E·S·P·Today
JOURNAL OF ENGLISH FOR SPECIFIC PURPOSES AT TERTIARY LEVEL

Gorica Tomić*

Center for Language and Literature Research
Faculty of Philology and Arts
University of Kragujevac, Serbia
gorica.tomic@filum.kg.ac.rs

A CORPUS-BASED ANALYSIS OF THE COLLOCATIONAL BEHAVIOR OF THE NOUNS *DEFORMITY* AND *MALFORMATION* IN MEDICAL ENGLISH

Abstract

This paper presents a corpus-based analysis of some aspects of the collocational behavior of the nouns *deformity* and *malformation* in the context of English for Medical Purposes (EMP). The aim was to determine which nominal and adjectival lexical items denoting or relating to the category of human body parts the two nouns exclusively or predominantly collocate with within three different patterns, namely '*modifiers of deformity/malformation*', '*deformity/malformation of*', and '*deformity/malformation in*' in the *English Web Corpus 2015* available in the *Sketch Engine* software. The results of the quantitative and qualitative analyses of the collocational information retrieved from all three patterns indicate that, despite the fact that there are collocates common to both *deformity* and *malformation*, *deformity* tends to co-occur with nominal and adjectival lexical items which denote or are related to parts of the human musculoskeletal system and the hard(-tissue) structures, while *malformation* has a tendency to co-occur with nouns or noun phrases and adjectives denoting or relating to (parts of) the internal organs of the human body and its soft(-tissue) structures. The paper also offers some pedagogical implications for EMP teaching and learning.

Key words

collocation, *deformity*, *malformation*, corpus, English for medical purposes.

* Corresponding address: Gorica Tomić, Filološko-umetnički fakultet, Centar za proučavanje jezika i književnosti, Jovana Cvijića bb, 34000 Kragujevac, Srbija.

1. INTRODUCTORY REMARKS

According to Panocová (2017: 11), "a central question in any subfield of English for Specific Purposes (ESP) is how it relates to the lexicon". In analyzing and ultimately establishing such relations, corpora represent an invaluable instrument, although it stands to reason that "no corpus (nor a dictionary or a usage manual) will ever be developed into such perfection that it would merit being examined alone" (Kaunisto, 2007: 301). Following these statements, as well as some of the assumptions and methods underlying corpus linguistics provided in Gries (2006: 4-6) and Biber, Conrad, and Reppen (1998: 4-5), the present research aims to determine which nominal and adjectival lexical items denoting or relating to the category of human body parts the two medical terms *deformity* and *malformation* exclusively or frequently and "by way of habit" co-occur with within three different patterns, namely '*modifiers of deformity/malformation*', '*deformity/malformation of*', and '*deformity/malformation in*', in a large corpus of electronic texts. Therefore, the research will attempt to provide answers to the following questions: Which quality makes some of the collocates retrieved from the corpus occur exclusively or predominantly with *deformity*, and which quality makes some other collocates occur exclusively or predominantly with *malformation*? In addition, the paper aims at determining the "common denominator" for their respective sets of collocates, thereby reducing the mass of the data produced about the two nouns and revealing some general tendencies in their collocational behavior. It must be remarked that the paper further attempts to accommodate those lexical items common to both *deformity* and *malformation* (if any). The obtained results are expected to be pedagogically relevant and help English for Medical Purposes (EMP) teachers to design such materials which may help EMP learners use the terms appropriately, as it is not uncommon for terms to be misused or used interchangeably in specialized or technical vocabulary.

The remainder of the paper is divided into five sections: Section 2 concerns the use of corpora in ESP/EMP vocabulary research whereas Section 3 provides an account of the methodology used in the study. Section 4 presents a detailed analysis of the collocational information derived from the corpus as well as a discussion of the results thereby obtained. Pedagogical implications are discussed in Section 5. Finally, the most relevant findings and recommendations for further research are summarized in the last section.

2. THE USE OF CORPORA IN ESP/EMP VOCABULARY RESEARCH

As in many other areas of language research, the use of corpora has considerably influenced ESP vocabulary research. A growing number of studies have used

various types of language corpora to analyze and describe discipline-specific vocabulary in English or to compare and contrast the frequency and distribution of a range of lexical items across disciplines, further discussing the pedagogical implications of such corpus linguistic approaches to vocabulary analysis. For instance, Nelson (2006) investigates the semantic associations (i.e. collocations and semantic prosodies) of words in the business English lexical environment using a corpus of both spoken and written business English and concludes that the prosodic categories identified should be conceptualized as tendencies rather than some absolute qualities (Nelson, 2006: 231). One other study of business vocabulary (Walker, 2011) shows how a corpus-based investigation of the collocational behavior of the key lexis of business English can help teaching English to those working or preparing to work in the field of business. A corpus-based study of the vocabulary of agriculture semi-popularization articles in English conducted by Muñoz (2015) aims to determine the high-frequency words appearing in this specific genre. A similar descriptive approach is adopted by Panocová (2016) for the characterization of medical English vocabulary using the medical subcorpus of *the Corpus of Contemporary American English* (COCA). Coxhead and Demecheleer (2018) use the written corpus and frequency principles to identify the technical vocabulary of plumbing, as well as the written and spoken corpora to create a plumbing word list. The vocabulary of civil engineering is investigated by Gilmore and Millar (2018) by means of a specialized corpus of civil engineering research articles, for the purposes of performing keyword analysis, that is, of identifying words related to civil engineering research articles that may have some pedagogical value, as well as comparing these words with the existing wordlists. One of the most recent corpus studies in ESP vocabulary research is Riccobono (2020). The aim of his study is twofold: to compile a corpus of baseball English and to use it to create specialized or technical vocabulary sets for teaching and learning English for Baseball Purposes, notably word and phrase lists. As mentioned above, there are, in addition, corpus studies which take a cross-disciplinary approach to researching ESP vocabulary. For instance, Ward (2007) studies and compares the frequencies of common nouns and their collocations in chemical engineering textbooks with those of the same collocations in four other disciplines (civil, electrical, industrial, and mechanical engineering), further demonstrating that such collocations are highly discipline-specific. Durrant (2009) explores the possibility and utility of creating a list of frequent cross-disciplinary collocations, therefore emphasizing the importance of expanding the existing wordlists by including their frequent collocations. The importance of collocation research is further highlighted by Peacock (2012) who performs a corpus analysis of the distribution of the high-frequency collocates of abstract nouns in research articles across eight disciplines, namely Chemistry, Computer Science, Materials Science, Neuroscience, Economics, Language and Linguistics, Management, and Psychology.

The growth of corpus-based research into ESP vocabulary has affected vocabulary teaching/learning as well, therefore attesting to "the close relationship between corpus linguistics and language teaching [/learning]" (Coxhead, 2002: 72). Specifically, many vocabulary studies using corpus-linguistic methodology have led to the writing of curricula, syllabi, coursebooks, or to compiling ESP wordlists whose number, as rightly remarked by Coxhead (2018: 4), has never been greater. Similarly, Smith (2020: 1) observes that "[o]ne of the principal applications of corpora in English language teaching and learning has been the compilation of vocabulary lists". One of the very first corpus-based studies motivated by the need to identify the academic vocabulary that could be used for designing language course materials was Coxhead's (2000) *Academic Word List* (AWL). Though this list became the norm in English language education, it later received some criticism for the methods adopted, such as determining word frequencies by means of word families (instead of lemmas, for example) or for its relation to the *General Service List* as a rather old list (Gardner & Davies, 2014: 307). Accordingly, the need for a new AWL was recognized, among others, by Gardner and Davies (2014), who established *A New Academic Word List* (ANAWL), using a notably larger and more up-to-date academic corpus than Coxhead (2000). Specifically, their corpus consisted of the written academic texts produced within nine disciplines, including, *inter alia*, medicine and health, whereas Coxhead (2000) used a written corpus of academic texts from the disciplines of arts, commerce, law, and science. Other authors such as Wang, Liang, and Ge (2008) utilized a corpus of online medical research articles to compile a list of the most frequent medical academic words – the *Medical Academic Word List* (MAWL), further confirming that medical academic vocabulary forms an important part of this genre (Wang et al., 2008: 442). Hsu (2013), for example, used a corpus of medical textbooks across thirty-one medical subject areas in an attempt to create a medical word list (MWL) which would make the difference between technical and non-technical vocabulary less sharp, as well as provide a valuable insight into the most-frequently used medical words to those which are new to the medical register (Hsu, 2013: 456, 468). One of Hsu's (2013: 470) suggestions regarding the use of the MWL in EMP classes concerns the provision of a glossary of these medical words, together with their most common collocates. Combining the methods of Coxhead (2000) and Gardner and Davies (2014), Lei and Liu's (2016) corpus study developed the *Medical Academic Vocabulary List* (MAVL), which had a wider coverage of medical English and was fairly shorter than the MAWL, therefore better serving the vocabulary needs of EMP learners. However, despite the good reception and value of such wordlists (Coxhead, 2018: 21-45; Nation, 2016: 3-13), they frequently prove to be insufficient for ESP students to rise to the challenges of learning the vocabulary of the specific discipline, as naturalness and proficiency in language use, including specialized languages such as medical English, require learners to acquire not only individual discipline-specific words, but also the typical collocations of these words. That is, in addition to the acquisition of individual words, learners need to

understand the way individual words naturally select each other and combine into larger recurrent formulaic or multi-word units (MWUs) like collocations. These can help learners to efficiently use and understand specialized materials, as well as to sound native-like when presenting their own ideas both in written and spoken English.

Regarding the use of language corpora in EMP vocabulary research in particular, different aspects of medical vocabulary in English have been the subject of a number of corpus studies, many of which aim at using the existing wordlists to quantitatively analyze the lexical coverage in various (sub)corpora and/or produce their own wordlists. For instance, Chen and Ge (2007) study the frequency and distribution of Coxhead's AWL word families in a corpus of 50 medical research articles (RAs), as well as in a sub-corpus consisting of the five sections of a medical RA (i.e. Abstract, Introduction, Materials and Methods, Results, and Discussion), with the aim of identifying the most prominent medical words in this list. They find that the text coverage of the AWL word families in their medical RAs is quite high (10.073%), therefore reaching a similar conclusion as Wang et al. (2008: 442) that academic words represent significant items in medical RAs (Chen & Ge, 2007: 508, 513). Similarly, Coxhead and Quero (2015) investigate the text coverage of some of the existing general and academic wordlists in the two corpora of medical textbooks with the intention of learning about the nature of high-frequency vocabulary in EMP. The main purpose of Quero's (2015) research is using a corpus of medical texts to identify the most frequent and relevant lexis EMP teachers and learners need for comprehending such texts more readily. In contrast to those corpus studies into medical vocabulary whose approach is principally pedagogical, Panocová (2017) aims at describing the vocabulary of medical English using the COCA corpus (which includes a medical subcorpus *ACAD: Medicine*) rather than producing a medical word list. She argues that the characterization of medical vocabulary is much more complex than is generally implied by a simple wordlist, as well as that the vocabulary of medical English is best regarded as a continuum based on absolute and relative frequency (Panocová, 2017: 41, 106). Similarly to Quero (2015), Quero and Coxhead (2018) attempt at identifying high-frequency medical vocabulary using multiple corpora of medical written texts, including some general wordlists, and integrating these corpus-based findings (in the form of a specialized wordlist) into an ESP (reading) course for medical students to help them start reading medical textbooks more efficiently. Hsu (2018) explores the vocabulary of English-medium traditional Chinese medicine (TCM) textbooks, with the aim of developing a TCM English wordlist as a reference for English for Chinese Medicine purposes. Last but not least, Le and Miller's (2020) corpus-based study has a somewhat more specific focus than the studies reviewed above, as it intends to produce a list of the most commonly occurring medical morphemes that could help EMP students improve their medical vocabulary and especially their morphological knowledge.

3. METHODOLOGY

The corpus utilized for the purpose of this research is the *English Web Corpus 2015* (EW15), which totals more than 15 billion words. It is accessed and processed via the *Sketch Engine* (SE) text analysis software. As stated on its website (https://www.sketchengine.eu/), the corpus has been crawled from the Internet via a web crawler designed for linguistic purposes and formed "using technology specialized in collecting only linguistically valuable web content". The choice of a non-specialized corpus over the others available in the software, including specialized corpora such as the *Medical Web Corpus* (MWC), does not endanger the validity of the results and conclusions, since the texts in which the words under analysis occur originate from medically relevant sources (e.g. academic journal articles, websites of healthcare facilities, etc.). The main reason for not using the MWC is the fact that the *Word Sketch Difference* in this corpus provides no collocational patterns that are the subject of this paper. In addition, the size of the EW15 is more impressive than that of the MWC, which totals 33,961,786 words.

The tools to work with in the EW15 include, *inter alia*, the *Word Sketch* and the *Word Sketch Difference*, of which the latter was used in this research. Namely, in contrast to the *Word Sketch* which "processes the word's collocates and […] can be used as a one-page [automatic, corpus-based] summary of the word's grammatical and collocational behaviour" (https://www.sketchengine.eu/), the *Word Sketch Difference* is designed and can be implemented for comparing the use of two (more often than not semantically related) lemmas via their collocates, the use of two different word-forms of the same lemma via their collocates, or the use of the same lemma in two different subcorpora of the same corpus. It therefore makes the comparison between the two lemmas (in one corpus or its subcorpora) or the two word-forms of the same lemma more effective by automatically generating both *word sketches* and highlighting those collocates that make the difference.

In each subcorpus, collocates are automatically grouped into those occurring exclusively with *deformity* (color-coded green), those being used with both *deformity* and *malformation* (appropriately color-coded gray), and those occurring exclusively with *malformation* (color-coded red).[1] Each collocate is accompanied by two numbers, the first of which indicates the number of times it occurs with *deformity* and the second one the number of times it occurs with *malformation*, as part of the selected collocational pattern in the whole corpus.

With regard to the lexico-semantic characterization of collocates the corpus is searched for nominal and adjectival lexical items (including terminological

[1] Note, however, that almost all three subcorpora display two extra areas: color-coded light green and light red, respectively. These contain collocates which, although used with both *deformity* and *malformation*, show a strong tendency towards one of the two nouns. For the sake of simplicity, I decided to consider them all part of the "gray area", but also to emphasize their original position where necessary in the paper.

syntagms) which denote or are related to human body parts. Positions of the relevant collocates within the keyword in context (KWIC) concordances are counted in one of the following two ways, depending on the collocational pattern. For instance, in the structurally similar patterns '*deformity/malformation of*' and '*deformity/malformation in*', the relevant collocate, be it a single-word unit or terminological syntagm such as *corpus callosum*, is positioned within the range of four words to the right of the KWIC. That is, the relevant collocate may take the position of one (or possibly more, if syntagmatic) of the three words that occur to the right of the preposition *of* and *in*, respectively, as in (1)–(4) below.

(1) In July 2008, the patient consulted our service for pain and **deformity** in her left **foot**.</s><s>On physical examination, a widen
(2) w.</s><s>Pancreatic disease.</s><s>Inflammation and **malformation** of the **pancreas** are readily identified by ultrasound, as are
(3) in the functional, and the esthetic rhino surgery.</s><s>The **deformities** of the **nasal septum** might be localized in the bone or in the ca
(4) base.</s><s>Individuals who have been diagnosed with **malformations** of the **corpus callosum**.</s><s>Schizophrenia is a debil

On the other hand, in the pattern '*modifiers of deformity/malformation*', the relevant collocate is a noun (phrase) or an adjective positioned within the range of one or two items (if it is a terminological syntagm such as *ductal plate*) to the left of the KWIC. Put differently, it takes the position of one (or possibly two words, if syntagmatic) occurring to the left of the keyword, as in (5)–(8) below.

(5) the base of the big toe.</s><s>The medical name for this **toe deformity** is hallux valgus.</s></s>The mean corrections in this study
(6) ally in the neonatal period, due primarily to severe
brain malformations.</s></s>Growth in the uterus is slow and the head is dis
(7) Longer term studies are warranted.<s><s>**Nasal septal deformities** in chronic rhinosinusitis patients: clinical and radiological asp
(8) as mild, moderate and severe), presence of **ductal plate malformation** is associated with a significantly poorer clinical outcome.</s><s> Results: The proporti

Despite the convenience of the automatic procedure described above, some manual manipulation of the results was still necessary. Specifically, it included a close inspection of the meaning and use of the retrieved collocates in the given co-text, as it appeared that some of them do not constitute part of medical vocabulary, that is, they do not represent a body part noun or related adjective or are too broad in meaning (e.g. *severe*, *cosmetic*, *system*, *structure*, *body*, *organ*, etc.), or simply because they denote various conditions themselves (e.g. *hallux valgus*, *hallux varus*, *flatfoot*, *Charcot foot*, *scoliosis*, etc.). To this end, the icon next to each collocate was first used to access the KWIC concordances, inspect the co-text of the given collocate, and check for its explicit reference to humans, as well as its

"medicalness" (Panocová, 2017: 73). Further relevance of each collocate to the research as well as the "common denominator" for each set of collocates were established by referring to the *Concise Medical Dictionary* (CMD, 2010) and especially to expert knowledge. To be specific, a specialist medically-trained informant (see Acknowledgements) was consulted on determining the quality shared by most (if not all) collocates of a particular set. That is, the informant was presented with the lists of relevant collocates from all three subcorpora in context and asked to find a major common denominator for those lexical items which exclusively or predominantly occur with *deformity* and *malformation*, respectively. The consultation with the expert about the meaning and use of those collocates attested to co-occur with both nouns as well as about the nature of their relationship with the two terms expressed by means of the statistical association measures (see the next paragraph) proved particularly helpful.

Consider, for instance, the collocate *neck*, which is obtained as part of the collocational pattern '*deformity/malformation of*'. Within this pattern, it is attested to collocate with *deformity* only. The frequency of its co-occurrence with *deformity* in the whole corpus is 10. However, if the concordance lines for this particular combination are displayed and sorted by the right context, it immediately becomes clear that one of them has to be removed from further consideration because the collocate is outside the collocational range specified. If the remaining concordances are further qualitatively or semantically analyzed, it appears that three more concordances have to be discarded, as the polysemous word *neck* therein contained refers to the *femoral neck* or that "narrowed end of the femur" (CMD, 2010: 273), and not to the part of the human body that connects the head with the trunk. Under this analysis, the number of concordance lines for the collocate *neck* being relevant to the research is 6. In addition to such in-depth qualitative analyses of the retrieved collocates, quantitative analyses were conducted to calculate the total number of co-occurrence of the relevant collocates of *deformity* and *malformation*, respectively, and in particular of those collocates attested to co-occur with both nouns. With regard to the latter, two statistical association measures, namely the *logDice* score and the *T-score*, were additionally used for determining more accurately the exclusivity or frequency of these combinations in the EW15 (regardless of the collocational pattern) and, therefore, for accommodating those collocates (the application and interpretation of these association measures are further discussed in Subsection 4.3.). The rationale behind choosing the *logDice* score and the *T-score* over more widely used *MI* score is the fact that the *logDice* score and the *T-score* highlight "exclusive but not necessarily rare combinations" and "frequent combinations of words", respectively, whereas the latter "favour[s] low-frequency collocations" (Gablasova, Brezina, & McEnery, 2017: 162, 163; cf. McGee, 2006: 119, who writes that "the highest MI scores are actually for very infrequent collocations"). Further, features that make the *logDice* score "directly comparable across different corpora and somewhat preferable to the MI-score [or *T-score*], neither of which have a fixed

maximum value", include the fact that it is a standardized measure with the highest value of 14, as well as the fact that it is independent from corpus size (Gablasova et al., 2017: 164). Consequently, the *logDice* score allows us to "see more clearly [...] how far the value for a particular combination is from the theoretical maximum, which marks an entirely exclusive combination" (Gablasova et al., 2017: 164).

4. RESULTS AND DISCUSSION

4.1. 'Modifiers of *deformity/malformation*' subcorpus

In this subcorpus, which provided the longest list of collocates of the three subcorpora or 92 different items in total,[2] I first removed from further consideration those nouns or noun phrases and adjectives whose semantics does not satisfy the requirements specified in Section 3, that is, those nominal and adjectival lexical items which do not denote part of the human body, which are too general in meaning (e.g. *truncular*), or those which refer to various conditions themselves (e.g. *hydrocephalus*). Accordingly, the collocates excluded from the list are: *hallux valgus*, *hallux varus*, *flatfoot*, *flexion*, *club foot*, *bunion*, *kyphotic*, *boutonnière*, *claw toe*, *beak*, *hideous*, *scoliosis*, *equinus*, *hammer toe*, *Charcot*, *flexural*, *rotational*, *scoliotic*, *pescavus*, *crippling*, *contracture*, *postural*, *cosmetic*, *cleft* (*palate*, *lip*), *angular*, *deformity*, *acquired*, *severe*, *skeletal*, *abnormality*, *congenital*, *fetal*, *malformation*, *Chiari-like*, *Dandy-Walker*, *cystic*, *adenomatoid*, *Galen*, *aneurysm*, *aneurysmal*, *Arnold-Chiari*, *cavernous*, *hydrocephalus*, *macrocephally-capillary*, *tract*, and *truncular*. Consequently, the lexical items attested to collocate exclusively with *deformity* include: ***chest* (*wall*, *cage*)** (155),[3] *dentofacial* (83), ***forefoot*** (33), ***hallux*** (25) (or *the big toe* [CMD, 2010: 328]), *jaw* (53), *nasal* (*septal*) (120), ***pectus*** (43) (or *the chest* [CMD, 2010: 548]), *penile* (29), ***spine*** (175), and ***toe*** (204). Some examples of their use in context are given in (9)–(11). A further qualitative analysis of these 10 items reveals that more than half of them (boldfaced) form part of the human musculoskeletal system, which provides mechanical support and movement to the human body as well as protection for its vital or internal organs, or are related to the hard(-tissue) structures (*dentofacial*, *nasal* (*septal*), and *jaw*), except for the adjective *penile*, which relates to the soft (erectile) tissue. With reference to this conclusion, it is interesting to observe that many of the items excluded from this subcorpus, which

[2] The spelling variants (e.g. *fetal* and *foetal*, *arterio-venous* and *arteriovenous*) as well as upper and lower case differences (e.g. *congenital* and *Congenital*) are counted only once.

[3] The figure in brackets indicates the frequency of occurrence of the collocate with the keyword within the specified range of the selected collocational pattern in the whole corpus. The lists are arranged alphabetically.

refer to conditions and collocate exclusively with *deformity*, can also be associated with parts of the musculoskeletal system, namely the toes and the spine.

(9) airo.<s><s>This is an operation to correct the severe **toe deformities** which occur in the feet of people with rheumatoid arthritis a
(10) /<s><s>A variety of inconsistent anomalies including **spine deformities**, cardiac malformations, anomalies of the genitourinary sys
(11) ter?<s><s>Bunions are one of the most common **forefoot deformities** and can be very painful.</s><s>Bunion pain can be very un

The collocates exclusive to *malformation* in this subcorpus include the following nouns or nominal phrases and adjectives: *anorectal* (199), *arterial* (11), *arterio* (-)*venous* (*AV*) (1,511), *capillary* (144), *cerebellar* (12), *cerebral* (26), *cerebrovascular* (21), *cortical* (55), *ductal plate* (15), *dural* (1), *foregut* (2), *fossa* (12), *genitourinary* (15), *intracranial* (4), *lymphatic* (175), *pancreatic* (4), *pulmonary* (*airway*) (33), *urogenital* (20), *uterine* (21), *vascular* (1,165), and *venous* (326). A semantic analysis of these lexical items shows that almost all of them denote or are related to (parts of) the internal organs or soft(-tissue) structures of the human body, with the exception of *fossa* ("[…] a hollow" [CMD, 2010: 288]), which here refers to the *posterior cranial fossa* or a bony structure of the cranial cavity in which the soft structures (brainstem and cerebellum) are located. Examples illustrating the use of some of the collocates exclusive to *malformation* are given in (12)–(14).

(12) umors of the pancreas, intrapancreatic metastasis,
pancreatic malformations and abnormalities.</s><s>The clinical and pathological charac
(13) evelop postnatal-onset microcephaly and have **cerebral malformations** that include hypogenesis of the corpus callosum and poly
(14) od vessel type they contain.</s> <s> The main
vascular malformations are:</s> <s>Capillary malformations – also known as po

In between these two groups are collocates common to both *deformity* and *malformation*, therefore constituting the so-called "gray area" of the subcorpus. It must be remarked, however, that this group is not homogeneous, either, in that its members also show an inclination towards one of the two nouns. That is, at the one extreme of this "gray area" continuum or much closer to the green group of collocates (which is why they are originally color-coded light green and why their frequencies of occurrence with *deformity* are comparable to those of the collocates attested to combine exclusively with *deformity*) are the following items: *bone* (358 : 52), *bony* (98 : 6), *extremity* (33 : 3), *facial* (1,027 : 82), *foot* (755 : 26), *skull* (58 : 12), and *spinal* (1,142 : 23). A further analysis of their semantics reveals that almost all of them form part of the musculoskeletal

system, with the exception of the adjective *facial* and the noun *skull* (a bony structure). Some examples of their use in context are given in (15)–(16). At the opposite extreme of this continuum, that is, much closer to the group of collocates exclusive to *malformation* (the reason why they are originally color-coded light red) are the following items: *brain* (15 : 387), *cardiac* (7 : 220), *genital* (17 : 491), and *ocular* (4 : 30), all of which denote or are related to the internal organs of the human body or its soft(-tissue) structures. Finally, collocates color-coded entirely gray within this continuum include: *cranial* (48 : 12), *craniofacial* (105 : 103), *limb* (207 : 92), and *vertebral* (64 : 23). With reference to this as well as other "fuzzy" areas discussed in the paper, see Table 1 below for an alphabetical list of all items attested to collocate with both *deformity* and *malformation* in the three subcorpora (some sharp differences are bolded).

(15) f complex foot problems, including pediatric and adult **foot deformities.**</s><s> The Yale-New Haven Hospital Primary Care Cent
(16) that affect the feet, such as diabetes, osteoarthritis, **foot malformations,** calluses, corns, bunions, hammer toes, ulcers and woun

4.2. '*Deformity/malformation of*' subcorpus

From the preliminary list of 81 different collocates in this subcorpus, I first excluded those items which do not fulfill the criteria for a noun (phrase) to be considered relevant in this research. Such items refer to instances of non-medical vocabulary, cover terms for organ systems of the human body, names of various conditions and the like (e.g. *sin*, *character*, *body*, *degree*, *skeleton*, *part*, *identity*, *system*, *CNS*, *tract*, *structure*, *development*, *baby*, (*young*, *new*, etc.) *leaves*, *organ*, *embryo*, *fetus*). Hence, the following nouns and nominal phrases are attested to collocate with *deformity* only: (*nasal*) *septum* (9), ***ankle*** (31), ***arm*** (16), ***ball***[4] (3), *breast* (7) (but see the entry *breast 2.* in the CMD, 2010: 96), ***chest*** (29), ***elbow*** (8), *eyelid* (8), ***femur*** (13), ***finger*** (37), ***forefoot*** (8), ***hip*** (23), ***knee*** (56), ***leg*** (37), *mouth* (9), ***neck*** (6), ***pelvis***[5] (7), ***shoulder*** (12), ***tendon*** (6), ***thumb*** (10), ***tibia*** (11), ***toe*** (49), and ***wrist*** (13). A further qualitative analysis of these lexical items indicates that the great majority of them denote parts of the musculoskeletal system (boldfaced), thus confirming the conclusion relative to those collocates of the green group in the previous subcorpus. Exceptions include: (*nasal*) *septum* (a

[4] It refers to *the head of the femur*, which is *ball*-shaped (CMD, 2010: 342).
[5] A relatively low frequency of this collocate with *deformity* might be accounted for by the fact that the collocates *hip* and *femur*, which are often used synonymously with *pelvis* (see the entry *hip* in the CMD, 2010: 342), have fairly high frequencies.

bony structure), *breast*[6], *eyelid*, and *mouth* (by which the jaws that "form the framework of the mouth" [CMD, 2010: 393] are meant in almost all of the concordances). However, their respective frequencies of occurrence with *deformity* are relatively insignificant (all below 10), especially when compared to those of *knee* or *toe*.

As concerns the collocates exclusive to *malformation*, the subcorpus gives the following nouns or noun phrases: (*adipose*, *(hard)dental*, *vascular*, etc.) *tissue* (8), (*bile*, *hepatic*, *thoracic*) *duct* (6), (*blood*, *lymphatic*) *vessel* (33), (*neural*) *tube* (6), (*oral*, *chest*) *cavity* (7), *artery* (8), *cochlea* (5), *corpus callosum* (15), *cortex* (9), *eye* (28), *genitalia* (10), *heart* (64), *kidney* (17), *lung* (22), *lymphatics* (6), *mandible* (8), *pancreas* (11), *spinal cord* (6), *tooth* (25), *uterus* (7), and *vein* (14). Similarly to those collocates constituting the red group in the first subcorpus, the significant majority of these lexical items denote (parts of) the internal organs or soft(-tissue) structures of the human body, excepting (*hard*) *dental tissue*, *tooth*, (*oral*,[7] *chest*) *cavity*, *mandible* (the only movable bone of the skull [CMD, 2010: 676]), and *cochlea* (one of the bony parts of the inner ear [CMD, 2010: 526]).

Finally, the collocates common to both *deformity* and *malformation*, similarly to those from the previous subcorpus, form a continuum of their own. At one end of this continuum (originally color-coded light green, since they display fairly strong tendencies to co-occur with *deformity*) are the nouns: (*acromioclavicular, ankle*, *elbow*, *finger*, *hip*, *interphalangeal*, *knee*, *limb*, *MCP* (*metacarpophalangeal*), *shoulder*, *toe*, *wrist*, etc.) *joint* (103 : 5), *foot* (153 : 13), *nose* (26 : 6), and *spine* (131 : 18), and at the other, the noun *brain* (9 : 120) (originally color-coded light red, since it shows a marked tendency to collocate with *malformation*). In between the two ends, belonging to the "gray area" proper, are located: (*abdominal*, *chest*, *nose*) *wall* (*of the thorax*) (11 : 9), (*AV*, *heart*, *mitral*, *pulmonary*, *tricuspid*) *valve* (6 : 13), (*spinal*, *vertebral*) *column* (13 : 9), *bone* (58 : 34), *ear* (35 : 30), *extremity* (18 : 8), *face* (67 : 24), *hand* (62 : 18), *head* (14 : 41)[8], *jaw* (11 : 9), *limb* (55 : 29), *penis* (15 : 11), *skin* (8 : 17), *skull* (30 : 25), and (*cervical*, *lumbar*, *spinal*, etc.) *vertebra* (6 : 6).

[6] One possible explanation as to why *deformity*, but not *malformation*, is used with this specific collocate may be the fact that in all these concordances the deformity of the breast occurred as a consequence of the mechanical factor involved in the treatment of breast cancer (i.e. breast cancer surgery), removal of breast tissue, and consequent hardening of breast tissue (cf. CMD, 2010: 97).

[7] The oral cavity (or *the mouth*, CMD, 2010: 522) contains the hard tissue (e.g. tooth enamel), too.

[8] A careful inspection of the concordances conducted in consultation with the informant showed that *head* in combination with *malformation* was primarily used to refer to the soft structures of the head, while *head* in combination with *deformity* was used in reference to the external (bony) structures of the head such as *the zygomatic bone*.

4.3. '*Deformity/malformation in*' subcorpus

The last subcorpus I examined concerns the collocations of the type *'deformity/malformation in'* plus a noun (phrase) denoting some part of the human body. As specified in Section 3, such nominal collocates are searched for within the range of four words to the right of the KWIC. Similarly to the previous two subcorpora, this one also gave three different groups of collocates, that is, the green, gray, and red group, totaling 55 different items. However, some of them were removed from further consideration according to the criteria defined above for a collocate to be considered relevant. They are: *scoliosis*, *larva*, *fish*, *arthritis*, *garb*, *bird*, *organism*, *species*, *plane*, *generation*, *shape*, *stage*, *people*, *addition*, *body*, *patient*, *adult*, *baby*, *child*, *population*, *frog*, *structure*, *infant*, *embryo*, *newborn*, *fetus*, *animal*, *offspring*, *woman*, *mouse*, *human*, *pregnancy*, *girl*, *rat*, *male*, *organ*, *syndrome*, and *boy*. The group of collocates exclusive to *deformity* therefore consists of the following items: ***arm*** (9), ***bone*** (9), *breast* (6), ***finger*** (8), ***foot*** (39), ***hand*** (23), ***joint*** (23), ***knee*** (7), ***leg*** (22), ***limb*** (18), ***spine*** (14), and ***toe*** (7). A semantic analysis of these 12 lexemes shows that almost all of them form part of the musculoskeletal system (boldfaced), with the exception of *breast* which is, according to the concordances provided in the subcorpus, likely to suffer from *deformities* usually after (breast cancer) surgery and radiotherapy or some other kind of reconstruction. Collocates exclusive to *malformation* in this subcorpus are: (*frontal*, *hepatic*) *lobe* (4), *brain* (33), *eye* (6), *lung* (10), and *skin* (6). Not unexpectedly, all of them denote either (parts of) the internal organs of the human body or its soft(-tissue) structures. Finally, no relevant collocates common to both *deformity* and *malformation* are attested in the "gray area" of this subcorpus.

collocates	*deformity*	*malformation*
bone	416	86
bony	**98**	**6**
brain	**24**	**507**
cardiac	**7**	**220**
(spinal, vertebral) column	13	9
cranial	48	12
craniofacial	105	103
ear	35	30
extremity	51	11
face	67	24
facial	**1,027**	**82**
foot	**908**	**39**
genital	**17**	**491**
hand	62	18
head	14	41
jaw	11	9
joint	**103**	**5**

limb	262	121
nose	26	6
ocular	**4**	**30**
penis	15	11
skin	8	17
skull	88	37
spinal	**1,142**	**23**
spine	**131**	**18**
(AV, heart, mitral, pulmonary, tricuspid) valve	6	13
vertebra	6	6
vertebral	64	23
(abdominal, chest, nose) wall (of the thorax)	11	9

Table 1. Items attested to collocate with both *deformity* and *malformation* in the EW15[9]

If the results for collocates of *deformity* and *malformation* obtained from the green and red areas above are summarized (keeping in mind the results from Table 1), it can be concluded that, in all three collocational patterns, *deformity* tends to co-occur with nominal and related adjectival items which denote parts of the human musculoskeletal system (given in boldfaced type in Table 2) or refer to the hard(-tissue) structures (i.e. the bony structures of the human body such as *dentofacial*, *nasal (septal)*, *jaw*, etc.), while *malformation* has a tendency to co-occur with nouns or noun phrases and related adjectives denoting either (parts of) the internal organs of the human body or its soft(-tissue) structures (given in italics in Table 2).

deformity	*malformation*
ankle	(*adipose*, dental, *vascular*, etc.) *tissue*
arm	(*blood*, *lymphatic*) *vessel*
ball (the femoral head)	(*frontal*, *hepatic*) *lobe*
breast	(*oral*, *chest*) cavity
chest (wall, cage)	*anorectal*
dentofacial	*arterial*
elbow	arterio(-)venous (*AV*)
eyelid	*artery*
femur	*corpus callosum*
finger	*capillary*
forefoot	*cerebellar*
hallux	*cerebral*

[9] As evidenced by the figures in Table 1, these items also tend to co-occur with one of the two nouns. The figures given in the table do not include those of collocates which also appear in the areas color-coded green or red of the other subcorpora. For instance, *hand* is attested to occur 23 times with *deformity* only in the third subcorpus, which means that this number is not represented in the table. If those figures were added, then the total number of times such collocates occur with the two nouns would be as follows: *bone* **425** : 86; *brain* 24 : **540**; *foot* **947** : 39; *hand* **85** : 18; *jaw* **64** : 9; *joint* **126** : 5; *limb* **280** : 121; *skin* 8 : **23**; *spine* **320** : 18.

hip	*cerebrovascular*
knee	cochlea
leg	*cortex*
mouth	*cortical*
nasal (septal)	(*bile*, *thoracic*) *duct*
(nasal) septum	*ductal plate*
neck	*dural*
pectus	eye
pelvis	*foregut*
penile	fossa
shoulder	*genitalia*
tendon	*genitourinary*
thumb	*heart*
tibia	*intracranial*
toe	*kidney*
wrist	*lung*
	lymphatic(s)
	mandible
	(*neural*) *tube*
	pancreas
	pancreatic
	pulmonary (*airway*)
	spinal cord
	tooth
	urogenital
	uterine
	uterus
	vascular
	vein
	venous

Table 2. Items attested to collocate exclusively with *deformity* or *malformation* in the three subcorpora of the EW15

As mentioned in the previous section, to try to accommodate those items attested to collocate with both *deformity* and *malformation* in the three subcorpora and therefore further reinforce the observed tendencies, the *logDice* score and the *T-score* were additionally calculated in the EW15, regardless of the collocational pattern (Table 3). In calculating these scores, that is, in identifying the collocability between the items from Table 1 and the nouns *deformity* and *malformation* using the two measures, the following criteria were applied: lemma (lowercase), range -2 -1 KWIC 1 2 3 4, minimum frequency in corpus 1, minimum frequency in given range 1. It must be noted, however, that the results are available only for the first 1,000 collocation candidates (the items for which the score values were unavailable are marked by *x* in Table 3).

collocates	*deformity*		*malformation*	
	T-score	*logDice*	*T-score*	*logDice*
bone	**24.25**	**4.99**	10.43	2.59
bony	**10.24**	**6.33**	3.46	3.52
brain	x	x	**27.97**	**4.45**
cardiac	3.80	1.39	**16.24**	**5.59**
chest wall	2.83	3.35	x	x
column	x	x	x	x
cranial	7.55	5.50	5.47	4.91
cranio(-)facial	15.58	11.07	14.36	10.83
ear	13.35	3.49	10.39	2.79
extremity	**7.41**	**4.80**	3.99	3.23
face	x	x	x	x
facial	31.50	7.45	10.38	4.32
foot	**31.65**	**4.09**	x	x
genital	4.46	3.34	**9.21**	**5.63**
hand	x	x	x	x
head	x	x	x	x
jaw	**9.62**	**4.29**	4.44	2.16
joint	**24.08**	**3.93**	x	x
limb	**21.06**	**6.33**	13.18	5.05
nose	9.91	3.57	x	x
ocular	2.81	2.06	**6.16**	**4.53**
penis	4.98	3.27	2.81	1.78
skin	9.38	1.27	x	x
skull	10.71	4.78	7.40	3.81
spinal	**37.76**	**8.21**	11.73	4.92
spine	**21.25**	**6.53**	6.06	3.00
valve	x	x	5.79	2.35
vertebra	**4.79**	**3.93**	3.31	3.14
vertebral	**11.13**	**6.61**	7.41	5.77

Table 3. The *T-score* and the *logDice* score as given by the EW15

Considering the *T-score* values in Table 3, it can be concluded that they confirm the tendencies observed earlier in the research analysis. Specifically, these values strongly indicate that combinations of the collocates forming part of the musculoskeletal system such as *bone*, *bony*, *extremity*, *foot*, *joint*, *limb*, *spinal*, *spine*, *vertebra*, *vertebral* and the noun *deformity* are more frequent than those of the same collocates and the noun *malformation*. The *T-score* values for combinations of *deformity* and the collocates such as *cranial*, *cranio(-)facial*, *ear*, *facial*, *jaw*, *nose*, *skull*, *chest-wall*, by which the bony structures of the human body are primarily referred to, are also higher than the values for combinations of these same collocates and *malformation*, implying their greater frequency and further reinforcing the tendencies observed earlier in the research. As further evidenced by Table 3, the combinations of collocates denoting or relating to the

internal organs of the human body or its soft(-tissue) structures such as *brain*, *cardiac*, *genital*, *ocular*, *valve* and the noun *malformation* are more frequent than those of the same collocates and *deformity*. The only two exceptions to these tendencies seem to be the collocate *penis*, whose combination with *malformation* is attested as less frequent than the one with *deformity*, and the collocate *skin*, which was not found among the first 1,000 collocation candidates of *malformation*, but only among the first 1,000 collocation candidates of *deformity*[10] (the *T-score* at 9.38 is rather low, though). That these tendencies are quite strong is further demonstrated by the *logDice* score values in Table 3, as they attest to the typicality of the above-mentioned combinations, that is, they serve as evidence of the extent to which the two (or more) words occur predominantly in each other's company in the given corpus.

5. PEDAGOGICAL IMPLICATIONS

It is argued in Section 1 of this paper that the corpus-based findings presented here may have some implications for EMP teaching/learning, especially in those cases where teachers or learners do not have free or full access to (large) electronic corpora. Namely, though the obtained results are seen as tendencies, they nonetheless may have important practical pedagogical implications for not only more effective teaching and learning of medical English vocabulary or raising learners' awareness of the importance of collocational knowledge, but also for pedagogical (specialized) lexicography of English. That is, they could be used as authentic language data for materials development such as compiling or updating collocation lists for EMP students, creating content of coursebooks for medical purposes, as well as for improving the dictionary entries for the two terms.

For example, by making the obtained concordance data an integral part of coursebooks or other teaching/learning materials (e.g. practice activities such as gap-filling or matching exercises), EMP learners are presented with the actual usage of the two terms in authentic medical (con)texts. Furthermore, integrating the tendencies observed here into the process of materials development, that is, presenting learners with the two nouns within their exclusive or frequent and typical lexico-semantic environments, may facilitate the acquisition of these collocations, therefore promoting their more active use. Specifically, when introducing learners to the appropriate use of the two terms, teachers should first inform learners of the fact that the nouns *deformity* and *malformation*, like many other words, tend to favor specific lexico-semantic environments. That is, when

[10] If the same criteria for calculating the *logDice* score are applied in the MWC, the obtained value for the combination between *penis* and *malformation* is stronger than the one with *deformity* (5.88 : 5.20). Unfortunately, the *logDice* score values for the combinations of *skin* and the two nouns are not available among the first 1,000 collocation candidates in this corpus.

deciding which of the two nouns to use within a given context, EMP learners should first carefully consider the given collocate by determining whether it denotes part of the human musculoskeletal system or a hard(-tissue) structure, or it refers to (part of) the internal organs of the human body and its soft(-tissue) structures. In case the lexical item denotes part of the human musculoskeletal system or relates to a hard(-tissue) structure, *deformity* is the preferred choice, with the exception of (*hard*) *dental tissue*, (*oral*, *chest*) *cavity*, *cochlea*, *fossa*, *mandible*, and *tooth*, which are attested as exclusive collocates of *malformation* within the three patterns. On the other hand, if the lexical item refers to (part of) the internal organs of the human body and its soft(-tissue) structures, preference should be given to *malformation*, with the exception of collocates *breast*, *eyelid*, *penile*, and *penis*, whose occurrences with *deformity* are recorded as exclusive or more frequent in the three subcorpora.

Regarding the possible pedagogical lexicographical use of the results of the present research, it is worth remembering that pedagogically-oriented dictionaries should, *inter alia*, make active use of corpora to meet the (vocabulary) needs of their users (Fuertes-Olivera & Arribas-Baño, 2008: 138). More specifically, the obtained results could be utilized for providing more precise definitions or usage notes of the two terms, as well as for their contextualizing at the level of collocations or example sentences. For instance, the identified collocations or concordance data could be used by lexicographers as illustrative or "live" examples (Fuertes-Olivera & Arribas-Baño, 2008: 138) for providing valuable information about the two nouns within their frequent or typical, as well as exclusive contexts of usage.

6. CONCLUDING REMARKS

In this corpus-based research, I made an attempt not only to compile lists of some of the significant collocates of the medical terms *deformity* and *malformation* within the three collocational patterns ('*modifiers of deformity/malformation*', '*deformity/malformation of*', and '*deformity/malformation in*'), but also to determine which qualities make some of these items combine exclusively or predominantly with *deformity* and *malformation*, respectively. The results of the qualitative and quantitative analyses of a number of the retrieved collocates suggest that the attested word combinations are best seen as a continuum of cases, with some collocates being exclusive or closer to the *deformity* end of the continuum, the others being exclusive or closer to the *malformation* end of the continuum, and some collocates showing both tendencies. Specifically, it has been shown that, regardless of the collocational pattern, *deformity* typically collocates with nouns and related adjectives which denote part of the human musculoskeletal system or refer to the hard(-tissue) structures of the human body, whereas the typical collocates of *malformation* are nouns or noun phrases

and adjectives denoting or relating to (parts of) the internal organs of the human body and its soft(-tissue) structures, with few exceptions. The number of exceptions has been significantly reduced by adopting the two collocation measuring methods, namely the *T-score* and the *LogDice score*, thereby making the findings more transparent, conclusive, or replicable, and further facilitating the process of teaching/learning these word associations. Although the findings should be conceptualized as tendencies, they still represent an up-to-date and authentic body of evidence of the ways the terms *deformity* and *malformation* are used within the EMP context. Finally, as the source of the collocational patterns and collocates used here was only one corpus – the *English Web Corpus 2015*, it would therefore be both beneficial and interesting for further research to compare the collocability or behavior of the two terms within the same or additional collocational environments across different (specialized) corpora, with a view to learning whether or not the results obtained therein confirm the findings and conclusions of this paper.

[Paper submitted 1 Jun 2020]
[Revised version received 21 Aug 2020]
[Revised version accepted for publication 15 May 2021]

Acknowledgements

I wish to thank Tanja Tomić, a senior physiotherapist at the Clinic of Rehabilitation "Dr Miroslav Zotović" in Belgrade, for her invaluable help in qualitatively analyzing the data and also for stimulating comments and discussions.

Sources

English web corpus 2015 (EW15). (n.d.). Retrieved from https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fententen15_tt21

Martin, E. A. (Ed.) (2010). *Concise medical dictionary* (*CMD*) (8th ed.). Oxford University Press.

Medical web corpus (MWC). (n.d.). Retrieved from https://app.sketchengine.eu/#dashboard?corpname=preloaded%2Fweb_med_1

Sketch engine (SE). (n.d.). Retrieved from https://www.sketchengine.eu/

References

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

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. https://doi.org/10.1016/j.esp.2007.04.003

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213-238. Retrieved from https://www.victoria.ac.nz/__data/assets/pdf_file/0020/1626131/Coxhead-2000.pdf

Coxhead, A. (2002). The academic word list: A corpus-based word list for academic purposes. In B. Ketteman, & G. Marks (Eds.), *Teaching and language corpora (TALC) 2000 conference proceedings* (pp. 72-89). Rodopi. https://doi.org/10.1163/9789004334236_008

Coxhead, A. (2018). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives* (1st ed.). Routledge.

Coxhead, A., & Demecheleer, M. (2018). Investigating the technical vocabulary of plumbing. *English for Specific Purposes*, *51*, 84-97. https://doi.org/10.1016/j.esp.2018.03.006

Coxhead, A., & Quero, B. (2015). Investigating a science vocabulary list in university medical textbooks. *TESOLANZ Journal*, *23*, 55-65. Retrieved from https://mk0tesolanzju6r70d8p.kinstacdn.com/wp-content/uploads/2019/10/TESOLANZ_Journal_Vol23_2015.pdf

Durrant, P. (2009). Investigating the viability of a collocation list for students of English for academic purposes. *English for Specific Purposes*, *28*(3), 157-169. https://doi.org/10.1016/j.esp.2009.02.002

Fuertes-Olivera, P. A., & Arribas-Baño, A. (2008). *Pedagogical specialised lexicography: The representation of meaning in English and Spanish business dictionaries*. John Benjamins Publishing Company.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning*, *67*, 155-179. https://doi.org/10.1111/lang.12225

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics*, *35*(3), 305-327. https://doi.org/10.1093/applin/amt015

Gilmore, A., & Millar, N. (2018). The language of civil engineering research articles: A corpus-based approach. *English for Specific Purposes*, *51*, 1-17. https://doi.org/10.1016/j.esp.2018.02.002

Gries, S. Th. (2006). Introduction. In S. Th. Gries, & A. Stefanowitsch (Eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis* (pp. 1-17). De Gruyter Mouton.

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: The establishment of a medical word list. *Language Teaching Research*, *17*(4), 454-484. https://doi.org/10.1177/1362168813494121

Hsu, W. (2018). The most frequent BNC/COCA mid- and low-frequency word families in English-medium traditional Chinese medicine (TCM) textbooks. *English for Specific Purposes*, *51*, 98-110. https://doi.org/10.1016/j.esp.2018.04.001

Kaunisto, M. (2007). *Variation and change in the lexicon: A corpus-based analysis of adjectives in English ending in -ic and -ical*. Rodopi.

Le, C. N. N., & Miller, J. (2020). A corpus-based list of commonly used English medical morphemes for students learning English for specific purposes. *English for Specific Purposes*, *58*, 102-121. https://doi.org/10.1016/j.esp.2020.01.004

Lei, L., & Liu, D. (2016). A new medical academic word list: A corpus-based study with enhanced methodology. *Journal of English for Academic Purposes*, *22*, 42-53. https://doi.org/10.1016/j.jeap.2016.01.008

McGee, I. D. (2006). *Lexical intuitions and collocation patterns in corpora* (Unpublished doctoral dissertation). Centre for Language and Communication Research, School of English, Communication and Philosophy, Cardiff University, Cardiff, UK.

Muñoz, V. L. (2015). The vocabulary of agriculture semi-popularization articles in English: A corpus-based study. *English for Specific Purposes*, *39*, 26-44. https://doi.org/10.1016/j.esp.2015.04.001

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company.

Nelson, M. (2006). Semantic associations in business English: A corpus-based analysis. *English for Specific Purposes*, *25*(2), 217-234. https://doi.org/10.1016/j.esp.2005.02.008

Panocová, R. (2016). A descriptive approach to medical English vocabulary. In T. Margalitadze, & G. Meladze (Eds.), *Proceedings of the XVII EURALEX International congress: Lexicography and linguistic diversity* (pp. 529-540). Ivane Javakhisvili Tbilisi University Press. Retrieved from https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202016/euralex_2016_058_p529.pdf

Panocová, R. (2017). *The vocabulary of medical English: A corpus-based study*. Cambridge Scholars Publishing.

Peacock, M. (2012). High-frequency collocations of nouns in research articles across eight disciplines. *Ibérica*, *23*, 29-46.
Retrieved from https://www.redalyc.org/pdf/2870/287024475003.pdf

Quero, B. (2015). *Estimating the vocabulary size of L1 Spanish ESP learners and the vocabulary load of medical textbook* (Unpublished doctoral dissertation). University of Wellington, Wellington, New Zealand.

Quero, B., & Coxhead, A. (2018). Using a corpus-based approach to select medical vocabulary for an ESP course: The case for high-frequency vocabulary. In Y. Kırkgöz, & K. Dikilitaş (Eds.), *Key issues in English for specific purposes in higher education* (pp. 51-75). Springer.

Riccobono, P. S. (2020). Triangulating diamond talk: Identifying technical spoken vocabulary in English for baseball purposes. *ESP Today*, *8*(2), 114-140. https://doi.org/10.18485/esptoday.2020.8.1.6

Smith, S. (2020). DIY corpora for accounting & finance vocabulary learning. *English for Specific Purposes*, *57*, 1-12. https://doi.org/10.1016/j.esp.2019.08.002

Walker, C. (2011). How a corpus-based study of the factors which influence collocation can help in the teaching of business English. *English for Specific Purposes*, *30*(2), 101-112. https://doi.org/10.1016/j.esp.2010.12.003

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4), 442-458. https://doi.org/10.1016/j.esp.2008.05.003

Ward, J. (2007). Collocation and technicality in EAP engineering. *Journal of English for Academic Purposes*, *6*(1), 18-35. https://doi.org/10.1016/j.jeap.2006.10.001

GORICA TOMIĆ is Research Trainee at the Center for Language and Literature Research at the Faculty of Philology and Arts, University of Kragujevac, Serbia. Her major research interests include English morphology and word-formation, lexicology, lexicography, as well as corpus linguistics. She has published in national and international peer-reviewed journals such as *Lexis: A Journal in English Lexicology*, *Nasleđe*, etc.