24	*from*	10487
25	*an*	10068
26	*were*	9740
27	*not*	9354
28	*have*	7131
29	*al*	7046
30	*et*	7040
31	*which*	6856
32	*can*	6330
33	*should*	6029
34	*s*	5854
35	*if*	5838

However, they can be useful for creating a *stop list* containing the words I wish to exclude from my analysis.

Notably, (Table 4.9) *the* is the most frequent word in the corpus and stands for 4.73% of the total tokens. In general language, the three most frequent words commonly reach an 11% of the whole, the 10 most frequent ones a 22%, the 50 most frequent words cover around the 80% (Schmitt, 2000). Those figures agree with the results obtained from the *MedEnCor* with only some slight differences:

Table 4.9. *Coverage of the most frequent words in the MedEnCor*

Most frequent words	*Coverage in general language*	*Coverage in MedEnCor*
3	11%	11.19%
10	22%	21.05%
50	37%	38.30%
100	44%	43.67%
2000	80%	72.76%

4.6.2 Generating keywords

Frequency wordlists, although an important first step in corpus analysis, in that they provide an immediate snapshot of the characteristics of a particular language variety, do not give useful indications of what is really important or unusually frequent in a corpus (Harvey, 2013, 58). *Keywords*, on the other hand, being word forms that

occur in a corpus with a greater significant frequency, provide a measure of saliency as opposed to pure frequency and thus, are a more sensitive measure of quantitative analysis than frequency lists (Baker, 2006, 125). Importantly, *keywords*[10], according to Scott (2001), are often likely to be words that human beings would identify as being *thematically* central to a text and are thus indicators of the 'aboutness'[11] of a particular corpus. Consequently, in order to identify the key health themes distributed across the *MedEnCor*, I supplemented the findings derived from the raw frequency lists with a keyword analysis. To detect the most outstanding or unexpectedly frequent words I used the *KeyWord* tool from *WordSmith* suite. I compared the 'purified' frequency wordlist with a wordlist from a larger and more general reference corpus: the *British National Corpus* (*BNC* available at http://ucrel.lancs.ac.uk/bncfreq /flists.html and http://www.natcorp.ox. ac.uk/). The idea was to contrast the frequency of the words in the *MedEnCor* with the frequency of the words in naturally occurring language (represented by the wordlist from the *British National Corpus*). Words which repeatedly appeared in the

[10] In their ability to indicate the propositional content of a text, keywords are an important tool for conducting discourse analysis, too. Indeed, since they are also an important indicators of style as well as content, keywords have been used by a number of researchers as a useful means of identifying writers' and speakers' positions in texts, revealing insightful information about the values and beliefs expressed by language users in a range of communicative contexts (Stubbs 2010, 24).

[11] Scott (1999) says that keyword lists tend to show up three types of words: (1) proper nouns; (2) 'aboutness' keywords, *i.e*. lexical words (nouns, verbs, adjectives, adverbs) which are generally those which are most interesting to analyse; (3) high frequency grammatical words, which may be more indicative of style than 'aboutness'.

MedEnCor but rarely in the reference corpus were probably 'candidate terms': *i.e.* words that "are used in a specialized domain and have a clearly identified meaning" (Bowker 2002, 145).

The program[12] generated a volume of 4,519 words, automatically reduced to 500 to be examined individually in contextual detail. The items yielded by *WordSmith* were identified as the most relevant single key words[13] (Appendix 2). Interestingly, the list contained far more lexical words than the already produced frequency list. It is clear, therefore, that keywords are a more efficient way of identifying those words that are most typical of a particular domain and deserve further analysis. Bondi (2010, 4) draws our attention on the underlying metaphor of the notion of a 'key': a tool that affords access to somewhere or something; a metaphor which suggests the power of opening and revealing what is unknown. Admittedly, detecting,

[12] *WordSmith* empirically compiles keywords according mechanical criteria: the program takes into account the size of each corpus and the frequencies of each word within them. It then carries out statistical tests on each word (the user can specify the *chi-squared* or *log-likelihood* test) which gives each word a *p* (or probability) value. The *p* value (a number between 0 and 1) indicates the amount of confidence that a word is key due to chance alone the smaller the *p* value, the more likely that the word's strong presence in one of the corpora is not due to chance but a result of the author's (conscious or subconscious) choice to use that word repeatedly. Because every word in the corpora is assigned a *p*-value, as corpus users it is up to us to decide how low the *p*-value needs to be before we label a word as a key (Scott, 2012).

[13] "There can be no guarantee that the *keywords* are 'key' in the sense which you may attach to 'key'[…] they are merely the words which are outstandingly frequent or infrequent in comparison with the reference corpus"(Scott Mike, 2012) (Available at http://www.lexically.net/downloads/version6/HTML/index.html? keywordsadvice info.htm)

categorizing and making connections between keywords constituted an important analytical step in this research, thus, I tried to open the door to the *MedEnCor* using these salient lexical items to gain access to aspects and features of the corpus that were hidden or not obvious.

4.6.3 Profiling keywords

After ranking words according to their 'keyness', the next step was to produce a lexical analysis that might help with the identification of those 'candidate terms' eligible as *core lexis.* I combined two kinds of approaches: statistical and linguistic. For the statistical approach, I completed my selection using Cobb's *Compleat Lexical Tutor*, a lexical profiler which arranged the *MedEnCor* top 500 keyword into four word K-lists[14], as shown in Table 4.10:

[14] All these K-lists encompass both the base and the derived forms of the words. The sources of the lists are: West's *General Service List* (1953) for the first 2,000 words and Coxhead's *Academic Word List* (1998) for the additional 570 academic word families. The 'Not in the list' contains those words that are not included in the previous three lists*, i.e.* specialized words.

Table 4.10 *Distribution of the 500 top keywords after running Compleat Lexical Tutor*

	Families	*Types*	*Tokens*	*Percent*
K1 word (1-1000):	63	78	78	16.18%
Function:	...	...	(6)	(1.24%)
Content:	...	...	(72)	(14.94%)
> Anglo-Sax. =Not Greco-Lat./Fr. Cog.	...	...	(19)	(3.94%)
K2 word (1001-2000):	37	47	47	9.75%
>Anglo-Sax.	...	...	(17)	(3.53%)
1k+2k		...	...	(25.93%)
K3 word (academic words):	47	58	58	12.03%
> Anglo-Sax.	...	...	(2)	(0.41%)
Not in the list:	?	299	299	62.03%
	147+?	482	482	100%

The detailed output of this analysis is given in Table 4.11:

Table 4.11. *Lexical profiling of the 500 keywords into Cobb's K-groups*

K1 word (1-1000):
activity, age, associated, based, bleeding, blood, care, cases, cause, center, changes, characterised, common, commonly, condition, controlled, due, effect, exercise, eye, failure, figure, findings, flow, follow, group, heart, high, include, increase, increased, is, levels, low, lower, may, measured, measurement, measures, observed, post, presence, pressure, prevention, rate, reduction, related, reported, respectively, results, secondary, should, signs, study, table, term, test, testing, total, trial, type, use, used, using, usually, values, vessels, weeks, with

.K2 word (1001-2000):
angle, blindness, bone, brain, compared, complications, decrease, delivery, disease, during, examination, frequency, health, healthy, lid, medicine, medicines, mice, mild, model, moderate, multiple, nurse, nursing, pain, patient, performed, practice, program, pump, ray, recommended, review, risk, self, severe, severity, skin, staining, swelling, treated, treatment

K3 word (academic words):
abnormal, analysis, approximately, assessment, computed, criteria, data, demonstrated, depression, device, duration, evaluated, evaluation, evidence, exposure, factor, function, functional, guidelines, incidence, indicated, induced, injury, intensity, intervention, journal, medical, methods, normal, occur, outcome, parameters, participants, phase, physical, positive, primary, procedure, ratio, response, significant, significantly, specific, stress, technique, thesis, topical, validation, vision, visual, volume

Not in the list
abdominal, abnormalities, abscess, acid, activation, acute, administered, adverse, airway, allergic, anaesthesia, analgesia, angina, angiography, ankle, anterior, antibiotic, aortic, aqueous, arterial, artery, arthritis, atopic, atrophy, attenuation, axial, bacterial, baseline, beta, bilateral, biopsy, blockers, breast, cancer, carcinoma, cardiac, cardiovascular, cataract, catheter, cell, cellular, cervical, chamber, chemotherapy, cholesterol, choroid, chronic, ciliary, clinical, clinically, clopidogrel, cohort, concentrations, congenital, conjunctivitis, cornea, corneal, coronary, defects, deficiency, degeneration, dermatitis, detachment, diabetes, diabetic, diagnosis, diagnostic, diastolic, dietary, disorders, dna, dose, drug, dysfunction, ecg, efficacy, elevated, epidural, epithelial, epithelium, estrogen, et, etiology, extraocular, eyeball, fetal, fluid, fractures, fundus, gene, genetic, glaucoma, glucose, haemorrhage, herpes, HIV, hypertension, hypotension, imaging, immune, implant, incision, infants, infarction, infection, inflammation, inflammatory,

infusion, inhibitors, injection, insulin, intake, intraocular, intravenous, invasive, iris, ischemia, ischemic, itching, kidney, lacrimal, lateral, lens, lesion, limbus, lipid, liver, macular, magnetic, malignant, maternal, medial, medication, membrane, meta, metabolic, metabolism, methotrexate, mitochondrial, morbidity, mortality, muscle, myocardial, myocardium, myopia, nasal, nausea, necrosis, neonatal, nerve, nutrition, obesity, obstruction, occlusion, ocular, oedema, ointment, onset, ophthalmic, ophthalmology, optic, oral, orbital, oxidative, oxygen, pathway, patients, pediatric, pelvic, perforation, perfusion, peripheral, pharmacist, physiological, plasma, posterior, postoperative, pregnancy, prescribing, prevalence, prognostic, prophylaxis, protein, psychosocial, pulmonary, papillary, randomized, receptor, rectus, recurrent, reductase, reflex, refractive, regression, renal, respiratory, retina, retinal, retinopathy, rheumatoid, rupture, sclera, score, serum, sodium, stenosis, stent, steroids, supplementation, surgery, surgical, symptoms, syndrome, systemic, systolic, tablets, therapeutic, therapy, thoracic, thrombosis, thyroid, tissue, tomography, tract, trauma, tumor, tumour, ulcer, ultrasound, urinary, urine, uterine, vaginal, van, vascular, venous, ventricular, versus, viral, vitamin, vitreous, vomiting

For the linguistic approach, I used *WordSmith Concord* function, which allowed me to scrutinize concordance lines and spot additional information on ambiguous terms that deserved critical attention, by simply studying the behaviour of the words in the vicinity (Table 4.12). For example, the program ranked the keyword ‘vessel’ in the *K-1 list*, but given its polysemic nature, a deeper investigation showed that the term in the corpus was used with the highly technical meaning of “artery or vein carrying blood” far distant from the more familiar term “large ship travelling on water” which West had obviously included in the first 1,000 words.

Table 4.12. *Concordance lines for 'vessel'*

as a result, blood	**vessel**	relax and blood pressure
the inflammation of blood	**vessel**	in the brain caused the
heart is healthy, the blood	**vessel**	flawless from thickening
inhibit large blood	**vessel**	and clots and soiled

Chapter 5
Results and discussion

5.1 Results and discussion

Scrutinizing large numbers of wordlists and keywords allowed me to disclose a unique linguistic repertoire mostly composed of lexical or content words related to medicine with only few instances of grammatical words. Not only did these lexical items provide informational insights into the 'aboutness' of the corpus, but they also helped me delineate the thematic focus of the *MedEnCor* itself. Accordingly, these keywords were able to reveal the prevalent concepts that are at the heart of medical communication and, implicitly, their distinctiveness guided me to detect an appropriate answer to my first research question: (1) exactly, what kind of words make up the medical lexis that medical undergraduates and practitioners need?

5.2 Arranging keywords into semantic domains

Looking back at the four groups of keywords generated by *Compleat Lexical Tutor*, I selected only those items included in *K3 word* (academic words) and *Not in the list* (specialized terms), leaving out *K1 word (1-1000)* and *K2 word (1001-2000)* being directed to lower intermediate learners. Then, I grouped this new block of 357 keywords into four semantic domains indicative of the traditional health issues, which I labelled:

(1) *Healthcare*: related to the health services offered to a patient population;

(2) *Medical science*: comprising the branches and specialties dealing with the maintenance of health and the prevention and treatment of diseases;

(3) *Clinical terms*: embracing all those terms referring to the study and practice of medicine by direct examination of patients; and

(4) *Body structure and functions*: describing the human organism and the related bodily activities (Table 5.1)

Table 5.1. *Keywords arranged into 4 semantic domains*

Healthcare
guidelines, journal, medical, participants, catheter, dietary, infants, injection, intake, maternal neonatal nutrition, ointment, patients, pediatric, perfusion pharmacist, prescribing, psychosocial, surgery, surgical, tablets,

Medical Science
computed, device, positive, thesis, technique, volume, acid, activation, anaesthesia, analgesia, antibiotic, bacterial, baseline, blockers, cholesterol, concentrations, congenital, drug, estrogen, fetal, fluid, gene, genetic glucose imaging, implant, insulin, inhibitors, lens lipid, magnetic, meta, metabolic, metabolism, ophthalmic, ophthalmology, optic, oxidative, oxygen, protein, reductase, sodium stent, steroids tomography, ultrasound, vitamin

Clinical terms
abnormal, abnormalities, abscess, acute, adverse, allergic, analysis, angina, angiography, arthritis, atopic, assessment, clinical, clinically conjunctivitis, criteria, data, defects, deficiency, degeneration, depression dermatitis, detachment, diabetes, diabetic, diagnosis, diagnostic, disorders, dose, dysfunction, duration, evaluation, evidence, exposure, efficacy, epidural, etiology, glaucoma, haemorrhage, herpes, hypertension, hypotension, immune, incision, infarction, infection, inflammation, inflammatory, infusion, injury, factor, incidence, intensity, intervention, invasive, ischemia, ischemic, itching, lesion, macular, malignant, medication, morbidity, mortality, myopia, nausea, necrosis, obesity, obstruction, occlusion, oedema, onset, pathway, perforation, physiological, postoperative, pregnancy, prevalence, prophylaxis, prognostic methods, normal, outcome, parameters, phase, primary, procedure, ratio, response, significant, significantly, specific, stress, topical, validation, randomized, recurrent, reflex, regression, retinopathy, rheumatoid, rupture, score, stenosis, symptoms, syndrome, systemic, systolic, therapeutic, therapy, thrombosis, tumor, tumour, ulcer, viral, vomiting, trauma

Body structure and functions
function, functional, physical, vision, visual, abdominal, airway, ankle, anterior, aortic, aqueous, arterial, artery, atrophy, axial, bilateral, breast, cardiac, cardiovascular, cell, cellular, cervical, chamber, choroid, ciliary, cornea, corneal, coronary, diastolic, epithelial, epithelium, extraocular, eyeball, fractures, fundus, intraocular, intravenous, iris, kidney, lacrimal, lateral, limbus, liver, medial, membrane, mitochondrial, muscle, myocardial, myocardium, nasal, nerve, ocular, oral, orbital, pelvic, peripheral, plasma, posterior, pulmonary, papillary, receptor, rectus, refractive, renal, respiratory, retina, retinal, sclera, serum, thoracic, thyroid, tissue, tract, urinary, urine uterine, vaginal, vascular, venous, ventricular, vitreous

Highly specialized terms were excluded from the lists (*ecg, clopidogre, DNA, herpes, HIV, methotrexate, oedema, ulcer)* because their technicality makes their meaning and usage easily manageable.

Studying the four semantic areas, it was evident that the keywords included in the domain labelled 'Clinical terms', immediately offered a reliable overview of a number of themes salient in scientific communication since they broadly corresponded to the pillars of the medical care process. Therefore, my attention focused on these keywords, which I considered as candidate terms of the *core lexis* of medicine, leaving the remainder of the terms to a further investigation at a later time.

After a more scrupolous scrutiny, it was interesting to notice how most of these keywords had Latin and Greek etymology. This was confirmed by Cobb's *Compleat Lexical Tutor* analysis (Table 5.2) which proved that 96.67% of these words were Latin and Greek cognates and

largely outnumbered the 3,33% of the English medical terms, labelled as Anglo-Saxon tokens (*detachment, itching, onset, outcome, pathway, randomized, score, stress*)

Table 5.2. *Etymology of clinical keywords*

Tokens per family:	30
Types per family:	30
Anglo-Sax Index: (A-Sax tokens + functors / onlist tokens)	3.33%
Greco-Lat/Fr-Cognate Index: (Inverse of above)	96.67%

Assuming that the first step towards the competent usage of these words might be eased by etymology awareness, I tried to identify and describe their prefixes and suffixes[1] (Table 6.3)

Table 5.3 *Prefixes and suffixes of 'Clinical terms'*

Keyword	*Prefix-suffix*	*Meaning*	*Etymology*
abnormal	-ab-	*away from*	Latin
abscess.	-ab-	*away from*	Latin
adverse	-ad-	*contrary, opposing*	Latin
allergic	-ic	*pertaining to*	Latin
analysis	-ana	*to loosen; to break up*	Greek

[1] I used as a reference the *McGraw-Hill Concise Dictionary of Modern Medicine.* Available at http://medical-dictionary.thefreedictionary.com/

angina	-ango	*strangling*	Greek
angiography	-angio, graphy	*blood vessel, recording*	Greek
arthritis.	-arthr,-itis	*joint; disease*	Greek
atopic	-a; -topos	*absence of; place,*	Greek
clinical	-clin-ic	*bed, pertaining to*	Greek
conjunctivitis	-itis	*inflammation, disease*	Greek
defects	-de	*away from*	Latin
deficiency	-de	*away from*	Latin
dermatitis,	-derma-itis	*skin, inflammation*	Latin
diabetes,	-dia	*across, passing through*	Greek
diagnosis,	-dia, gnosis	*through, knowledge*	Greek
disorders,	-dis	*bad*	Greek
dysfunction,	-dys	*abnormal*	Latin
epidural	-epi	*upon*	Latin
evaluation,	-ex	*beyond*	Latin
exposure,	-ex	*beyond*	Greek
etiology,	-logy	*cause, study of*	Greek
glaucoma,	-glauco, -oma	*bluish, tumor*	Greek
haemorrhage,	-haem, gia	*pertaining to blood*	Greek
hypertension,	-hyper	*above, over*	Latin
hypotension,	-hypo	*under*	Latin
immune,	-im	*free from*	Latin
incision,	-in	*into*	Latin
infarction,	-in	*into*	Latin

inflammation	-in, -flam	*fire*	Latin
injury,	-in	*opposite*	Latin
intervention,	-inter	*between*	Greek
ischemia,	-isch-emia	*restriction, blood condit.*	Greek
macular	-macula	*spot*	Latin
malignant	-mal	*bad*	Greek
myopia,	-my	*near*	Greek
necrosis,	-necro, -osis	*dead, condition*	Greek
obesity,	-ob	*over*	Latin
obstruction,	-ob	*against*	Latin
parameter	-para	*alongside, besides*	Greek
perforation,	-per	*through*	Latin
physiologic	-phys, logy	*body, study*	Latin
postoperative,	-post	*after*	Latin
pregnancy	-pre	*before*	Latin
prevalence,	-pre	*before*	Latin
prognostic	-pro, gnosis	*in advance, knowledge*	Greek
prophylaxis,	-pro	*in advance*	Greek
primary,	-prim	*first*	Latin
recurrent,	-re	*back*	Latin
reflex	-re	*back*	Latin
regression,	-re	*back*	Greek
rethinopathy	-pathy	*disease*	Latin
rheumatoid	-oid	*resembling*	Greek

rupture	-rupt	*break*	Latin
stenosis,	-steno, sis	*narrow, condition*	Greek
symptoms,	-sym	*likeness*	Greek
syndrome,	-syn	*together*	Greek
systemic,	-system	*body*	Greek
topical	-topos, -ic	*local place, pertaining to*	Greek
thrombosis	-thromb, osis	*blood clot, disease*	Greek
trauma	-traum	*-wound*	Greek

5.3 Identifying semantic sub-layers

Visualizing etymology and relating the meaning of the words parts to the meaning of the words opened up a new scenario where I could spot four new semantic layers, hierarchically identifiable inside the original domain of 'Clinical terms', which, this time, I labelled as:

(1) C*linical practice* (Table 5.4)

(2) *Signs and* s*ymptoms* (Table 5.5)

(3) *Pathology and disorders* (Table 5.6)

(4) *Cure* .(Table 5.7)

Table 5.4. *Keywords related to 'Clinical practice'*

Nouns	*Adjectives*	*Adverbs*
analysis, assessment, criteria, data, dose, duration efficacy, evaluation, evidence factor, incidence, intensity, methods, mortality, parameters, phase, prophylaxis, procedure, ratio, regression, response, therapy, validation	*clinical, normal, postoperative, recurrent, significant, specific, therapeutic*	*clinically, significantly*

Table 5.5. *Keywords related to 'Signs and symptoms'*

symptoms	*-itis (inflammation)*	*-flam (fire).*
symptom	*arthritis, conjunctivitis, dermatitis*	*inflammation, inflammatory*

Table 5.6. *Keywords related to 'Pathology and disorders'*

Abnormalities (prefixes of movement)	*Diseases (-pathy)*	*Adjectives*
abnormalities, abscess, defects, deficiency degeneration, depression, diabetes, disorder, dysfunction hypertension, hypotension, injury, myopia, obesity	*retinopathy*	*abnormal, adverse, allergic atopic, invasive macular, malignant, viral*
	(-osis)	
	necrosis, thrombosis	
	(tumor)	
	glaucoma, tumor, tumour	
	(restriction)	
	angina,obstruction, occlusion, stenosis, ischemia,	
	(breaking)	
	haemorrhage, infarction, lesion, rupture perforation, trauma	

Table 5.7 *Keywords related to 'Cure*

nouns	*Adjectives*	*adverbs*
incision, epidural infusion intervention medication prophylaxis	*primary, physiological, postoperative, topical, systemic*	

The compilation of these new semantic sub-layers highlighted two distinctive traits of the *core lexis* of medicine under investigation: (1) at any level of specialization or technicality medical language is characterized by compression and precision, or, as Halliday notes (2004, 656) by "information density" which condenses a great deal of concepts into few words; (2) since, by definition, medicine is the science encompassing a variety of practices aimed at maintaining and restoring health by the prevention and treatment of illnesses in human beings; no wonder if these keywords and semantic domains and sub-domains all rely on the pillars of the medical discipline that are and, have always been: prevention, diagnosis, treatment.

Unsurprisingly, these new findings seem to bring the situation back to the compilation of the *MedEnCor*, when, for reasons of balance and representativeness, it was decided to include in the corpus all those texts that covered the central issues of medicine and were related to diagnosis, anatomy, etiology, pathology and treatment. Well, then, in the middle of more than three million words, those guidelines have materialized in real and concrete word lists: this is clear evidence that in corpus linguistics, nothing is created and nothing is destroyed, but everything always returns.

Finally, my last point is a question of method: it must be stressed that the findings from my exploration constitute just a sample of the health science discourse and, however representative of a particular collection

of medical texts, they are just observations, and as such they “must be dealt with as deductions rather than as facts” (Hunston, 2002, 23).

Chapter 6

Pedagogical implications

6.1 Learning the language of medicine

The study of medicine is certainly one of the most demanding educational programs in terms of amount and complexity of subject matter, sheer time involved as well as aptitudes and abilities encompassing inductive reasoning, finger dexterity and memory. The latter, *de facto*, constitutes more a challenge than a real skill since medical students, junior clinicians, biomedical researchers and senior physicians are not only required to describe the human body, its various parts, symptoms, diseases and treatments but, more diligently, they are also expected to constantly update the ever-expanding number of medical terms used to describe their everyday practice. With the rapid establishment of English as the *lingua franca* of scientific communication, medical students and healthcare practitioners are also faced with the need to use those terms in English, which unquestionably constitutes a crucial problem, especially with non-native speakers.

To meet these ends, in Chapter 4 and 5 I explained how semantic groups of keywords extracted from a specialized corpus deserved attention as representative of the *core lexis* of medicine. In this chapter,

instead, I introduce the *MedEnCor-Lex*, a web-based lexical database devised to help non-native researchers from different medical fields learn those keywords in context in order to produce appropriate and efficient native-like writings.

6.2 Compiling a web-based glossary to learn the *core lexis* of medicine

The word glossary originates from the Latin word *glossarium* which in turn is derived from the ancient Greek γλῶσσα (glossa) indicating an explanatory note attached to the side of a term difficult to understand. Today a glossary is essentially "an alphabetical list of technical terms, different from a dictionary because it is usually limited to some specialized fields of knowledge"[1]. The amount of information and details contained in a glossary can vary greatly depending on the purpose for which it is intended. Thus, it may be a simple collection of terms and meanings or a richly detailed inventory containing definitions, related terms, usage notes and examples. This was exactly the idea of glossary I had in mind when, after identifying the *core lexis* of medicine, I tried to find an answer the second research question of this study: how should such vocabulary be learned and taught?

[1] http://dictionary.cambridge.org/dictionary/british/

6.2.1 Why a web-based glossary?

Given the increasing number of electronic dictionaries providing easy and updated resources, what is the point of compiling a medical monolingual glossary? The reason is that, although dictionaries can be invaluable for solving some types of language problems, they are not always sophisticated enough to meet all the needs of non-native health care users, as they may not contain specific contextual and phraseological information. Even monolingual specialized dictionaries dealing with specific domains, sometimes tend to concentrate on providing information about the meaning rather than the usage of terms, which can be problematic particularly in cases where a word has more than one meaning.

Conversely, medical experts and novices must pay attention to how terms are used and how terms combine in sentences if they want to write native-like scientific texts and share the results of their research and practices with the scientific community. They are already familiar with the terminology of their discipline in their language but they need to acquire how to use it appropriately, in English. These kinds of information can be provided by presenting words in context instead of in isolation. Thus, sets of concordance lines, displaying target words *i.e*. the words being studied, surrounded by their context (sorted to the left or to the right, or both) become easy and useful instruments to learn how words are frequently used.

6.3 Corpora and pedagogy: the Data Driven Learning approach

The idea of using corpus evidence in the teaching and learning of languages is not new. There is indeed a wide range of fully corpus-based reference works available to learners and teachers, and a number of concrete suggestions on how concordances and corpus-derived exercises could be used in the language teaching classroom, thus significantly "enriching the learning environment" (Aston 1997, 51). Indicative of the popularity of pedagogical corpora use is the considerable number of books and edited collections that have recently been published on the topic, either using corpora as tools, *i. e*. the actual text collections and software packages for corpus access; or as methods, *i. e*. the analytic techniques that are used while working with corpus data (Bernardini 2002; Ghadessy, Henry, Roseberry, 2001; Aston, Bernardini, Stewart 2004; Sinclair 2004b; Gavioli 2006).

Of particular relevance to my study was the direct application of corpora to pedagogy introduced by the pioneering works of Tim Johns and Philip King, who, back in the 1980s, started the 'Data-Driven Learning'(DDL) or 'classroom concordancing' approach. It referred to the use of computer-generated concordances in the classroom to get students to "explore the regularities of patterning in the target language" with the development of "activities and exercises based on concordance output. (Johns & King,1991, iii). Today, the term remains controversial with researchers applying the label 'DDL' almost at random to a range

of activities, although in the founders' intentions it was associated with "an inductive, discovery-based approach to learning in which students work out rules of probabilities from the examples provided" (Flowerdew, 2012, 197). Bernardini (2000) proposes the notions of "serendipity learning"; "learner-as-researcher" and "hands-on-activities" as characteristic advantages of the DDL approach, with students working with worksheet output of concordance data thereby familiarizing with corpus methodologies and interpretation of frequency. Boulton (2011) defines DDL as "the freest form of corpus consultation, where learners take on complete responsibility for their learning" especially for "error-correction or written production" though mostly at higher level of education. Boulton insists that the non-dogmatic nature of DDL makes this approach fully compatible with learner-centered learning and learning by doing with an emphasis on authentic language, which, in turn, make DDL particularly useful in language for specific purposes.

6.4. The *MedEnCor-Lex* database

In view of the above considerations, it seemed reasonable to adopt the DDL approach even for the core lexis of medicine. By uncovering patterns of real language use stretched in concordance lines I did not only expect to ease the effective use of *keywords* in the medical specialized register, but also transform students from language learners

to autonomous language users. Basically, five were the distinctive objectives I had in mind while creating the *MedEnCor-Lex* database: (1) to help users learn English and use the language for professional purposes; (2) to raise users' awareness towards the importance of authentic medical documents written in English; (3) to make users proficient in using resources readily available online; (4) to make students autonomous in their learning; (5) to foster attitude towards language use, especially after graduation when there are fewer opportunities to receive language training at the workplace.

In point of fact the *MedEnCor* is not only an instrument likely to satisfy the diverse language needs of its users either for distance education or self-learning but also an e-learning system intended to maintain, improve and broaden the medical linguistic knowledge and skills as well as to develop a positive orientation towards continuing specialized development.

To begin with, my intention was to exploit the data already collected in the *MedEnCor* Corpus, simply moving them to an open source content management platform[2], where information could be easily stored and retrieved, considering my limited knowledge in programming and computational analysis.

My choice fell on *Drupal* (https://www.drupal.org /about): a free software package equipped with a powerful blend of features for building dynamic web sites. It offers a broad range of services including

managing online structures; designing and editing flexible and engaging contents; adapting and customizing materials. Additionally, being web-based, it is also compatible with all operating systems.

6.4.1 Designing the logo and the homepage

In order to configure and customize the *MedEnCor-Lex* domain (www.medencor.com), I also devised a logo which might be visually appealing, clearly organized and effectively relevant. After visiting a large number of existing medical websites, two were the factors that I elected as distinctive of healthcare: (1) neat reassuring colours and (2) engaging design. Long hours of drawing and assembling led to the creation of a stethoscope embracing the *Med En Cor* distinguishing acronym purposefully printed in simple but professionally suggestive characters (Figure 6.1)

Figure 6.1- *MedEnCo-Lex: the logo*

To complete the homepage graphic layout and to catch the potential users' attention on the purpose of the database, I chose a quotation by Sinclair (1991) whose meaning also summarizes the mission of this writing tool: "the language looks rather different when you look at a lot of it at once"(100).

The next step was to choose the homepage settings which I kept as relevant and user-friendly as possible: a central query box with five tabs below (Figure 6.2). Also, in view of the fact that, as target users are learners from the health sciences background, they may not necessarily be familiar with linguistic terminology

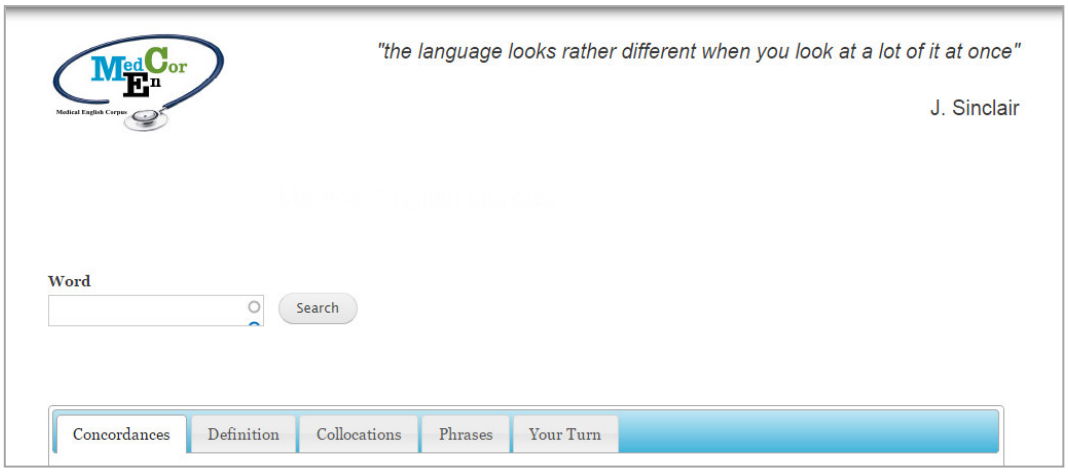


Figure 6.2- *MedEnCor-Lex: the homepage*

6.4.2 Creating keyword slots

In order to collect extra information about the keywords selected from the *MedEncor* corpus to be exploited as definitions and usage notes for the database, I used *WordSmith*'s *Concord* tool and presented each keyword in *KWIC* (*Key Word In Context*) concordances to be scrutinized vertically, horizontally and in depth. Reading the concordances from top to bottom and from left to right allowed me to look more closely at the environment surrounding the selected terms and to observe which recurrent patterns preceded or followed them. Expanding the context and moving back to source texts, instead, enabled me to get more information on the meaning of the node words and their usage. I tried to pick up as many clues as possible, making notes about all those prepositions, nouns, adjectives and verbs that recurrently co-occurred with the node terms.

Despite the wealth of details offered by the concordance lines, I reserved to verify my provisional results by consulting online monolingual dictionaries and thesauri, both medical (http://medical-dictionary.thefreedictionary.com/) and advanced (http://dictionary. cambridge.org/).

6.4.3 Arranging the results

While collecting data, I began to record my findings systematically in order to facilitate the categorization and retrieval of information. To be consistent with my terminology work, for each keyword entry I devised a standardized record sheet in electronic format whose modifiable structure allowed me to add supplementary slots whenever I found new information. In so doing the data was accessible, transparent and easily transferable to the database.

6.4.4. Database taxonomies and queries

I organized and indexed contents in the *MedEnCor-Lex* database according to specific taxonomies so that each document can be identified or retrieved by its acronym; genre or text-type; domain; year of publication; *uniform resource locator* (url); web-source and original *pdf*-file.

Queries in the *MedEnCor-Lex* database can be performed for forms, lemmas or grammatical categories (the results can be printed or copied to a file for later use). For any searched word, the software automatically returns all the existing occurrences, specifically highlighted and presented in concordance lines. Documents can be queried in block or individually and they can be freely selected. Thanks to the auto-complete

query field, if a term is a keyword, by simply typing the beginning of the word, the software directs the query to the keywords itself, allowing the opening of other five tabs: (1)Concordance; (2) Definition; (3); Collocation (4) Phrases; (5) Your turn (Figure 6.3)

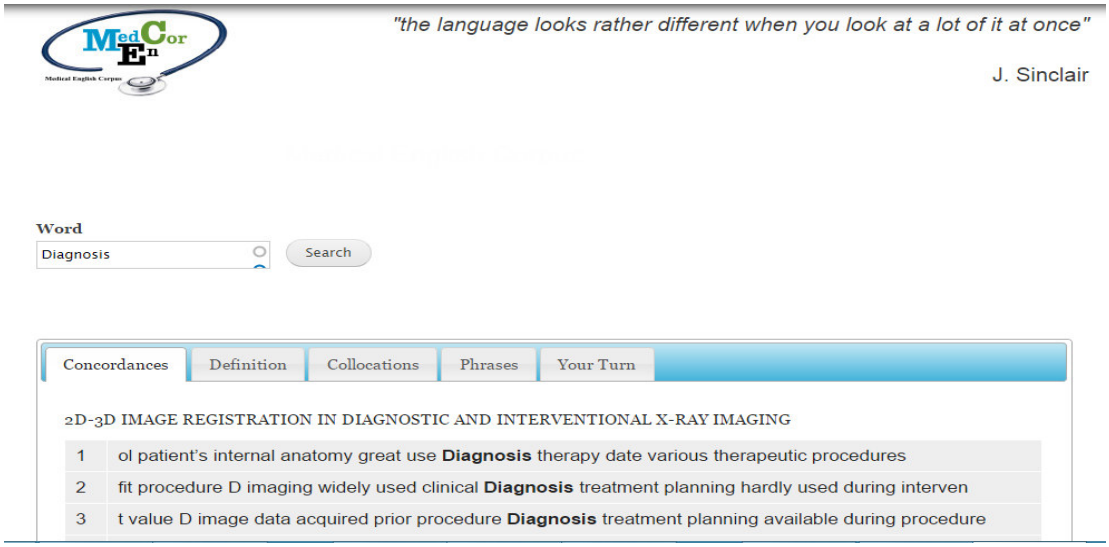


Figure 6.3. *MedEnCor-Lex tabs*

As stated above, the concordance tab (1) shows all the occurrences of the keyword, as shown in Figure 6.4, for the keyword 'diagnosis':

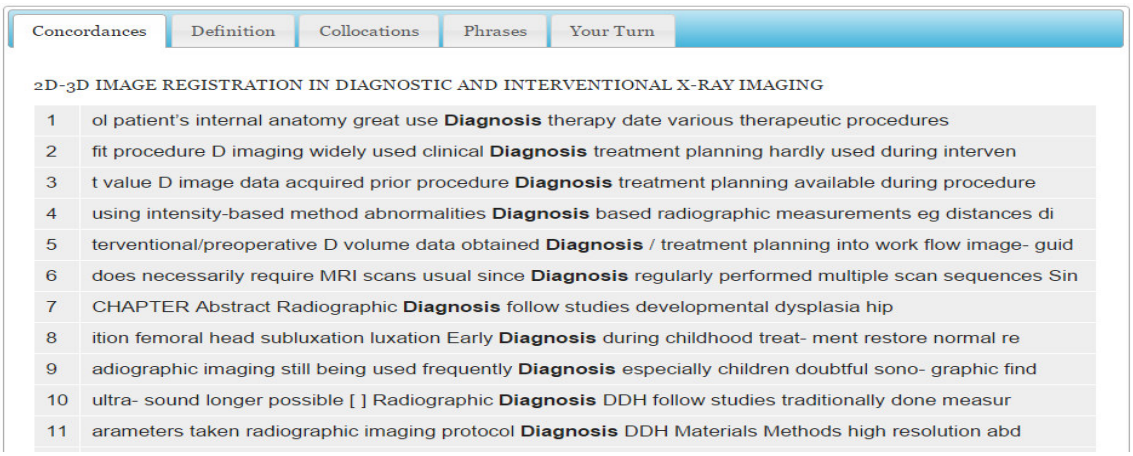


Figure 6.4. *KWIC, Key Word In Context*

The definition tab (2) displays meaning, grammatical category, synonyms, if any, and abbreviated forms (Figure 6.5):

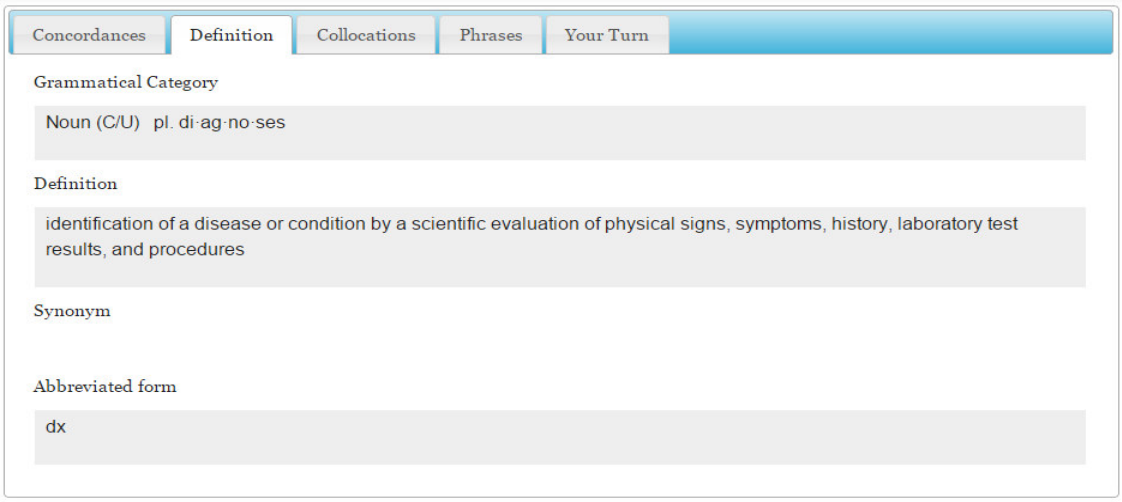


Figure 6.5. *Definition for the keyword'diagnosis'*

The collocation tab (3) shows all the grammatical and lexical words that match with the node word (Fig. 6.6)

Figure 6.6 *Examples of prepositions that collocate with 'diagnosis'*

The phrase tab (4) displays clusters of words containing the selected keyword (figure 6.7) :

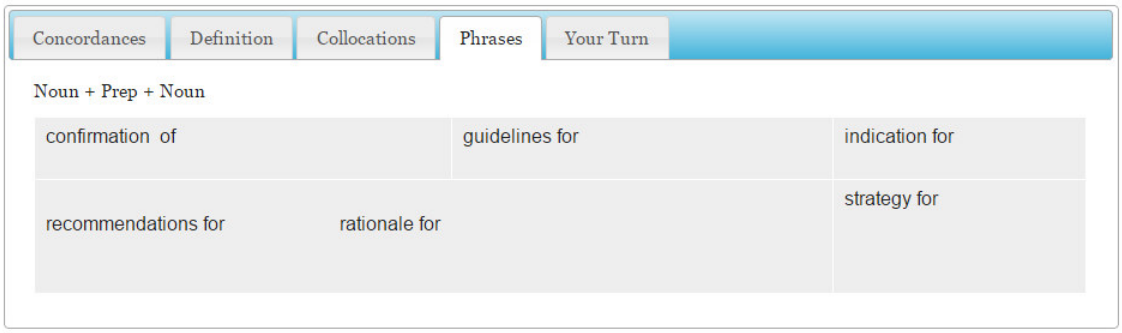


Figure 6.7. *Clusters of words containing 'diagnosis'*

Finally the fifth tab (5) labelled as 'Your Turn' proposes activities meant to quickly revise the usage of the keyword (Fig. 6.8) with multiple choice options.

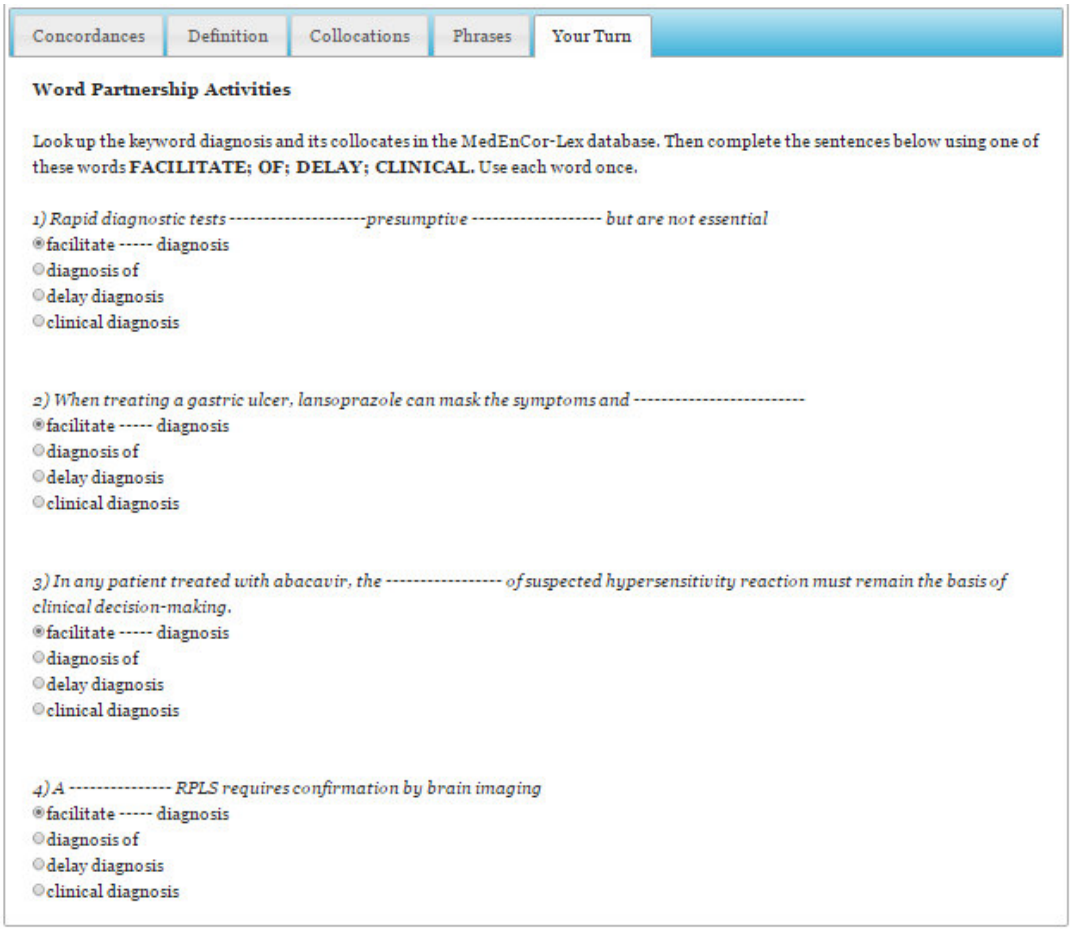


Figure 6.8. *Activities using 'diagnosis' in context*

6.5. Final considerations

Trying to find an effective and satisfactory answer to my second research question I devised the *MedEnCor-Lex* database as a didactic tool addressing non-native researchers' lexical needs and 'disorders'. At the moment only a limited amount of keywords has been scrutinized and inserted in the *MedEnCor-Lex* database which has provided me with just

a preliminary view of the wider research I mean to carry out, especially in view of a more effective pedagogical intervention.

I hope that when the *MedEnCor-Lex* is released, feed backs from medical users will prove my expectations.

Chapter 7
Conclusions

7.1 Conclusions

After evaluating how written Medical English constitutes a real challenge to non-native scientific researchers, the primary consideration of the present dissertation - clearly expressed in my first research question - has been to search for the *core lexis* of medicine, *i.e.* to define *what* vocabulary non-native medical graduates and undergraduates need in written professional communication. Aware of the fact that a knowledge of specialized terminology alone is not a sufficient condition for successfully coping with written medical discourse (Salager 1983, 54-55), my goal has been to extrapolate a selection of specialized keywords, analyze them in context and, by collating all the results in a web-based monolingual glossary (the MedEnCor-Lex database, available at www.medencor.com), try to overcome what constitutes a major obstacle to written communication for those involved in the scientific community.

A corpus-based approach has not only directed the compilation of a balanced and representative corpus, the *MedEnCor*, but has also suggested which specialized keywords to extrapolate. The analysis of

500 top-keywords has led to the identification of useful semantic groups and sub-groups fundamentally related to the pillars of medicine which are, and have always been: prevention, diagnosis, treatment. Concordance lines, instead, have allowed me to examine these words in real contexts and see what patterns of lexis, grammar and meaning surrounded them, yielding valuable insights into the structure and usage of medical language.

7.2 Future steps to be taken

At the moment only a limited amount of keywords has been scrutinized and inserted in the *MedEnCor-Lex* database which has provided me with just a preliminary view of the wider research I mean to carry out, especially in view of a more effective pedagogical intervention. I expect that once the *MedEnCor-Lex* is completed and released I will be in a better position to address my second research question and understand whether this writing aid is a useful pedagogical tool for the acquisition of the medical core lexis.

Since my research has been conducted as a health care 'outsider' I expect that the collaboration of subject field experts may lighten the burden during the future term selection process and help me identify the relationships between the selected terms, certainly with new and more valuable findings.

Finally I hope that by releasing the *MedEnCor-Lex* database, my data will prove constructive to those involved in the medical community and will also make the English used in medicine accessible to both health care professionals and medical undergraduates alike, who will thus familiarize themselves with terms and expressions relevant to their field of specialization.

My final consideration concerns more the approach that I have used rather than the results I have obtained: my interest in corpus linguistics as a vehicle to better understand language has blossomed with this research and with that, the conviction that I have used the language that medical students, junior clinicians, biomedical researchers and senior physicians all alike will encounter when they step outside the keywords of the database and step into the real world of language use.

Appendix 1

Case study

Abstract

ESL and EFL university students need to have good receptive and productive knowledge of general and academic English if they want to have access to the literature pertaining to their discipline and acquire the distinctive linguistic features of academic discourse (Hinkel, 2004; Paquot, 2010).

This case-study reports on an exploratory investigation carried out within the Department of English Studies at the University of Milan. The case-study, which is part of a PhD research project on medical lexis, was aimed at investigating the productive knowledge of vocabulary of a group of 100 medical students. The students were involved in two writing tasks as part of an English language test. A learner corpus of 200 written texts (60-80 words each) was compiled and data was analysed with reference to the General Service List (West, 1953) and the Academic Word List (Coxhead, 2000). Preliminary findings pointed to a limited size and range of general and academic English vocabulary which also resulted in inappropriate lexical choices. The corpus data was used to produce in-house EAP materials to foster autonomous learning and active participation in first-year medical students. Pedagogical applications of corpus work are discussed.

1 Background to the study

In Italy ESL instruction officially starts as early as in the first grade. At primary and middle school (Grade 1-8), ESL instruction covers various domains of language skills and aims to familiarize L2 learners with basic English sentence structures and the most commonly used words. During secondary school education (Grade 9-12) English is taught in academic and vocational strands and, following the launching of CLIL (Content and Language Integrated Learning) methodology[1], one widely implemented educational policy is that at least one subject area is taught in English[2]. In addition, thanks to the recent reforms, intermediate to advanced CEFR certifications (*Common European Framework of*

[1] CLIL is an umbrella term adopted by the European Network of Administrators, Researchers and Practitioners in the mid 1990s. It encompasses any activity in which a foreign language is used as a tool in the learning of a non-language subject in which both language and subject have a joint role. CLIL operates along a continuum of the foreign language and the non-language content without specifying the importance of one over another. CLIL is flexible and dynamic and gives both language and non-language subject matters a joint curricular role in the domain of mainstream education, pre-schooling and adult lifelong education. (Ranieri, 2013).

[2] The Italian Ministry of Education, in line with European Union policies and as in most European countries, has mandated CLIL as an approach to be adopted for teaching non-linguistic subjects in the last year of Italian secondary school and Italian technical high school by 2013. This policy moves Italian educators beyond traditional teacher-centered lecturing towards learner-centered ways of learning, promoting the proficiency of L2. (Di Martino & Di Sabato, 2012, 74-78).

Reference for Languages from B1 to C1 levels)[3] have been introduced in the national high school system so as to give ESL students the opportunity to achieve more versatile academic topics such as culture, business, science and technology. The study of English as first foreign language is also compulsory for most university students regardless of their discipline.

Despite the latest revisions of the national curricula and the ESL teachers' efforts to comply with the new European guidelines recommended by the Ministry of Education in matter of high school and university language teaching[4], it is a verifiable and established fact that Italian L2 freshmen and undergraduates still need to develop those academic writing skills expected to achieve degree programs. In particular what they need is to become relatively good at displaying academic knowledge within the genres, formats and vocabulary required in academic discourse. ESL learners' academic survival will depend on

[3] Ministry of Education, Universities and Research: Decrees of 12th July 2012; 28th Jan. 2013; 21st May 2013 and 14th July 2014 (Retrieved 14th Sept. 2014 from http://www.istruzione.it/allegati/2014/ ddg_prot5541_14.pdf).

[4] Council conclusions on language competences to enhance mobility (Council of the European Union, 2009; 2011); VET (*Vocational Education and Training*); Strategic Framework for European Cooperation in Education and Training (*ET* 2020). (*The Official Journal of the European Union*, *C 119*, 2-10).

their ability to construct passable written prose employing appropriate words common in the English-speaking environments and academies.

2 Aims of the study

This case-study aims to investigate the written academic discourse of a group of one hundred novice ESL students attending the Faculty of Medicine at the University of Palermo. Two are the driving forces behind the research: *1)* 'academic vocabulary' relating to

> that set of lexical items that are not core words but, unlike technical terms, are frequent in academic texts, regardless of the discipline […] sometimes used as a synonym for subtechnical vocabulary or discourse organizing vocabulary (Paquot, 2010, 9)

and *2)* 'academic writing' referring to the ability of organizing writing

> 'to convey major and supporting ideas […] demonstrating command of standard written English including grammar, phrasing, sentence structure, spelling, punctuation and a range of vocabulary appropriate for the topic (Hinkel, 2004, 18-19)

These two interconnected and interdependent forces are situated at the front line of EAP (English for Academic Purposes) a broad term defined as 'teaching English with the aim of assisting learners' academic communicative practice' (Flowerdew and Peacock, 2001, 8; Jordan, 1997, 1) and covering such areas as undergraduate and postgraduate teaching; research genres (from journal articles to conference papers and grant proposals); student writing (from essays to papers and graduate theses) and administrative practice (from course documents to doctoral forms) (Hyland, 2006, 1).

To accomplish these goals a trial English language test was anonymously administered to the sample group of non native undergraduates involved in the study in order to examine how the size and range of their general and academic English vocabulary might affect effectual academic writing. The data was collected in two learner corpora specifically compiled to analyse the lexical features and choices employed by the students. In addition, a post-test questionnaire was conducted gathering the candidates' feedbacks both to define the linguistic difficulties encountered by the students in completing the test

tasks and to understand the situation of English teaching and learning in the Italian ESL university classrooms.

Needless to say that examining the features of EFL writings and the students' problems in performing the assigned tasks would certainly be pedagogically beneficial, since the ultimate purpose and achievement of this study is to provide some adjustments and corrections, where necessary.

2.1 Research questions

Starting from the widely recognized assumption that unfamiliar vocabulary is one of the major barriers to comprehension and that learning the most frequent lexical terms is fundamental to successful communication either general or academic, with grammar judged as secondary to the enterprise (Read, 2000; Nation, 2001; Schmitt & Clapham, 2001; Cobb & Spada, 2001; Webb, 2005) three were the research questions which guided my study:

1. Are ESL undergraduates using an appropriate variety of vocabulary in their written works?
2. Is there an academic vocabulary distinctive for writing academic texts?
3. How large a vocabulary do Italian ESL undergraduates need to master to write academic texts (curricular/extracurricular)? Which words to focus on?

It should be made clear that the second research question is in no way meant to overshadow the importance of Coxhead's established *Academic Word List* whose 570 word families offer a wide coverage of academic texts. Rather, it aims at specifically identifying what are the academic words that non-native medical learners are likely -or not likely- to use in their academic texts.

3 Methodology

Primarily intended to provide a practical and comprehensive overview of the written outputs produced by the sample group of EFL learners

under investigation, this study mostly relied on computer learner corpus (CLC) research to find potentially useful insights worth examining. One of the main distinguishing features of computer learner corpora –and indeed one of their main strengths– is that they represent added value for EFL analysis. As explained by Flowerdew (2001), Granger (2002) and Paquot (2010) when educators need to ascertain the type of errors the learners make or the items they tend to under- or overuse, learner corpora are the most valuable resources for addressing the specific problems that non-native students encounter because, by using authentic texts and directly showing the context where the learners' main deficiencies lie, learner corpora may really help to improve students' writing skills.

3.1 Computer Learner Corpus

Computer learner corpus (CLC) research is a fairly recent phenomenon, as it started to emerge in the late 1990s. Despite its relative youth as a field of scientific enquiry, learner corpus study, however,

presents so many crucial advantages compared to the traditional procedures usually adopted in foreign language analysis that it seems worth recalling some of its remarkable benefits.

Learner corpora are collections of authentic texts produced by foreign/second language learners, stored in electronic format (Granger, 2004, 124). Learner corpora can be defined as 'systematic computerized collections of texts', where 'systematic' refers to the texts included in the corpus normally selected on the basis of 'some criteria (e.g. the learners' L2 proficiency level) representative of a certain learner group' (Nesselhauf, 2005, 40-41). Thus, one of the most important advantages of learner corpora is that by including real production data, they yield valuable information on what learners can actually deliver in a given situation. Moreover, unlike predefined elicitation tests (such as fill-in-the-blank tests or judgment tasks) predominantly used in EFL teaching environments, learner corpora can investigate more data; take better account of what learners actually want to express and also generate new hypotheses or valid statements on the learners' productive skills.

3.2 Compilation of the Learner Corpora *Med-A* and *Med-B*

Conceived as "a useful resource for anyone wanting to find out how people learn languages and how they can be helped to learn them better" (Leech, 1998, xiv-xx), the learner corpora *Med-A* and *Med-B* have been compiled in the hope that they will become a useful resource to understand the learning process of Italian medical EFL undergraduates, and in the hope that with corpus-based research findings, future EFL academic teachings will be tailored to the undergraduates' real learning needs.

3.3 Corpus design criteria

Med-A and *Med-B* are two small corpora collecting 100 handwritten texts each, with a total of 14,434 tokens and 1,452 types. The texts, all belonging to the e-mail genre, constitute the written session of a more comprehensive trial language test anonymously administered at the Faculty of Medicine at the University of Palermo. The participants did

not have any help of third parties throughout the writing process and were not allowed to make use of dictionaries or grammar books. The tasks aimed at testing the candidates' ability to produce two different e-mails (80-60 and 50-60 words each) in a fairly short period of time (40 minutes). The types of compositions were mainly descriptive, expository and argumentative and were related to contexts and situations easily encountered in academia, precisely: *1)* delaying the deadline of an assignment; *2)* obtaining information for a medical summer course (Fig. 1).

e-mail type A (60-80 words)	***e-mail type B (50-60 words)***
Write an email to your lecturer/professor in which you: ♦ introduce yourself ♦ explain what your project is about ♦ say why you were not able to complete the project last week ♦ ask for permission to hand in the completed project at a later date	*Write an e-mail to the Faculty of Medicine and Health Science of the University of Nottingham in which you:* ♦ indicate what subjects you are currently studying at your university; ♦ ask information about the subjects taught during the summer school; ♦ enquire about costs and special rules for non-UK students.

Figure. 1. *Two different timed writing tasks (40 min.) included in the test*

3.3.1 On small corpora

The choice of two collections of small corpora was mainly methodological: *1*) corpus browsing is user-friendlier; *2)* small corpora are put together quickly and, unlike large corpora which are designed for late human intervention, they allow early and instant intervention; *3)* the shrinking of the size has nothing to do with the quality of the data and the results achieved from their analysis.

3.3.2 Learners' profile

The participants, aged 18-19 years, were all non-native speakers in their first year at the Medicine & Surgery Faculty. Their English proficiency varied from elementary to advanced, with a majority of B1/B1+ intermediate level (CEFR) as declared in the feedback questionnaire kept as learners' supportive information (Table 1):

Table 1. *Learners' English language level and skills (CEFR)*

	Elementary A1-A2	*Intermediate B1/B1+*	*Advanced C1-C2*	*No answer*
Reading	11%	62%	21%	6%
Listening	35%	52%	6%	7%
Speaking	25%	62%	7%	6%
Writing	18%	68%	7%	7%

3.3.3 Data collection

After data collection, all the handwritten texts were converted into *text* machine readable format by means of the *Dragon dictation* application: a speech synthesizer which, relying on voice recognition, automatically transcribes sound messages of variable length running five times faster than regular typing on the keyboard.

While compiling *Med-A* (including 100 emails type A) and *Med-B* (including 100 emails type B) every effort was made to preserve the authenticity of the learning context and to ensure that the digital version of the emails matched with the participants' hand-written accomplishments. Grammatical mistakes were maintained because they were representative of the original writings. Spelling mistakes, instead,

were corrected to avoid problems with computer reading and data analysis and for wordlist conformity. Other transcriptional alterations included: proper names (which were omitted for anonymity); numbers and symbols (not considered as proper lexical units) and overused repetitions (if they were mere copies or transcriptions of the assigned prompts).

4 Data analysis and discussion

Analysis was first performed with *WordSmith Tools 6.0* (Scott, 2014) whose functions like the components of a Swiss army penknife helped to cut out the *Med-A* and *Med-B* corpora into wordlists, to chop frequencies and to uncork concordance lines (Scott, 2001, 47). In the second stage, *Range* software was used: a lexical profiler designed by Nation, Hatley and Coxhead (2002) able to group word-families into four base-word lists created by splitting the GSL (West's *General Service List)* into the first thousand (Base 1) and the second thousand (Base 2) commonly used

words, adding the AWL (Coxhead's *Academic Word List*) as the third (Base 3) and technical words as the fourth band (Base 4) (Figure 2):

BASEWORD 1	**includes the most frequent 1000 words of English**
BASEWORD 2	includes the second 1000 most frequent words of English
BASEWORD 3	comprises words frequent in university texts from a wide range of subjects
NOT IN THE LIST-4	contains (highly) technical words

Figure. 2. *Base word lists in Range.exe software*

The core academic vocabulary contained in the *Academic Word List* has been identified by Coxhead's *AWL Highlighter*, a program that returns words arranged by frequency in ten sublists.

4.1 Quantitative analysis

A major advantage of the corpus approach lies in the usefulness for conducting quantitative analysis. The quantitative features of a corpus provide a basic but global view of the characteristics of the learners'

writings. Thus, the following findings depict the characteristics of the *Med-A* and *Med-B* as learner corpora.

After running the *WordList* tool from *WordSmith,* a list of the most frequent words in the corpora was obtained. The list allowed the identification of more than 800 *hapax legomena* (words or forms occurring only once in a corpus) which, given their low frequency were excluded from the analysis. Of the remaining words, the first eight most frequent words were function words (Table 2) which held little or no meaning but were necessary to the grammatical structures. It was decided to exclude these words incorporating them in a special 'stop list'.

Table 2. *First nine most frequent words*

N.	***Word***	***Freq.***	***%***
1	*I*	756	5,24
2	*the*	647	4,48
3	*to*	502	3,48
4	*and*	483	3,35
5	*of*	458	3,17
6	*a*	325	2,25
7	*in*	302	2,09
8	*am*	275	1,91

The remaining list of 403 tokens was analysed with *Range* tool and the results are illustrated in Table 3:

Table 3. *Distribution of the 403 most frequent tokens after running Range.exe*

	Families	*Types*	*Tokens*	*Percent*
Base 1 (1-1000):	228	282	**295**	**73.20%**
Function:	...	...	(72)	(17.87%)
Content:	...	...	(223)	(55.33%)
> Anglo-Sax. =Not Greco-Lat./Fr. Cog.	...	...	(114)	(28.29%)
Base 2 (1001-2000):	29	36	**36**	**8.93%**
>Anglo-Sax.	...	...	(16)	(3.97%)
1k+2k		...	...	(82.13%)
Base 3 (academic words):	20	24	**25**	**6.20%**
> Anglo-Sax.	...	...	(1)	(0.25%)
Not in the list:	?	46	**47**	**11.66%**
	277+?	388	403	100%

What struck most about this data was that the majority of the words (73.20 %) used by the candidates extensively belonged to the first 1000 most frequent words thus confirming that they mastered only half of the expected 2000 words from West's *General Service List.* Conversely, only a minority (about 9%) was comprised in *Base 2* group with an even

smaller percentage (6.20%) included in *Base 3* which contains Coxhead's *AWL* word families, as visualized in the pie chart in Figure 3.

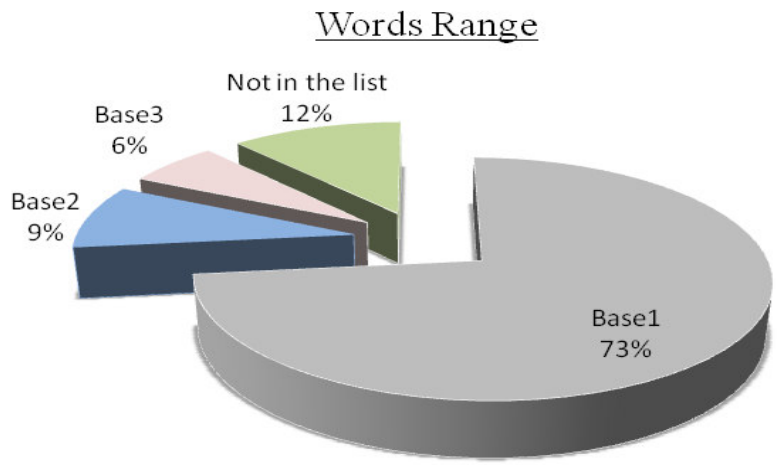


Figure. 3. *Percentage of the words included in Med-A and Med-B corpora distributed across the Range-exe Base groups -*

Another surprising discovery was the 12% of highly technical or low frequency words numbered in the fourth group (*Not in the list*). Excluding geographical names (*Italian, Nottingham, Palermo, Sicily, UK)* and e-mail salutations or introductions (*dean, email, exam, faculty, informatics, laboratory, maths, professor, undergraduate)* which were required to compile the assigned tasks, a closer analysis of the remaining highly technical words *(anatomy, biology, bowel, cancer, casualty, cells,*

chemical, chemistry, Crohn, diffuse, embryology, physician, genetic, HIV, irritable, histology, oncology, patients, physiology, surgery, syndrome) showed that they mostly belonged to medical terminology and were therefore easily learnt and appropriately used by the candidates given their monosemic nature; their Latin or Greek origin and their pertinence to the specialized disciplines relevant to the Italian students' medical studies.

Activating the *AWL Highlighter* software, the 25 types belonging to the AWL group (*Base 3*): *affected, aid, analysis, computer, consequences, data, finally, incident, involved, link, linked, medic, methodology, methods, option, period, positive, previously, project, research, researching, response, statistic statistics* were redistributed across seven sublists which highlighted a rather low level in vocabulary diversification (Table 4).

Table 4. *Distribution of the 25 AWL words across Coxhead's 7 sublists*

Sublist 1	**Sublist 2**	**Sublist 3**	**Sublist 4**	**Sublist 5**	**Sublist 6**	**Sublist 7**
analysis, data involved, methodology methods, period, research, researching, response	*affected, computer, consequence finally, positive, previously*	*link, linked*	*options, project, statistic, statistics*	*medic, medical*	*incident*	*aid*

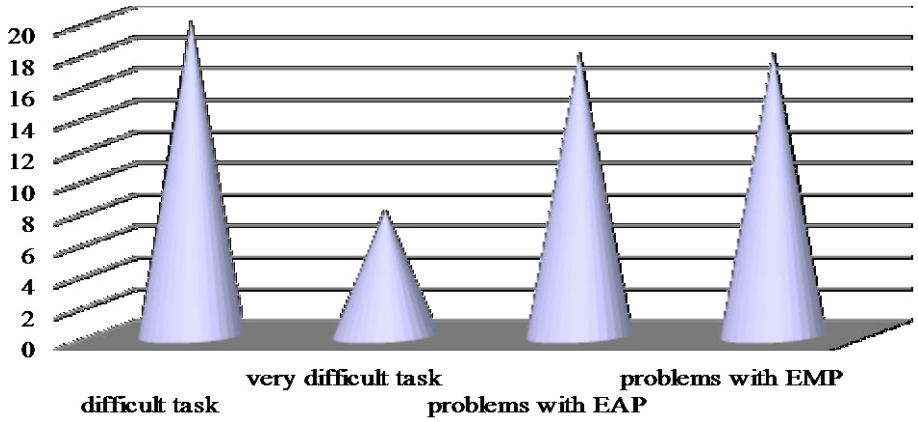
These findings were also reflected in the type/token ratio (0.96) and in the content word/token ratio (0.82) -also called lexical density- both indicative of a limited lexical range.

4.2 Feedback questionnaire

An external confirmation of the reduced vocabulary load mastered by the participants was also provided by the follow-up questionnaire compiled by the candidates to gain feed-back into the problems encountered during the test: 20% of the students rated the writing tasks as "difficult", with a peak of 8% who rated them as "very difficult".

Among the reported difficulties: 18% lamented “problems with EAP and EMP word choice” or felt handicapped by “a less rich vocabulary” and “less facility in expression” (Table 5)

Table 5. *Language difficulties encountered by the participants*



4.3 Qualitative analysis

Analysing the content words (154+ 69) included in *Base 1* and *Base 2* respectively it was noticed that only 12 words out of 223 were adverbs: *actually, already, also, just, likely, now, soon, too*, *again, currently, faithfully, gently.* A deeper investigation by means of *WordSmith’s Concord* function across the co-text surrounding these words showed

that they were mainly used as ‘circumstantial adverbs’, *i.e* adding some kind of information (manner, place, time, frequency, or other circumstances) to the verb or verb phrase expressed in the main clause. They were not used as ‘connecting adverbs’ *i.e*. adverbs that serve to connect two stretches of discourse showing cause and effect, sequence, contrast, comparison or other relationships (Culpeper, J., Katamba, F., Kerswill, P., Wodak, R. & McEnery T., 2009, 121-123) which, instead, help smoothness and cohesion, especially in academic writing.

Consistently, the scarcity of linking devices and discourse markers was ample evidence of lack of command of formal academic writing, resulting in unbalanced and poorly coherent production.

4.3.1 Odd register

A careful analysis of the abundance of exclamation marks “*!*” (241) and the overuse of informal words connected to salutations and greetings: *hey* (21); *hi* (22); *good morning* (46) (mostly used to address university professors or faculty lecturers, as required by the test prompts)

suggested that the participants –against any academic formality or 'netiquette'[5]– wrote in a way which was more similar to the spoken interaction, the register that contains the highest occurrences of these items. Interestingly, even attempts to formality failed, as it is shown by the high occurrences of the personal pronoun "him" (11) improperly used instead of the corresponding "you" which is to be preferred when addressing interlocutors or recipients (L1 transfer).

5 Pedagogical implications

Consistent with the aims set down at the beginning of this case-study, I decided to turn the linguistic data resulting from the analysis of *Med-A* and *Med-B* learner corpora into effective teaching materials, thus to

[5] The term *netiquette* is the abbreviation for 'network etiquette'. It is derived from two French words: '*net*' meaning 'bucket' and '*iquette*' meaning 'of doilies' and it refers to the online good manners that should be kept in virtual correspondence by the Internet users, especially academicians. Appearing to be badly behaved in emails and other forms of online communication may offend recipients or damage reputation, therefore 'avoid using exclamation marks as far as possible […] Too many exclamation marks are not acceptable in formal conversation. They make the email seem casual and the exclamation is believed to be only an exaggeration or a pure scolding'. (Pinge, D., 2007, *Netiquette: What not to do online.* Retrieved 16 th Sept. 2014 from http://www.rediff.com/getahead/2007/jul/10netiquette.htm).

provide pedagogical solutions to the lexical impairments of the future healthcare professionals. Hence, I moved the research from a corpus-based approach to a genre-based approach, specifically focusing on e-mail as a genre.

5.1 E-mail as a genre

Following Swales (1990, 58) who identified a genre as a 'cultural and interpersonal event related to a social purpose, making use of language with a recognizable form and structure […] a proper register of language associated to it' and Bathia (1993, 16) and Miller (1984, 158) who stated that learning a genre implies learning how to participate in the actions of a community, I recognized e-mails as fundamental genres in the candidates' academic life[6] and therefore, learning e-mail patterns, moves, register and cohesive devices became a priority. Accordingly, firstly I identified the academic words (from AWL sub-lists) significant

[6] Much of the academic life is through academic correspondence: student/university administration, student/supervisor(s), student/university staff (Wallwork, 2011)

for each move of the genre; secondly, I tried to recycle (mistakes included) the vocabulary already used by the students and, finally, I devised vocabulary learning activities to help English language undergraduates carry out everyday correspondence, *i.e*. write different e-mail types (request, reply or application e-mails); adopt the right level of formality; use standard words and phrases appropriately (Figure 4 and Figure 5).

	Academic e-mail writing: 10 useful tips
01	adopt the right level of formality: deemphasize the conversational level
02	be synthetic: keep sentences short
03	limit the number of clauses
04	be specific, never vague
05	avoid words that add no value for the reader
06	use modals to soften claims
07	use link words (in longer mails) to show connections
08	check your spelling and grammar
09	mind punctuation and smileys (only if recipient used, first)
10	be careful how you use pronouns

Figure.4. *Tips to follow to write effective academic e-mails*

Academic e-mail writing: *gene*ric conventions	
SUBJECT LINE	**DO:** always specify purpose and context **DON'T:** leave the box empty
TO	**DO:** always check your recipient's email address **DON'T: '***to whom it may concern'* ***use CC** (*Carbon Copy*), if you want to send your email to two (or more) different people. ***use BCC** (*Blind Carbon Copy*), if you do not want the original recipient to know that you are sending an email to someone else**.**
SALUTATIONS	**DO:** Dear Miss/Prof Brown **DON'T:** Dear Teacher**,** Teacher, Hi! Hey!
OPENINGS (first sentence) (next sentence)	**DO:** always announce who you are **DON'T:** use your nick-name **DO:** explain why you are writing; state problems; use modals to soften claims e.g. *Could you please .....?* **DON'T:** use imperatives
CLOSING	**DO:** use a sign off word before you sign your name. Always sign your name. You can use your first name only, or you can use both/all of your names. e.g. (very formal): *Sincerely*, e.g (regular): *Thank you*, e.g (a bit informal): *Have a nice day/evening/etc*., e.g (a bit informal): *Thanks* **DON'T:** use your nickname or any words about love e.g (too informal): *Love*, *Much love*, e.g (too personal) *Lots of love*
SIGNATURE	**DO:** Always sign your name. You can use your first name only, or you can use both/all of your names. e.g. *James*, or *James Brown* **DON'T:** nickname or no name

Figure 5. *Tips related to e-mail writing as a 'genre'*

6. Conclusions

This case-study has shown how learner corpus-based research and foreign language teaching are closely interconnected in that corpus evidence suggests which language items and processes are most likely to be encountered by language users (what is frequent and typical) and may thus deserve more time in classroom instruction. Accordingly, *Med-A* and *Med-B* learner corpora have turned out to be valuable resources because they have helped *a)* identify typical language difficulties (lack of appropriate variety of vocabulary in written academic works); *b)* provide a correction to the frequently occurring mistakes in learner language and use (academic e-mail writing); *c)* of a certain learner group (medical undergraduates); *d)* of a certain native language (Italian non-native speaker of English) whose academic survival will much depend on the ability to construct satisfactory prose of adequate quality.

Appendix 2

List of 500 most relevant keywords

PATIENTS	STUDY	BASELINE	COMPLICATIONS
DIAGNOSIS	CELLS	CHRONIC	RESULTS
CLINICAL	LENS	LESIONS	SKIN
TREATMENT	SURGERY	SYMPTOMS	EYE
PRE	DOSE	CORNEA	FETAL
DISEASE	ACUTE	RETINAL	CONJUNCTIVA
PATIENT	SURGICAL	ARTERY	INFLAMMATORY
ANESTHESIA	CATARACT	INTERVENTION	DUE
BLOOD	GLAUCOMA	OPTIC	SYSTEMIC
ANTERIOR	TISSUE	STUDIES	VASCULAR
NON	MYOCARDIAL	DIAGNOSIS	DRUG
RISK	POSTERIOR	OCULAR	SYNDROME
ASSOCIATED	INFECTION	RENAL	IMAGING
CARDIAC	MAY	PERFORMED	CANCER
CORNEAL	MUSCLE	DATA	UVEITIS
THERAPY	EFFECTS	PAIN	PERFUSION
CORONARY	CELL	PRESSURE	CONGENITAL

ANTI	FLUID	DECREASED	SERUM
OUTCOMES	HYPERTENSION	INTRAOCULAR	ORAL
NUTRITION	MEDICAL	TRAUMA	USED
DRUGS	PLASMA	EXAMINATION	AGE
NURSE	PULMONARY	OCCUR	RANDOMIZED
ANAESTHESIA	PROTEIN	CILIARY	CHOLESTEROL
EPIDURAL	TABLE	INFLAMMATION	FIGURE
PARTICIPANTS	PRESCRIBING	RELATED	FUNCTION
DISEASES	SEVERE	CAUSES	CONJUNCTIVAL
PROGRAM	USING	MORTALITY	FACTORS
DISORDERS	CARDIOVASCULAR	MIN	LOWER
COHORT	VITAMIN	ARTHRITIS	ADVERSE
CONJUNCTIVITIS	VITREOUS	METHOTREXATE	CARE
PREGNANCY	ANALYSIS	WITH	PELVIC
INCREASED	RETINA	ANGIOGRAPHY	NORMAL
SIGNIFICANT	LEVELS	PRIMARY	RUPTURE
VS	CATHETER	BREAST	IRIS
HEART	RESPIRATORY	COMPARED	PHARMACIST
OCCURS	INCLUDE	INFARCTION	THYROID
NERVE	INFECTIONS	PERIPHERAL	INTERVENTIONS
DIABETES	EVALUATION	ACID	DECREASE

LACRIMAL	VOLUME	OXIDATIVE	CARCINOMA
INDUCED	LATERAL	BLINDNESS	MRI
EXERCISE	METABOLISM	INCISION	MEMBRANE
VISION	MEDICINE	POST	LIVER
GUIDELINES	RECOMMENDED	ITCHING	MATERNAL
TEST	ABDOMINAL	ULCER	GUIDELINE
ORBITAL	DIAGNOSTIC	FLOW	META
ACTIVITY	AQUEOUS	MEDICINES	MULTIPLE
BLEEDING	ARTERIAL	GLUCOSE	VAGINAL
VISUAL	ANKLE	OEDEMA	DYSFUNCTION
INCIDENCE	INVASIVE	EPITHELIUM	HEALTHY
FREQUENCY	ULTRASOUND	TOPICAL	OPHTHALMOLOGY
DURING	FOLLOWING	FINDINGS	GROUP
SUPPLEMENTATION	ETIOLOGY	CASES	NASAL
LOW	VENTRICULAR	UTERINE	SYSTOLIC
MEASURES	EYEBALL	REPORTED	MED
ISCHEMIA	INJURY	DNA	MACULAR
BONE	COMMON	STENOSIS	ELEVATED
POSTOPERATIVE	RETINOPATHY	ANGLE	HEALTH
LESION	TUMOR	VENOUS	AXIAL
INFANTS	USUALLY	RATE	OUTCOME

DEPRESSION	MILD	ONSET	URINARY
JOURNAL	VESSELS	HIGH	HYPOTENSION
MEDICATIONS	DEFICIENCY	PUPILLARY	COMMONLY
IMPLANT	MYOCARDIUM	TUMOUR	INSULIN
CAUSE	ANGINA	CERVICAL	OPHTHALMIC
TUMORS	LID	ANTIBIOTICS	CHAMBER
DELIVERY	FOLLOW	DIABETIC	EPITHELIAL
INTENSITY	RESPONSE	CHANGES	SPECIFIC
SCLERA	HYPERTENSIVE	SWELLING	REFRACTIVE
TRACT	ANALGESIA	CATHETERS	INFUSION
PROGRAMS	MORBIDITY	ECG	OCCLUSION
TUMOURS	SIGNIFICANTLY	ISCHEMIC	PROPHYLAXIS
DOSES	DIETARY	INCREASE	TOMOGRAPHY
ASSESSMENT	RECEPTOR	ANESTHETIC	RECTUS
MICE	EVIDENCE	GENE	MYOPIA
URINE	METABOLIC	ATROPHY	ANTIBIOTIC
PREVENTION	FUNDUS	SECONDARY	MALIGNANT
TABLETS	OBSERVED	DURATION	ANESTHETICS
PRACTICE	PATHWAY	MEASURED	SHOULD
BASED	DIASTOLIC	THORACIC	RATIO
PUMP	PRACTICES	LIPID	SCORE

RHEUMATOID	VALIDATION	MEASUREMENT	INDICATED
OBSTRUCTION	ANALYSES	USE	ATOPIC
ADMINISTERED	TECHNIQUE	CENTER	REDUCTION
MEDICATION	REGRESSION	HAEMORRHAGE	MODERATE
EXPOSURE	EFFECT	BACTERIAL	METHODS
CONDITION	PHYSICAL	INTRAVENOUS	NEONATAL
CRITERIA	POSITIVE	TRIAL	PRESENCE
THERAPEUTIC	APPROXIMATELY	INHIBITORS	PROCEDURES
CONCENTRATION	MAGNETIC	AIRWAY	BETA
ABNORMAL	SIGNS	COHORTS	OBESITY
SELF	STRESS	VAN	CHARACTERISED
NURSING	HIV	WEEKS	HERPES
MITOCHONDRIAL	ABNORMALITY	PREVALENCE	TRIALS
ATTENUATION	TISSUES	NECROSIS	MMHG
IMMUNE	GENETIC	CLINICALLY	RAY
CELLULAR	AORTIC	IMPLANTS	BLOCKERS
PHASE	PARAMETERS	DEFECTS	CONTROLLED
COMPUTED	RECURRENT	DEGENERATION	SEVERITY
THROMBOSIS	OXYGEN	FRACTURES	STEROIDS
PHYSIOLOGICAL	ALLERGIC	TESTING	STAINING
INTAKE	TYPE	EFFICACY	BIOPSY

EVALUATED
REFLEX
VOMITING
INJECTION
DERMATITIS
MEDIAL
VIRAL
MODEL
THESIS
OINTMENT
TOTAL
DEVICE

REVIEW
ESTROGEN
EXTRAOCULAR
ABSCESS
BRAIN
PERFORATION
KIDNEY
BILATERAL
TERM
SODIUM
VALUES
CLOPIDOGREL

RESPECTIVELY
DETACHMENT
VERSUS
FAILURE
PROGNOSTIC
FUNCTIONAL
REDUCTASE
PSYCHOSOCIAL
CHEMOTHERAPY
STENT
PEDIATRIC
LIMBUS
FACTOR

TREATED
DEMONSTRATD
ACTIVATION
LIDS
TESTS
PROCEDURE
NAUSEA
PROTEINS
CHOROID

References

Adams Smith, D. E. (1984). Medical discourse: Aspects of author's comment. *The ESP Journal*, *3*(1), 25-36.

Allison, D., Corcos, R., and Lam, A. (1994). Laying down the law? Reflecting on course design in progress. Hong Kong Papers. *Linguistics and language teaching*, *17*, 1-11.

Allum, V., Wright, R., & McCullagh M. (2013, 2nd-6th-April). *Introduction to teaching English for Medical Purposes.* Lectures presented at the EMP Teacher Training Conference, Exeter University, Exeter, UK. Retrieved 14th Dec. 2013, from http://www.e4nursingandhealthcare.com/uploads/9/9/1/8/9918627/itempe xeter 2013.pdf

Allwright, J. & Allwright, R. (1977). An approach to the teaching of medical English. In S. Holden (ed.), *English for specific purposes*, 58-62. London: Modern English Publications

Ainsworth-Vaughn, N. (2001). The discourse of medical encounters. In Schiffrin, D., Tannen, D., & Hamilton, H. E., *The handbook of discourse analysis*. Malden, Mass.: Blackwell Publishers, 453-469.

Anderson, W., & Corbett, J. (2009). *Exploring English with online corpora: An introduction*. Basingstoke, England: Palgrave Macmillan.

Atkins, S., Levi, B. & Zampolli, A.(1992). Corpus design criteria. *Literary and linguistic computing*, *7*(1), 1-16.

Atkinson, D. (1992). The evolution of medical research writing from 1735 to 1985: the case of the Edinburgh Medical Journal. *Applied Linguistics*, *13*(4), 337–74.

Baethge, C. (2008). The Languages of Medicine. *Dtsch Arztebl International, 105*(3), 37–40. Retrieved 14th Aug. 2013 from http://www.ncbi.nlm.nih. gov/pmc/articles/PMC2696676/ pdf/Dtsch_Arztebl_Int-105-0037.pdf

Baker, P. (2006). *Using corpora in discourse analysis*. London: Continuum.

Basturkmen, H. (2006). *Ideas and options in English for specific purposes*. Mahwah, New Jersey: Lawrence Erlbaum Associates.

Bauer, L. & Nation, P.(1993). Word families. *International Journal of Lexicography*, *6*(4) 253–279.

Bhatia, V. K. (1993). *Analysing genre. Language use in professional settings*. London, UK: Longman.

Belcher, D. D. (ed). (2009). What ESP is and can be: an introduction. *English for specific purposes in theory and practice*. Ann Arbor: University of Michigan Press, 1-14.

Benfield, J. R. and Howard, K. M. (2000). The language of science. *European Journal of Cardiothoracic Surgery, 18, 642–8.* Retrieved 26th August 2012 from http://ejcts. oxfordjournals.org/content/18/6/642.full.pdf+html

Benfield, J. R. & C. B. Feak (2006). How authors can cope with the burden of English as an international language, *CHEST*, *129*, 1728–1730. Retrieved 26th August 2012 from http://www.chestjournal.org/content/129/6/1728.full.html

Bergammer, B. (2008). English as the lingua franca of science: A translator's view on what's lost–and what's gained–in translation. *The Write Stuff* , *17*(4), 213-217.

Berruto, G. (1980). *La variabilità sociale della lingua*. Torino: Loescher.

Berwick, R. (1989). Needs assessment in language programming: from theory to practice. In: Johnson, R. K. (ed). *The second language curriculum* , 48-62. Cambridge: Cambridge University Press .

Biber, D., (1993) Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 244-257. Retrieved 16th Sept. 2012 from http://otipl.philol. msu.ru/media/biber930.pdf

Biber, D., Conrad, S. & Rippen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Birch-Bécaas, S. (1994). Writing scientific articles in English: Solutions for French researchers?. *ASp, (*Actes de l'atelier Langue de spécialité du 33e congrès de la SAES), *1994*(3), 2-6. Retrieved 28th Sept. 2014 from file:///C:/Users/francesca67/ Desktop/asp-4234-3-writing-scientific-articles-in-english-solutions-for-french-researchers.pdf

Björkman, B. (2013). *English as an academic lingua franca. an investigation of form and communicative effectiveness*. Berlin, Boston: De Gruyter Mouton, 6-13.

Bondi, M & Scott, M. (2010). *Keyness in texts: Corpus Linguistic investigations*. Amsterdam: Benjamins.

Bosher, S. (2013). English for Nursing. In B. Paltridge & S. Starfield (eds), *The handbook of English for Specific Purposes*. Chichester: Wiley-Blackwell, 263-282.

Bowker, L., & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora*. London: Routledge.

Brindley, G. (1989). The role of needs analysis in adult ESL program design. In: Johnson, R .K. (ed). *The second language curriculum,* 63-78. Cambridge: Cambridge University Press.

Brown, J. D. (2009).Foreign and second language needs analysis. In Long M. H. &. Doughty C. J. (eds), *The Handbook of Language Teaching.* Chichester, U.K.: Wiley-Blackwell, 269-293.

Brumfit, C. (1995) 'Teacher professionalism and research', in Cook, G. & Seidlhofer, B. (eds.) *Principles and practice in Applied Linguistics*. Oxford: Oxford University Press, 27-42.

Burns, A. (2012). *The Cambridge guide to pedagogy and practice in second language teaching*. New York: Cambridge University Press.

Candlin N.C., Leather J.H. & Bruton C.J. (1976). Doctors in casualty: applying communicative competence to components of specialist course design. *IRAL 14*(3), 245-272.

Carter-Thomas, S., Rowley-Jolivet, E. (2008). If-conditionals in medical discourse: from theory to disciplinary practice. *Journal of English for Academic Purposes,7*(3), 191–205.

Chen, Q., Ge, G. C.(2007). A corpus-based lexical study on frequency and distribution of Coxhead's AWL word families in medical research articles (RAs). *English for Specific Purposes*, *26*(4), 502-514.

Chung, T. M., Nation, P. (2003). Technical vocabulary in specialised texts". *Reading in a Foreign Language, 15*(2), 103-116. Retrieved 16th June 2012 from http://nflrc.hawaii.edu/rfl/October2003/ chung/chung.html

Cobb, T. (1997): Is there any measurable learning from hands-on concordancing? *System*, *25*(3), 301-315.

Cobb, T. (1999). Vocabulary sizes of some City University students. *Journal of the Division of Language Studies of City University of Hong Kong, 1*(1), 59-68.

Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., Ferrara, J. and Fine, J. (1979). Reading English for Specialized Purposes: discourse analysis and the use of student informants. *TESOL Quarterly, 13*(4), 551-564.

Conrad, B. (1999). The importance of corpus-based research for language teachers. *System, 27*(1), 1–18.

Cook, V. (2004). What is Applied Linguistics?. In *Starting Applied Linguistics research.* University of Newcastle upon Tyne. Retrieved 14th Dec. 2012, from http://homepage.ntlworld.com/vivian.cWritings/Papers/WhatIsAppliedLinguistics .htm

Cooter, M. (2000). News notes. *European Science Editions, 26*(2), 67.

Corbett, J., Lu, P. (2012). *English in medical education an intercultural approach to teaching language and values.* Bristol: Multilingual Matters.

Coxhead, A. (2000) A new Academic Word List. *TESOL Quarterly*, *34*(2), 213-238.

Crystal, D. (1980) *A first dictionary of linguistics and phonetics*. London: André Deutsch.

Crystal, D. (1995). *The Cambridge encyclopedia of the English language*. Cambridge: Cambridge University Press, 367-375.

Crystal, D. (2003). *English as a global language.* (Second edition). Cambridge: Cambridge University Press, 11-24.

Culpeper, J., Katamba, F., Kerswill, P., Wodak, R. & McEnery T.(eds) (2009). *English Language: description, variation and context.* Houndmills, Basingstoke: Palgrave Macmillan.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualis, 19*, 121-129.

Cummins, J. (1984*). Wanted: a theoretical framework for relating language proficiency to academic achievement among bilingual students*. from The Ontario Institute for Studies in Education. Retrieved 9th Jul. 2012 from http://dyna2.nc.hcc.edu.tw/dyna/data/user/hs1283/ files/201204092037390.pdf

Davies, A. (2007). *An introduction to Applied Linguistics from practice to theory* (2nd ed.). Edinburgh: Edinburgh University Press.

Di Martino, E., & Di Sabato, B. (2012). CLIL implementation in Italian schools: Can long-serving teachers be retrained effectively? The Italian protagonists' voice. *Latin American Journal of Content and Language Integrated Learning*, *5*(2), 73-105. Retrieved 29th January 2014 from http://www.unifg.it/sites/default/files/allegatiparagrafo/20-01-2014/di_martino_di_sabato_clil_implementation_in_italian_schools.pdf

Dudley-Evans, T. & John, M. (1998). *Developments in English for Specific Purposes: A multi-disciplinary approach*. Cambridge: Cambridge University Press.

EL-Sakran, A. T.(2013). A Review of Problems Arab Students Encounter in Academic Writing. *English for Specific Purposes World*, *14*(38). Retrieved 4th Dec. 2013 from http://www.esp-world.info/Articles_38/Abstracts/Ayah_ Problems_Arab_Students_Encounter_in_Academic%20Writing.htm

Farrell, P. (1990). Vocabulary in ESP: a lexical analysis of the English of electronics and a study of semi-technical vocabulary. *CLCS Occasional Paper*, 25. Dublin: Trinity College.

Fatah-ELrahman Dafa-Allah .A.M., (2012). ESP learners' needs: A case study of medicine students at some Sudanese universities. *English for Specific Purposes World*, *36*(12). Retrieved 13th Sept. 2013 from http://www.esp-world.info/Articles_36/Daffa-Allah.pdf

Ferguson, G. (2001). If you pop over there: a corpus-based study of conditionals in medical discourse. *English for Specific Purposes, 20*(1), 61-82.

Ferguson, G.. (2013). English for medical purposes. In B. Paltridge & S. Starfield (eds), *The handbook of English for Specific Purposes*. Chichester: Wiley-Blackwell, 243-261.

Ferguson, G., Perez-Llantada, C., Plo, R. (2011). English as an international language of scientific publication: a study of attitudes. *World Englishes*, *30*(1), 41–59.

Finney, D. (2002). The ELT curriculum: A flexible model for a changingworld. In: Richards, J. C. & Renandya, W. A. (eds). *Methodology in language teaching : An anthology of current practice,* 69-79. Cambridge: Cambridge University Press.

Fleischman, S. (2001). Language and Medicine. In Schiffrin, D., Tannen, D., & Hamilton, H. E., *The handbook of discourse analysis*. Malden, Mass.: Blackwell Publishers, 470-502.

Fløttum, K., Dahl, T., Didriksen A. A., & Müller Gjesdal, A. (2013). KIAP – reflections on a complex corpus. *The many facets of corpus linguistics in Bergen -*

in honour of Knut Hofland Bergen Language and Linguistic Studies (BeLLS), *3*(1), 137-150. Retrieved 10th March 2014 from https://bells.uib.no/bells/article/view/367/381

Flowerdew, J. (2000). Discourse community, legitimate peripheral participation, and the nonnative-English-speaking scholar. *TESOL Quarterly, 34*(1), 127–50.

Flowerdew, J. (2001) Attitudes of journal editors to nonnative speaker contributions. *TESOL Quarterly, 35*(1), 121–50.

Flowerdew, J. and Peacock, M. (2001). *Research perspectives on English for academic purposes*. Cambridge: Cambridge University Press

Flowerdew, L. (2001).The exploitation of small learner corpora in EAP material design. In Ghadessy, M., Hendry a. & Roseberry R. L. (eds) *Small Corpus Studies and ELT. Theory and practice*. John Benjamins Publishing Company, 363-379.

Fraser, S. (2006). The nature and role of specialized vocabulary: What do ESP teachers and learners need to know? *Hiroshima Studies in Language and Language Education, 9*, 63-75.

Fraser, S. (2009). Breaking down the dision between general, academic and technical vocabulary: the establishment of a single, discipline-based word list for ESP learners. *Hiroshima Studies in Language and Language Education, 12*, 151-167.

Fryer, D. L. (2012). Analysis of the generic discourse features of the English-language medical research article: a systemic functional approach. *Functions of Language*, *19*(1), 5-37.

Fuertes Olivera, P.A. (2005). Specialized communication and English studies: research proposals on specialized lexicography and English for Specific Purposes. *Atlantis, 27*(2), 41–55.

Fulcher, G. (1999). Assessment in English for academic purposes: putting content validity in its place. *Applied Linguistics*, *20*(2), 221-236.

Garzone, G. (2006). *Perspectives on ESP and popularization*. Milano: CUEM.

Ghadessy, M., Henry, A. & Roseberry, R. L. (eds.) (2001), Small Corpus Studies and ELT. Theory and Practice. Amsterdam: John Benjamins.

Giannoni, D.(2008). Medical writing at the periphery: the case of Italian journal editorials. *Journal of English for academic Purposes, 7*(2), 97-107.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes, 19,* 115-135.

Gotti, M. (2006).*Advances in medical discourse analysis oral and written contexts*. Bern: Peter Lang.

Gotti, M. (2006). *Specialized discourse: linguistic features and changing conventions.* Bern: Peter Lang.

Gotti, M. (2007). CADIS - A Corpus for the Analysis of Identity Traits in Academic Discourse. Paper accepted for presentation at *6th European Symposium on Language for Special Purpose*s. University of Hamburg (Germany), August 27-31.

Gotti, M. (2011). Insights into medical discourse in oral and written contexts. In Loiacono, A., Iamartino, G. and Greco, K. (eds), *Teaching Medical English. Methods and Models.* Monza: Polimetrica. 29-52.

Gotti, M. (2014). Cooperative Meaning-Making Strategies in ELF University Courses. *Textus, XXVII*(1), 17-33.

Graddol, D. (2000). *The Future of English? A guide to forecasting the future of English in the 21st ce*ntury. London: The British Council, 42-47. Retrieved 10th Nov. 2012 from http://www.britishcouncil.org/learning-elt-future.pdf

Graddol, D., (2008). *English next. Why global English may mean the end of 'English as a foreign language*. London: The British Council, 106-110. Retrieved 11th Nov. 2012 from http://www.britishcouncil.org/learning-research-english-next.pdf

Granger, S. (2002). A bird's-eye view of computer learner corpus research. In S. Granger, J. Hung, S. Petch-Tyson, & J. Hulstijn (eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam & Philadelphia: John Benjamins, 3–33.

Granger, S. (2004). Computer learner corpus research: current status and future prospects. In U. Connor & T. Upton (eds.), *Applied corpus linguistics: a multidimensional perspective*. Amsterdam & Atlanta: Rodopi, 123–145.

Halliday, M. A. (2005). Lexicology . In. Halliday M.A.K, Cermáková A., Tuber, W., Yallop, C.(eds) *Lexicology and corpus linguistics an introduction*. London: Continuum, 1-22.

Halliday, M. A. K. (2004). *An introduction to functional grammar*. Revised by Matthiessen, C.M.I.M., (3 ed.). London: Arnold.

Harvey W., K. (2013). *Investigating adolescent health communication: A corpus linguistics approach*. London, Bloomsbury.Publishing Plc.

Hinkel, E. (2004). *Teaching academic ESL writing practical techniques in vocabulary and grammar*. Mahwah, N.J.: L. Erlbaum Associates.

Hudson, D. (1999). *Applied Linguistics*: *an attempt at definition*. Retrieved 17th Nov. 2012 from http://www. phon.ucl.ac.uk/home/dick/AL.html

Hamp-Lyons, L. (2001). English for academic purposes. In: Carter, R. and Nunan, D. (eds). *The Cambridge guide to teaching English tospeakers of other languages,* 126-130. Cambridge: Cambridge University Press.

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: the establishment of a medical wordlist. *Language Teaching Research, 17*(4), 454-484.

Hunston, S., & Oakey, D. (2010). *Introducing applied linguistics: concepts and skills*. New York, NY: Routledge.

Hutchinson T., & Waters, A. (1987). *English for specific purposes: a learning-centred approach* (25th ed.). Cambridge: Cambridge University Press.

Hyland, K. (1995). The author in the text : hedging scientific writing. Hong Kong papers in linguistics and language teaching, *18*, 33-42. Retrieved 13th July. 2013

from http://www.unibuc.ro/prof/zafiu_r/docs/2014/mar/05_11_58_17Hyland_author_1995.pdf
Hyland, K. (2002). Specificity revisited: How far should we go now? *English for Specific Purposes, 21*(4), 385-395.
Hyland, K. (2009). Specific purpose programs. In Long M. H. &. Doughty C. J. (eds), *The Handbook of Language Teaching.* Chichester, U.K.: Wiley-Blackwell, 201-217.
Hyland, K. (2011a). Disciplinary specificity: discourse, context and ESP. In Belcher, D. D., A. M. Johns,.& B. Paltridge (eds), *New directions in English for specific purposes research*. Ann Arbor, MI: University of Michigan Press, 11-12.
Hyland, K. (2011b). *English for Academic Purposes: an advanced resource book.* New York: Routledge Applied Linguistics.
Hyland, K. & Tse P. (2007) Is there an "academic vocabulary"?. *TESOL Quarterly*, *41*(2), 235-253.
Iwai, T., Kondo, K., Limm, S. J. D., Ray, E. G., Shimizu, H., and Brown, J. D. (1999). Japanese language needs analysis. Retrieved 13th Sept. 2013 from http://www.nflrc.hawaii.edu/ Networks/NW13/NW13.pdf
Johns, T. (1986). Micro-Concord: a language learner's research tool. *System*, *14*(2), 151-162.
Johns, A. M. & Dudley-Evans, T. (1991). English for Specific Purposes: international in scope, specific in purpose. *TESOL Quarterly, 25*(2), 297-314.
Johns, A.M., & Price-Machado, D. (2001). English for Specific Purposes (ESP): Tailoring courses to student needs —and to the outside world. In M. Celce-Murcia (Ed.), *Teaching English as a second or foreign language* (3rd ed.). Boston: Heinle & Heinle, 43-54.
Jordan, R. R. (1997). *English for academic purposes: A guide and a resource book for teachers*. Cambridge: Cambridge University Press.
Kaewpet, C. (2009). A Framework for Investigating Learner Needs: Needs Analysis Extended to Curriculum Development. *Electronic Journal of Foreign Language Teaching, 6*(2), Centre for Language Studies National University of Singapore, 209-220. Retrieved 09th Sept. 2013 from http://e-flt.nus.edu.sg/v6n22009/kaewpet.pdf
Knight, K., Lomperis, A. E., van Naerssen, M. & Westerfield, K.(2010). English for Specific Purposes. An overview for practitioners and Clients (Academic & Corporate). In *ESP at a Glance.* Retrieved 8th Jul. 2012 from http://www.academia.edu/3841239/ESP_at_a_Glance
Krashen, S., & Lee, S.Y. (2004). Competence in foreign language writing: progress and lacunae. *Literacy across cultures*, (*12*)2, 10-14. Retrieved 9th Jul. 2012 from http://www.sdkrashen.com.
Krashen, S., & Lee Brown, C. (2007). *What is Academic Language Proficiency*? University of Southern California.Singapore Tertiary English Teachers Society

(STETS copyright). Retrieved 9th July 2012 from http://www.sdkrashen.com/content/articles/krashen_brown_alp.pdf

Kūbler, N. & Aston, G. (2010). Using corpora in translation. In In Keeffe, A. & McCarthy M. (eds). *The Routledge handbook of corpus linguistics*. London: Routledge. 501-515.

Laso, N. J. and Salazar D. (2013). Collocations, lexical bundles and *SciE-lex.* A review of corpus research on multiword units and meaning. In Verdaguer, I, Laso, N. J. and Salazar D. (eds) Biomedical English. A Corpus-based Approach. Amsterdam: John Benjamins. 1-20.

Laufer, B. (1996). The lexical threshold of L2 reading: where it is and how it relates to L1 reading ability. In *Approaches to second language acquisition*, Sajavaara K. and Fairweather. C. (eds.) Jyvaskyla: Jyvaskyla Cross Language Studies 17, 55-62.

Laufer, B.(1997). What's in a word that makes it hard or easy? Intralexical factors affecting the difficulty of vocabulary acquisition. In Schmitt, N., & Mccharthy, M. (eds), *Vocabulary: description, acquisition and pedagogy*. Cambridge: Cambridge University Press, 140-155.

Laufer, B. (2001). Quantitative evaluation of vocabulary: how it can be done and what it is good for. In *Experimenting with uncertainty*. Elder C.,Brown A., Grove E., Hill K., Iwashita N., Lumley T., McNamara T., O'Loughlin K. (eds). Cambridge University Press, 241-250.

Laufer, B. and Nation P.. 1995. Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics 16,* 307-322.

Lee, S.H. (2003) ESL learners' vocabulary use in writing and the effects of explicit vocabulary instruction. *System*, *31*, 537-561.

Leech, G. (1992). Corpora and theories of linguistic performance. In Startvik J. (ed.), *Directions in corpus linguistics.* Berlin: Mouton de Gruyter, 105-122.

Leech, G. (1998). Preface. In S. Granger (ed.), *Learner English on Computer* . London, New York: Longman, xiv-xx.

Leech, G. (2007). New resources or just better old ones?. In Hundt, M., Nesselhauf, N. & Biewer, C. (eds), *Corpus Linguistics and the web.*Amsterdam: Rodopi. 134-49.

Leighton C., Leslie C., Juli C., Frank S. D., Robert E., Michelle E., et al. (2010). Authors' submission toolkit: a practical guide to getting your research published. *CMRO* (*Current Medical Research and Opinion) Informa Healthcare. 26*(8), 1967-1982.

Llach, M. A. (2011). *Lexical errors and accuracy in foreign language writing*. Bristol: Multilingual Matters.

Li, L.J., & Ge, G.C. (2009). Genre analysis and structural evolution of the English medium medical research article 1985-2004. *English for Specific Purposes*, *28*(5), 93-104.

Loiacono, A., Iamartino, G. and Greco, K. (2011). *Teaching Medical English. Methods and Models.* Monza: Polimetrica.

Löning, P. (1981): Zur medizinischen Fachsprache. Stilistische Gliederung und Textanalysen". In: Muttersprache 91, 79-92. Retrieved 11th Nov. 2011 from http://www.translationdirectory.com/articles/article2053.php

Lowe, I. (2010). Identifying technical vocabulary. An assessment of the views of. TM Chung and P Nation (2003, 2004). Retrieved 13th Apr. 2012 from http://www.docstoc.com/docs/45918328/ Identifying-technical-vocabulary-An-assessment-of-the-views-of-TM

Luzón Marco, M. J. (2000). Collocational frameworks in medical research papers: a genre-based study. *English for Specific Purposes, (19),* 63-86. Retrieved 14th Apr. 2012 from http://www.corpus4u. org/forum/upload/forum/2005062814544143.pdf

Maclean, J. (1977). Professional preparation: a technique for LSP teachers education. In R. Howard and G.Brown (eds), *Teacher education for LSP.* 158-75.

Maglie, R. (2009). *Understanding the language of medicine* (1st ed.). Roma: Aracne.

Magnet, A., Carnet, D. (2006). Letters to the editor: still vigorous after all these years? A presentation of the discursive and linguistic features of the genre. *English for Specific Purposes* 25(2)*, 173-199.*

Maher, J. (1986a) The development of English as an international language of Medicine. *Applied Linguistics 7*(2), 206-218.

Maher, J. (1986b). English for Medical Purposes. *Language Teaching*, *19*(2), 112-145.

Martin, A. (1976). Teaching academic vocabulary to foreign graduate students. *TESOL Quarterly, 10*(1), 91-97.

Mauranen, A. (2009). *English as a lingua franca: studies and findings*. Newcastle upon Tyne: Cambridge Scholars Publishing, 1-3. Retrieved 31st Oct. 2011 from http://www.cambridge scholars.com/download/sample/58659

Mauranen, A. (2011). English as a lingua franca in the academic world. In Belcher, D. D., A. M. Johns,.& B. Paltridge (Eds). *New directions in English for specific purposes research*. Ann Arbor, MI: University of Michigan Press. 94-117.

McAlister, F. A. et al (1999). The Medical Review Article Revisited: Has the Science Improved?. *Annals of Internal Medicine*, *131*(12), 947-951.

McDonough, J. (1984). ESP in perspective: A practical guide. London: Collins ELT.

McCarten, J. (2007). *Teaching vocabulary: Lessons from the corpus, lessons for the classroom*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (1996). *Corpus linguistics*. *An Introduction* Edinburgh: Edinburgh University Press.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice.* Cambridge: Cambridge University Press.

Meara, P. (1997). Toward a new approach to modelling vocabulary acquisition. In Schmitt, N., & McCharthy, M. (eds), *Vocabulary: description, acquisition and pedagogy*. Cambridge: Cambridge University Press, 109-121.

Meara, P. & Bell, H. (2001). P-Lex: a simple and effective way of describing the lexical characteristics of short l2 tests. *Prospect*, *16*(3), 5-19.

Miller, C. T (1984). Genre as social action. *Quarterly Journal of Speech, 70*(2), 151-167.

Milton, J. & Meara P. M. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL Review of applied linguistics*, *107/108,* 17-34.

Molhim, M. A. (2011). English for medical purposes course design for Arab university students. *English for Specific Purposes World 32*(10), 1-21. Retrieved 20th Dec. 2011 from http://www.esp-world.info/Articles_32/DOC/Molhim.pdf

Munby, J. (1978). *Communicative syllabus design*. Cambridge: Cambridge University Press.

Mungra, P. (2007). A research and discussion note: The macrostructure of consensus statements
in different genres of writing. *English for Specific Purposes. 26*(*1*)*,* 79–89.

Mungra, P., Webber, P. (2010). Peer review process in medical research publications: language and content comments. *English for Specific Purposes.* (*29),* 43-53.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press

Nation, P. & R. Waring. 1997. Vocabulary size, text coverage and word lists. In: Schmitt, N. and M. McCarthy (eds). *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press. 6–19.

Nation, P., & Waring, R. (2004). *Vocabulary size, text coverage and word lists.* Retrieved 5th Dec. 2012 from http://www.fltr.ucl.ac.be/ fltr/germ/etan/bibs/vocab/cup.html

Nation, I., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle, Cengage Learning.

Nattinger, J. (1988). Some current trends in vocabulary teaching. In Carter, R. & McCarthy, M. (eds). *Vocabulary and language teaching*. London: Longman. 62-81.

Nesselhauf, N. (2005). *Collocations in a Learner Corpus.* Amsterdam & Philadelphia: John Benjamins Publishing.

Ng Yu, J. (2013). Corpus-based Studies on Nursing Textbooks. *Advances in Language and Literary Studies,(4)*2, 22-28. Retrieved 13th Aug. 2014 from file:///C:/Users/francesca67/Desktop/172-342-1-SM-libre.pdf

Nwogu, K. (1997). The medical research paper: structure and functions. *English for Specific Purposes, 16*, 119-38.

O'Grady, W. et al., (2005). Contemporary linguistics: An introduction. New York: St. Martin's Press

O'keefe, A., McCarthey, M., & Carter, R. (2007*) From corpus to classroom. Language use and language teaching.* Cambridge: Cambridge University Press.

Paquot, M. (2010). *Academic vocabulary in learner writing from extraction to analysis*. London: Continuum.

Pearson, J. (1998) *Terms in Context*. Amsterdam: John Benjamins Publishing Company

Pinge, D. (2007). *Netiquette: What not to do online.* Retrieved 16th Sept. 2014 from http://www.rediff.com/getahead/2007/jul/10netiquette.htm

Ranieri, T. (2013). *Cross that bridge when you come to English: CLIL as a new challenge in Italy.*The European Conference on Language Learning 2013, University of Bari, Italy. Retrieved 13th Sept. 2014 from http://www.unifg.it/sites/default/files/allegatiparagrafo/21-01-2014/ranieri_clil_as_a_new_challenge_in_italy.pdf

Read, J. (2000). *Assessing vocabulary*. Cambridge: Cambridge University Press.

Règent, O. (2000), Pratiques de communication en medicine. Langages, *105*, 66-75.

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge: Cambridge University Press.

Richards J.K. (1989), *The context of language teaching.* Cambridge: Cambridge University Press.

Richards, J.C. (1976). The role of vocabulary teaching. *TESOL Quarterly 10,* 77-89.

Rizzo, C. R. (2010). Getting on with corpus compilation: from theory to practice. *ESP World, 1*(27)-9, 1-23. Retrieved 12th March 2013 from http://www.esp-world.info/articles_27/camino%20rea.pdf

Roberts, C. and Sarangi, S. (2003) Uptake of discourse research in interprofessional settings: reporting from medical consultancy. *Applied Linguistics* 24(3), 338–59.

Robinson, P. (1991). *ESP today: A practitioner's guide*. Prentice Hall, UK: Prentice Hall International (UK) Ltd.

Salager-Meyer, F. (1983). The lexis of fundamental medical English: classificatory framework and rhetorical function (a statistical approach). *Reading in a Foreign Language*. *1*(1), 54-64.

Salager-Meyer, F. (1990). Discoursal flaws in medical English abstracts: an analysis per research and text-type. *Text 10,* 365-84.

Salager-Meyer, F. (1992). A text-type and move analysis study of verb tense and modality distribution in medical English abstracts. *English for Specific Purposes, 11,* 93-113.

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes*. *13*(2), 149-171.

Salager-Meyer, F. (1997). Scientific multilingualism and "lesser languages". *Interciencia 22*(4), 197-201. Retrieved 11th Oct. 2012 from http://www.interciencia.org/v22_04/ensayo.html

Salager-Meyer, F., Defives, G., Jensen, C., & De Filipis, M. (1989). Principal component analysis and medical English discourse: An investigation into genre analysis. *System, 17*(1), 21–34.

Salager-Meyer, F., Alcaraz Ariza M.A. & Zambrano N. (2003) The scimitar, the dagger and the glove: intercultural differences in the rhetoric of criticism in Spanish, French and English medical discourse (1930-1995). *English for Specific Purposes*, *22,* 223-247.

Salager-Meyer, F., Ariza, M. A. & Pabón M. (2007).Collegiality, critique and the construction of scientific argumentation in medical book reviews: a diachronic approach. *Journal of Pragmatics 39,* 1758-1774.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.

Schmitt, N. (2002). *An Introduction to Applied Linguistics*. Hodder Arnold Press.

Schmitt, N.(2010) *An Introduction to Applied Linguistics*. Routledge Press. (2nd ed.).

Schmitt, N. (2010) *Researching vocabulary. A vocabulary research manual*. Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Schmitt, N., & McCarthy, M. (1997). *Vocabulary: Description, acquisition and pedagogy*. Cambridge: Cambridge University Press.

Schmitt, N., Schmitt, D. & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55-88. Retrieved 11th Nov. 2012 from http://www.norbertschmitt.co.uk/uploads/schmitt-n-schmitt-d-and-clapham-c-(2001)-developing-and-exploring-the-behaviour-of-two-new-versions-of-the-vocabulary-levels-test-language-testing-18-1-55-88.pdf

Schutz, N. & Derwing, B. (1981). The problem of needs assessment in ESP: some theoretical and practical considerations. In Mackay, R., & Palmer, J. D.(eds.) *Languages for specific purposes: program design and evaluation*. Rowley, Mass.: Newbury House Publishers. 29-44.

Scott, M. (1999). *WordSmith tools help manual.* Version 3.0. Oxford: Oxford University Press.

Scott, M. (2001). Comparing corpora and identifying key words, collocations, and frequency distributions through the WordSmith Tools suite of computer programs. In Ghadessy, M., Henry, A. & Roseberry, R. L. (eds.), *Small Corpus Studies and ELT. Theory and Practice*. Amsterdam: John Benjamins. 47- 67.

Scott, M. (2012). *Version 6.0 WorSmith Tools.* Oxford: Oxford University Press.

Scott, M. Tribble, C. (2006). *Textual Patterns. Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins

Seedhouse, P. (1995). Needs analysis and the general English classroom. *ELT Journal*, *49*(1), 59-65.

Seidlhofer, B. (2011). *Understanding English as a lingua franca: a complete introduction to the theoretical nature and practical implications of English used as a lingua franca*. Oxford: Oxford University Press.

Sinclair, J. (2004) Intuition and annotation - the discussion continues. Advances in corpus linguistics. Papers from the *23rd International Conference on English Language Research on Computerized Corpora* (ICAME 23). Aijmer, K and Altenberg, B. (eds.). Amsterdam/New York: Rodopi.

Songhori, M. H.(2008). Introduction to needs analysis. *English for Specific Purposes world*, *4*. Retrieved 13th Sept. 2013 from http://www.esp-world.info/articles_20/doc/introduction %20to%20needs%20analysis.pdf

Skelton, J. (1994) Analysis of the Structure of Original Research Papers: an Aid to Writing Original Papers for Publication. *British Journal of General Practice*. *44(*387), 455 – 459.

Skelton, J. (2012). English for Medical Purposes. *The Encyclopedia of Applied Linguistics.* Blackwell Publishing Ltd

Skelton, J. & Whetstone, J. (2012). English for Medical Purposes and Academic Medicine: looking for common ground. *Ibérica 24*, 87-102. Retrieved 07th Sept. 2013 from http://www.aelfe.org/documents/08_24_Skelton.pdf

Stevens, V. (1995): Concordancing with language learners: Why? When? What?. *CAELL Journal 6*(2), 1-10

Strevens, P. (1988). ESP after twenty years: a re-appraisal. In M. Tickoo (Ed.), *ESP: State of the Art,* Singapore: SEAMEO Regional Language Centre, 1-13.

Stubbs, M. (2010).Three concepts of keywords. In M Bondi & M Scott (eds) *Keyness in texts: Corpus Linguistic investigations*. Amsterdam: Benjamins. 21-42. Retrieved 17th Oct. 2013 from http://www.uni-trier.de/fileadmin/fb2/ ANG/Linguistik/Stubbs/stubbs-2008-keywords.pdf

Swales, J. M. (1985). *Episodes in ESP: a source and reference book on the development of English for Science and Technology*. Oxford: Pergamon Institute of English.

Swales, J. M. (1990). *Genre analysis: English in academic and research settings*. Cambridge: Cambridge University Press.

Swales, J. M. (1997). English as a Tyrannosaurus Rex. *World Englishes*, *16*(3), 373-382.

Swales, J. (2004). *Research genres: explorations and applications*. Cambridge, UK: Cambridge University Press.

Tarone, E. and Yule, G. (1989). *Focus on the Language Learner*. Oxford: Oxford University Press.

Taylor, L., Leech G.& Fligelstone S. (1991). A survey of English machine-readable corpora. In Johansson and Stenstrom (eds.), *English Computer Corpora*. *Selected Papers and Research Guide*. Berlin, New York: Mouton de Gruyter. 319-354.

Teubert, W. & Cermàkovà, A. (2007). *Corpus linguistics a short introduction*. London: Continuum.

Tognini-Bonelli (2010). Theoretical overview of the evolution of corpus linguistics. In Keeffe, A. & McCarthy M. (eds). *The Routledge handbook of corpus linguistics*. London: Routledge. 14-27.

Trimble, L. (1985). *English for science and technology: a discourse approach*. Cambridge: Cambridge University Press.
Vartalla, T. (1999). Remarks on the communicative functions of hedging in popular scientific and specialist research articles on Medicine. *English for Specific Purposes*, *18*(2), 177–200.
Wallwork, A. (2011), *English for academic correspondence and socializing*, New York: Springer.
Wang, J., Liang S.L & Ge, G.C (2008). Establishment of a medical academic wordlist. *English for Specific Purposes 27*, 442-458.
Webb, S., (2005). Receptive and productive vocabulary learning: The Effects of Reading and Writing on WordKnowledge. *Studies in Second Language Acquisition, 27*(1), 33-52.
West, M. P. (1953). *A general service list of English words: with semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longmans. Green.
West, R. (1998). ESP- State of the art. Retrieved 11th Sept. 2013 from www.man.ac.uk /CELSE/esp/west.htm
Widdowson, H. G. (1983). *Learning purpose and language use*. Oxford: Oxford University Press.
Wulff, H. R. (2004). The language of medicine. *Journal of the Royal Society of Medicine, 97,* 187-188.
Zoltán, D. (2007). *Research methods in applied linguistics : quantitative, qualitative, and mixed methods*. Oxford: Oxford University Press.
Zweigenbaum, P., . Jacquemart, N., & Grabar, B. (2001) Building a text corpus for representing the variety of medical language. *Health Technology and Informatics,* Ebook , *84*, 245-254.

ONLINE CORPORA

BNC, *British National Corpus*. Available at http://ucrel.lancs.ac.uk /bncfreq/flists.html and http://www.natcorp.ox.ac.uk/
CADIS Corpus (1980-2011), CERLIS *Research Centre on Languages for Specific Purposes*, University of Bergamo. Available at http://dinamico.unibg.it/ cerlis/public/CADIS_Corpus.pdf
CEEM, *Corpus of Early English Medical Writing* (2005). Available at http://www.helsinki.fi/ varieng/CoRD/corpora/CEEM/Emodcontext.html)
THF Corpus, *Teenage Health Freak Corpus* (2007). Available at http://www.nottingham.ac.uk/research/groups/cral/projects/thf/thfcorpusoverview.aspx

WEB SITES:

Council of Europe. (2001). *Common European Framework of Reference for Languages*. Available at http://www.coe.int/T/DG4/Linguistic/Source/ Framework_EN.pdf

DART-Europe E-theses Portal. Available at (http://www.dart-europe.eu/basic-search.php).

doc2doc. Available at (http://doc2doc.bmj.com/):

Drugs.com. Available at http://www.drugs.com/

Expert Advisory Group on Language Engineering Standards (EAGLES). Available at http://www.ilc.cnr.it/ EAGLES/intro.html).

MedicalStudent.com. Available at http://www.medicalstudent.com/#Learn

MedlinePlus. A service of the U.S. National Library of Medicine. Available at http://www. nlm.nih.gov/medlineplus/encyclopedia.html

PubMed. A service of the US National Library of Medicine and National Institutes of Health. Available at http://www.pubmed.gov

The British National Journal. Available at http://www.bmj.com/about-bmj/resources-authors/forms-policies-and-checklists/bmj-right-journal-my-research-article

The Lancet. Available at http://www.thelancet.com/

The Ministry of Education, Universities and Research. Available at http://hubmiur.pubblica. istruzione.it/ web/istruzione/home

The Official Journal of the European Union. Available at http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri OJ:C:2009:119:0002:0010:EN:PDF

US Department of Health-National Guidelines. Available at http://www.guideline.gov/

WEB TOOLS AND FREE-WARES:

Acronym Generator freeware. Available at https://www.cs.uoregon.edu/ research/paracomp/anym/

Coxhead, A. (2000) *Academic Word List*. Available at http://www.victoria.ac.nz/lals/resources/ academicwordlist/ information.aspx

Coxhead, A. (2000) *AWL Highlighter*. Available at http://www.nottingham.ac.uk/ alzsh3/acvocab/wordlists.htm

Dragon Dictation PC Application. Available at http://download.cnet.com/Dragon-Dictation/3000-2064_4-75000022.html

Cobb,T. (2005) *Web Vocabprofile,* an adaptation of Heatley, Nation & Coxhead's (2002) *Range*. Available at http://www.lextutor.ca/vp/

Heatley, A., Nation, I.S.P. & Coxhead, A. (2002). *RANGE and FREQUENCY programs*. Available at http://www.victoria.ac.nz/lals/staff/paul-nation.aspx
WordSmith 6.0 (2005) Available at http://www.liv.ac.uk/~oms2928/homepage.html

WEB DICTIONARIES

TheFreeDictionary's Medical dictionary. Available at http://medical-dictionary.thefreedictionary.com/
Latin dictionary. Available at http://latindictionary.wikidot.com/noun:corpus
McGraw-Hill Concise Dictionary of Modern Medicine. (2002). Available at http://medical-dictionary.thefreedictionary.com/
Miller-Keane Encyclopedia and Dictionary of Medicine, Nursing, and Allied Health, Seventh Edition. (2003). Available at http://medical-dictionary.thefreedictionary.com/ intraprofessional +team
Segen's Medical Dictionary. (2011). Available at http://medical-dictionary.thefreedictionary.com/ intraprofessional+team