Review of the habilitation thesis
entitled The vocabulary of medical English: A Corpus-based study
by mgr Renáta Panocová, PhD

The thesis entitled The vocabulary of medical English: A Corpus-based study by mgr Renáta Panocová, PhD fully complies with the academic and procedural requirements specified for habilitation theses.

The very idea of the research project undertaken by dr Panocová is commendable. She subjects to scrutiny two extensive corpora with a view to devising accurate and effective methodology that could be applied so as to delimit and characterize medical vocabulary in English. The corpora in question are the subcorpus of the Corpus of Contemporary American English (COCA) based on medical research papers, amounting to nearly 9 million words, and the WIMECO corpus compiled by the Author on the basis of Wikipedia medical articles, amounting to 1 111 735 words. This is a theoretically and empirically oriented research project, which will have practical applications in as important and relevant an area as medical vocabulary. With English taking the lead in international communication, and in academic discourse in particular, there is a strong need for a comprehensive compendium of specialized words and terms. The research project grew out of the dissatisfaction of the Author with random lists of medical words compiled randomly to serve pedagogical purposes. It is a response to the growing need to identify and characterize the whole of medical terminology in English on the basis of extensive corpora that are now available thanks to enormous progress made in the area of computer technology, data processing, storage and retrieval. Developing an optimal methodological approach is as urgent as it is difficult. Dr Panocová makes two important and original claims. First, she argues for the necessity of investigating the correlation of absolute and relative frequency in characterizing medical vocabulary. Second, she demonstrates the crucial role of corpus selection in determining medical vocabulary.
The thesis consists of four chapters and the conclusion. There are also abstracts in English and Slovak with lists of key words, lists of figures and tables, a resumé in Slovak, bibliography and two appendices. In terms of construction and organization of material the thesis is complete and logically structured.

In chapter 1 Introduction: defining medical English, dr Panocová contrasts two perspectives in the study of medical English, i.e. pedagogical, which involves compiling lists of high frequency words for teaching materials, and descriptive, which strives to characterize medical English as a linguistic subsystem. It also provides the definition of medical English and brings up the question of the relationship of medical English to general English. We learn that medical English can be viewed in terms of register and language for special purposes or as a sublanguage with a part-whole relation with no clear-cut boundaries. Delimiting a subset in mathematics yields a closed domain. Delimiting the subset within language would amount to pointing to a limited set of sentences, if Harris’ (1968) mathematical closure were to be observed. Harris himself discerns the difference between language and mathematics and argues that a more feasible task is to arrive at a vocabulary closure. Following Kittredge (1983), dr Panocová demonstrates that defining a subset of a particular language is not a trivial matter and there is no unanimously accepted definition and criteria for distinction in characterizing the sublanguage as a linguistic system. The register perspective makes reference to linguistic competence and relies on intuitions, which can be misleading, and she argues in favour of approaches which deploy linguistic corpora, since they are a tangible source to rely on. The chapter concludes with a list of research questions that will be pursued in the thesis and the presentation of the lay-out of the thesis.

One minor critical comment that might be raised with reference to chapter 1 is that one can get the impression that English for special purposes and register are one and the same phenomenon, whereas specialized vocabulary may be one of the many distinctive traits that define a register. This distinction could have been more explicitly brought out.

Chapter 2, entitled Determining the vocabulary of medical English, is an excellent overview of previous research attempts at characterizing and defining medical English. We can trace the development of corpus-based research of specialized vocabulary and are offered a critical evaluation of the methodologies applied. This chapter starts with the explanation of key notions such as lemma/lexeme, word family, specialized vocabulary and terminology, which are instrumental in understanding the difference in the methods applied by various researchers. The terms are clearly explained, the distinction between specialized vocabulary and terms in the narrow sense, in particular. This point is discussed with reference to ten
Hacken’s (2010, 2015) observation that the former are grounded in prototypes and are gradient, whereas the latter show clear-cut boundaries. Consequently, it is not sufficient that the word is used in a specialized context, as Coxhead (2013) would have it. A proper characterization of specialized vocabulary in the lexicon calls for a more elaborate and fine-grained classification. Consequently, the Author chooses to adopt the system developed by ten Hacken (2010, 2015) since it allows for the scalar nature of terms (in a broad sense). In view of the results obtained in the research part, this move seems well justified. The only source mentioned with reference to the theory of prototypes is Labov (1973). However, a word of comment might have also been devoted to cognitive psychologists such as Rosch or Mervis.

The Author discusses in turn and pinpoints the problems with methods used to delimit specialized vocabulary which are not corpus-based such as consulting experts and technical dictionaries, and Chung and Nation’s (2003) four step scale. They are dismissed as inadequate since they involve subjective decision making. Dr Panocová also pinpoints methodological shortcomings of early corpus-based approaches. The starting point for the discussion of the use of corpora in identifying vocabulary for ESP is Coxhead’s (2000) Academic Word List (AWL). It is rightly criticised for making use of word families, used in the sense of Bauer and Nation (1993), which obscures the distinction between homonymous but distinct lexical categories. Another problem is that only word families outside the first 2000 most frequently occurring words of English as specified in West’s (1953) GSL can be treated as specialized. This has the undesirable effect of leaving out specialized lexical items only on grounds that they are also frequent in the general language. The same design flaws are identified in Wang, Liang and Ge’s (2008) attempt at characterizing medical vocabulary. Coxhead’s study takes into account specialized occurrence, range and frequency, prioritizing range to frequency. In the most recent study aiming at devising a new academic vocabulary list, Gardner and Davis (2013) propose to use lemmas instead of word families and to use a large and representative corpus of contemporary English. In their approach the new list must be statistically derived and tested against other corpora. The resulting academic vocabulary list (AVL) can definitely be considered an advancement on the academic word list (AWL) devised by Coxhead a good few years before. Dr Panocová sets out to demonstrate in the following chapter how the latest state-of-the-art methodology used to characterize academic vocabulary in general could be applied in the characterization of specialized vocabulary relating to medicine. The selection of this model is appropriate and justified, since it allows the Author to get round the
methodological issues raised with reference to the previous accounts, i.e. the use of word families, the use of West’s GSL (1953) and problems with the structure of the corpus.

Chapter 3 entitled *Methodological considerations* is most prominent in the thesis since it contains the Author’s original methodological proposals with regard to the use of corpora in the characterization of medical vocabulary. Dr Panocová provides convincing arguments in favour of using the COCA to serve the purpose of characterizing and describing medical English. This simply is the largest corpus of medical texts available. The corpus compiled for Wang et al.’s (2008) study was eight times smaller. She gives a detailed description of the medical subcorpus ACAD: Medicine in COCA and we learn how the structure of the corpus influences the characterization of medical vocabulary. The procedure of characterizing medial vocabulary is presented. The overall argumentation is convincing. The use of lemmas is justified. The replacement of the GSL with the notion of relative frequency makes good sense. She also rightly observes that in the study of a specialized language there is no need for a strict compartmentalization of the corpus, as was the case in studies of general academic vocabulary. Words are selected when their frequency in the specialized corpus is significantly higher than in a general language corpus. The relative frequency is calculated by taking the normalized frequency of the medical corpus and by dividing it by the normalized frequency of the general corpus. This measure shows how typical a word is in medical vocabulary. If the relative frequency value is 1, the word is just as frequent in general as in medical texts. The higher the relative frequency the more specialized the word is considered to be. Another dimension that should be taken into consideration is the absolute frequency of the word in medical texts. These words must be frequent enough in the corpus of medical texts. In the proposed model, medical vocabulary should be viewed in terms of a cline or continuum based on the interaction of absolute and relative frequency. The methodology adopted seems sound and it gives satisfactory results. However, one of the weak points of relative frequency, which the Author does not sweep under the proverbial rug, is that the relative frequency is dependent on the coverage of other domains used in the general corpus. As a result the relative frequency of *experiment* is lower than that of *extension*, as discussed on p. 68 of the thesis. Maybe a prospective path for future research would be to calculate relative frequency not with respect to normalized absolute frequency, but to the normalized frequency of words in the academic subcorpus only. Then we would arrive at the vocabulary list which is devoid of lexical items that characterize specifically academic discourse. The key issue in characterizing the medical vocabulary is to appropriately set the threshold values of the absolute and relative frequency that would qualify a given item for inclusion. The Author criticises Coxhead (2000) and
Wang et al. (2008) for stipulating threshold values. She herself does indeed show what the effect of different values would be, but stops short of giving any definite threshold values leaving that to future research.

Chapter 4 is a discussion of results yielded by the analysis of an alternative corpus compiled by the Author herself on the basis of medical texts from the Wikipedia. The purpose of this chapter is to substantiate the claim that the selection of the corpus may affect, in a statistically significant way, the characterization of medical vocabulary. At first, I did not see the point of compiling a corpus based on the Wikipedia, because it is up to the user to select the key words in the construction of the corpus. By selecting the top ten lexical content words form the Academic Medicine COCA wordlist we are prejudging the result, confirming the results of the previous analysis. Another disadvantage is that the initial selection involves some arbitrary decision making on the part of the compiler and is not based solely on frequency. The ten seed words for the construction of the corpus are not function words or general academic words. In the Author’s view, however, the bias in data selection is only apparent, since in a control corpus with a given seed word omitted, the list of 500 keyword nouns would return the omitted seed word, albeit with a different ranking and absolute frequency. I find this explanation convincing. Additionally, this state of affairs does not affect the Author’s main line of argument that the corpus with different text types is bound to give different results, i.e. the nature of the corpus influences what vocabulary emerges as medical vocabulary.

The research results are summarized in the conclusion. This is not a mere recapitulation of conclusions. The Author returns to the research questions posed in the Introduction and answers each of them in turn to show that the aims set at the beginning have been reached.

As is usually the case with committing to paper the results of an extensive research project, there are bound to be some minor editorial and stylistic slip-ups, which, out of duty, I enumerate below. Eliminating these shortcomings is desirable before publishing the results in a book format, which, in my opinion, should definitely be carried out.

1) Cowan (1974) and Farrell (1990) referred to on p.17 are not listed in the references
2) p. 24 l. 4 is amd should be and
3) p.42 l. 12 is attained should be acquired, l. 14 is list should be lists
4) p. 59 reference is made to McEnery and Hardy (2011) whereas in the references there is Hardie
5) p. 67 l.7 typical for as opposed to typical of in l.18, such inconsistent use of prepositions is also found in other places (e.g. p. 111)

6) p. 133 is Appendix 2 should be Appendix 1

To sum up, this is a well thought-out research project, extensive in its scope, maintaining high academic standards. The subject matter of the thesis concerns a relevant area in the investigation of the lexicon/terminology, which is medical vocabulary. The methodological proposals put forward in the thesis are original and valid. The theoretical background is accurately and comprehensively described. The sources cited (121 altogether) are relevant, complete and up-to-date. The language of the text is appropriate and arguments are presented clearly in a lucid style. The thesis is an important voice in the debate concerning the corpus-based description and characterization of medical vocabulary of the English language and the conclusions reached by dr Panocová should be taken into account in the proper description and characterization of medical English.

I hereby express my positive opinion of dr Panocová’s habilitation thesis and propose that she should be admitted to further stages of the habilitation process.