Academic Collocations in Egyptian Medical Abstracts: A Corpus Based Study

Waleed S. Mandour

Center of Preparatory Studies, Sultan Qaboos University, Oman
E-posta: w.abumandour@squ.edu.om
Orcid ID: 0000-0002-9262-5993

Research Article Received: 12.03.2022 Accepted: 25.04.2022 Published: 30.04.2022

Abstract:

The present study aims primarily at investigating academic collocations included in research abstracts submitted by Egyptian medical authors in Egypt, in a way of evaluating their academic literacy. Retrieved collocations are compared to the Academic Collocation List (Ackermann & Chen, 2013) that comprises 2,468 academic multi-words written by English-speaking natives. The researcher collected 795 medical abstracts from a renowned medical journal in Egypt which were published over six years (2013-2018). The corpus data renders 216,842 words in cross-sectionally annotated file compilation. Results illustrate the non-native speakers' tendency of including plenty of collocations of statistical connotations rather than those of medical references at the expense of including academic collocational components in other abstract moves. Despite the accepted academic language enclosure, Egyptian authors exclusively use few academic collocations of statistical significance. Thus, that habitual linguistic phenomenon of the non-native writings indicates an epistemological need for learning academic collocations to improve publications' quality.

Keywords:

Medical Discourse, Academic Collocations, Corpus, Academic Purposes.

For **Citation /Atif** İçin / للاستشهاد: Mandour, Waleed. (2022). Academic Collocations in Egyptian Medical Abstracts :A Corpus Based Study. Arap Dilbilimi ve Edebiyatı Dergisi – DÂD. Volume/Cilt 3, Issue/Sayı 5, 268-300 https://www.daadjournal.com/

دراسة المتصاحبات اللغوية الأكاديمية في مستخلصات الأبحاث الطبية المصرية على أسس الذخائر اللغوية

وليد مندور

جامعة السلطان قابوس، عمان

البريد الإلكتروني: w.abumandour@squ.edu.om

معرف (أوركيد): 0000-0002-9262-5993

بحث أصيل الاستلام: ١٢-٣-٢٠٢٢ القبول: ٢٥-٤-٢٠٢٢ النشر: ٣٠-٤-٢٠٢٢

الملخص:

تهدف الدراسة الحالية في المقام الأول إلى تمحيص المتصاحبات اللغوية الأكاديمية المدرجة في المستخلصات البحثية المقدمة من قبل مؤلفين مصريين في المجالات الطبية بمصر، من أجل تقييم المضمون الأكاديمي لديهم، اذ تمت مقارنة نتاج المتصاحبات اللغوية الأكاديمية بقائمة المتصاحبات اللغوية الأكاديمية التي تضم ٢٤٦٨ مُرَكبًا لفظيًّا أكاديميًّا كتبها باحثون يتحدثون الإنجليزية الأم، حيث جمع الباحث ٧٩٥ مستخلصًا طبيًّا من مجلة طبية شهيرة في مصر نُشِرَت على مدار ست سنوات (٢٠١٣-٢٠١٨)، تألفت الذخيرة اللغوية من ٢١٦,٨٤٢ كلمة جُمعت في ملفاتٍ إلكترونية مشروحة لُغويًّا، وتوضح النتائج ميل المتحدثين غير الناطقين بالإنجليزية إلى تضمين الكثير من المتصاحبات ذات الدلالات الإحصائية بدلاً من تلك الخاصة بالأمور الطبية أو المهنية الأكاديمية، وعلى الرغم من استخدام المؤلفين المصريين لكمٍّ مقبول من المفردات الأكاديمية، فإنهم لم يستعملوا سوى القليل من المتصاحبات الأكاديمية بشكلٍ كبير إحصائيًّا، من ثَم تشير هذه الظاهرة اللغوية المعتادة لكتابات المتحدثين غير الناطقين بالإنجليزية إلى حاجتهم المعرفية لتعلم المتصاحبات الأكاديمية لتحسين جودة المنشورات.

الكلمات المفتاحية:

الخطاب الطبي، المتصاحبات اللغوية الأكاديمية، الذخائر اللغوية، الأغراض الأكاديمية.

Mısır Tıbbi Araştırmalarının Özetlerindeki Akademik Dilsel Bağlantıları Dilsel Derleme Temelinde İnceleme

Waleed S. Mandour

Sultan Qaboos Üniversitesi, Umman

E-mail: w.abumandour@squ.edu.om

Orcid ID: 0000-0002-9262-5993

Araştırma Makalası Geliş : 12.03.2022 Kabul : 25.04.2022 Yayin: 30.04.2022

Özet:

Mevcut çalışma, öncelikle Mısır'daki tıp alanlarında Mısırlı yazarlar tarafından sunulan araştırma özetlerinde yer alan akademik dilsel bağıntıları, akademik içeriklerini değerlendirmek için incelemeyi amaçlamaktadır.

Akademik dilsel bağlantıların sonuçları, ana dili İngilizce olan araştırmacılar tarafından yazılan 2468 Akademik Sözlü Bileşik içeren Akademik Dilsel Bağlantılar Listesi ile karşılaştırılmıştır. Araştırmacı tarafından, Mısır'da altı yıllık bir süre boyunca (2013-2018) yayınlanan ünlü bir tıp dergisinden 795 tıbbi alıntı toplanmıştır. İncelenen dilsel derleme, dilsel olarak açıklamalı elektronik dosyalarda toplanan 216.842 kelimeden oluşmaktadır. Sonuçlar, İngilizce bilmeyenlerin tıbbi veya profesyonel akademik konulardan ziyade istatistiksel olarak daha fazla çağrışım içerme eğiliminde olduklarını göstermektedir. Mısırlı yazarlar kabul edilebilir miktarda akademik kelime dağarcığı kullanmalarına rağmen, yalnızca istatistiksel olarak anlamlı birkaç akademik bağıntı kullanmışlardır. İngilizce bilmeyenlerin yazılarındaki bu yaygın dilsel alışkanlık, yayınların kalitesini artırmak için akademik bağıntıları öğrenmeye yönelik bilişsel ihtiyaçları olduğunu göstermektedir.

Anahtar Kelimeler:

Tıbbi Söylem, Akademik Eşdizimler, Derleme, Akademik Amaçlar.

Introduction:

The term *collocation* is concisely defined in the *Oxford Dictionary of Advanced Learner* (Hornby, 2015) as "the habitual juxtaposition of a particular word with another word or words with a frequency greater than chance", marking its lexico-grammatical and semantic features and referring to its traditional identification method of frequency measurement. Other dictionaries explain the term's attributes in more detail. The *Longman Dictionary of Teaching and Applied Linguistics* (Richards, 2013) defines the noun *collocation* and its corresponding verb *collocate* as:

> Collocation refers to the restrictions on how words can be used together, for example, which prepositions are used with particular verbs, or which verbs and nouns are used together. For example, in English the verb *perform* is used with the operation, but not with discussion:
>
> *The doctor performed the operation.*
>
> ** The committee performed a discussion.* Instead, we say:
>
> *The committee held/had a discussion.*
>
> perform is used with (collocates with) operation and *hold* and *have* collocate with *discussion*….

Richard's definition highlights the proper usage of that governs the collocation process. *Cambridge Advanced Learner's Dictionary* (2008) further elucidates that restriction behavior that writer or speakers should follow as:

> [C]… *collocate,* a word or phrase that is often used with another word or phrase, in a way that sounds correct to people who have spoken the language all their lives, but might not be expected from the meaning: In the phrase "a hard frost", "hard" is a collocation of "frost" and "strong" would not sound natural.
>
> [C] the combination of words formed when two or more words are often used together in a way that sounds correct: The phrase "a hard frost" is a collocation.

> [U] the regular use of some words and phrases with others… (Walter, 2008)

On the other hand, "recurrent combinations" or "fixed combinations" are what they are called in the *BBI Combinatory Dictionary of English* (Benson et al., 2010); it further stresses its ubiquity nature when described as "many fixed, identifiable, non-idiomatic phrases and constructions" (Benson, Benson & Ilson, 2010, p. XIX). All definitions underscore the phraseological phenomenon in which collocations establish conventionally used phrases far from being formed accidentally. They also indicate the fact that misplacing word order results in an unnatural flow of language that may, in turn, distort communication. Accordingly, looking into the collocational forms produced by non-native speakers of English should adhere to their selection correctness.

Although the neo-Firthians, Halliday and Sinclair, agreed on redeeming collocations the aspect of being grammar-independent, Sinclair stated that lexical items can be identified within their intrinsic environment, i.e., grammatical signs (as cited in Krishnamurthy, 2006, p. 596). *Cambridge History of the English Language, Vol.1* (Hogg, 1992) states that the lexical field theory has showcased the significance of studying collocations – despite the term *collocation* was not mentioned explicitly. As cited, Jost Trier, who suggested studying the linguistic phenomenon (1931) reasoned the fact that words carry their meaning through establishing syntagmatic relationships with other words within the same word field (Hogg, 1992, pp. 425–430). Hence, one word constricts the meaning of neighboring words in a field fitting neatly together like a *mosaic* (as known as the *mosaic concept*). If single lexis undergoes a *semantic change,* then the entire structure of the lexical field changes.

In fact, the neo-Firthians laid immense emphasis on studying collocations that dominate lexical concordances and proposed coexistence of lexical and syntactical patterns. Sinclair et al. (2004) conceived collocations as fundamental in language theory for they denote semantic preferences and discourse prosodies. They differentiated between two types of collocations: *Significant Collocations* which refer to the co-occurrence of items more frequently than their respective text lengths would anticipate; and *casual collocations* where the co-occurrence phenomenon is considered insignificant as the term reflects (Sinclair et al., 2004, p. 10).

The gravity of conducting studies about collocations in the language learning scope entails looking through the historical background of Palmer's work on which Firth relied in his presentation (Firth, 1957) that became a source of inspiration to many phraseologists, lexico-grammarians, and pedagogists as well (e.g., Cowie, 1998; Granger, 1999; Howarth, 1998; Sinclair, 1991; Sinclair et al., 2004). As cited by Firth (Firth, 1957), Palmer's contribution which was published in Japan (1933) did highlight the importance of collocations in English Language Teaching (ELT) and learning acquisition. He stressed the learners' phraseological knowledge of L1 collocations that leads to native-like production. Similarly, O'Dell & McCarthy (2008) have manifested the naturalness of speaking and writing skills a learner can possess among the top benefits of learning language collocations. An example was given of using the adjective *great* instead of *big* or *high* to collocate with *importance*. It can be noticed by native English speakers that the collocation *great importance* sounds more natural and fluent. Studying a language, as set by Wray (2005, p. 143), involves "not only its individual words but also how they fit together." That collocation attribute, in particular, leads us to confer about what the concept of collocation implies in relation to learner research and pedagogical issues.

Since collocations are considered a cornerstone for English Language Learners, a similar interest is shown in that formulaic language within research publications (see Durrant, 2009; Durrant & Schmitt, 2009; Gledhill, 2000; Hyland & Shaw, 2016; Schmitt, 2012). Durrant & Schmitt postulated the susceptibility of non-native speakers to "over-rely on forms which are [...] common in English" (2009, p. 174). In other words, even advanced learners of English (academic authors of non-native English language herewith) may fail to intuitively and equally provide correct collocations (Hyland & Shaw, 2016; Schmitt, 2012).

Subsequently, this paper tries to shed light on the academic collocational spectrum in a collection of medical abstracts extracted from an Egyptian medical journal (Menoufia Medical Journal, to be named here as *MMJ*). The inspection is performed in comparison with Ackermann & Chen's Academic Collocation List (ACL) which encompasses multi-word combinations from different epistemological domains, including the medical ones (2013). The corpus data includes 795 abstracts that represent 36 medical categories all investigated collectively to evaluate the academic collocation items they use in terms of the near nativity.

Research Questions:

The present study attempts to answer the following questions:

1. What are the academic collocations used by Egyptian medical researchers? And what are the distinct features in terms of lexical and grammatical structures?
2. To what extent do they match the Academic Collocation List (ACL)?
3. What are the areas of strength and others of weaknesses that medical researchers in Egypt can be provided with to sustain their academic literacy and near-nativity production?

Literature Review:

The medical genres in English for Medical Purposes (EMP) are categorized by Ferguson (2013, p. 243ff) into two perspectives: one deals with enhancing English skills for NNS in the medical fields, and the other looks into health care communication. It is not confined to looking into professionals' communications but rather it includes public access to the medical field (e.g., patients) (Ordóñez-López & Edo-Marzá, 2016). Ordóñez-López & Edo-Marzá aregued that specialized discourse such as the medical domain has become unrestricted to practitioners and academicians due to the prominent status of medicine to people establishing a dynamic dialogue between science, including the field of medicine, and society (2016). And according to Swales (2000), medical research findings are communicated via an "open-genre network", such as posters, blogs, authentic medical websites, etc. along with the conventional means of academic communication, e.g., research articles, conference proposals.

In his investigation of the phraseological patterns the medical discourse possesses, Marco (1998) referred to a medical corpus sampled from the *British Medical Journal* which has previously denoted the written medical discourse can be considered a "result of social construction" in addition to reporting scientific findings. He stated that research articles are empowered with persuasion tools (especially editors and referees) which reveal three main qualitative aspects: originality, reliability, and significance of subject matters (Marco, 1998). Ordóñez-López & Edo-Marzá (2016) have added other features after inspecting large medical corpora, spoken and written. These notable features include metaphorical language, exploitation of visuals, such as graphs, pictures & formulas.

Staicu (2017) stressed the fact that the nature of the medical language used as an ESP is influenced with typically two main characteristics: its paradigmatic and syntagmatic aspects due to the lexical-semantic fields as well as the "specific combinatorial features" which are observed in the collocational patterns followed. *Noun* + *adjective*, *noun* + *noun,* and *verb* + *noun* collocational patterns are found the most dominant paradigms in the medical discourse (Staicu, 2017). Nadja Nesselhauf, in her book, *Collocations in the English of Advanced Learners: A Study Based on a Learner Corpus*, concluded that observations in many studies about phraseological patterns in academic writing of L2 authors mostly indicate the fact that "collocation production presents a problem for second language learners". She argued that learners prefer to produce overly fewer collocations than native speakers (Nesselhauf, 2003, p. 8) – a similar observation noted by Schmitt (2012). The latter study proves the epistemological challenges of the non-native production of collocations that may hinder genuine fluency.

A paramount attempt in analyzing the medical academic discourse was done by Wang, Liang & Ge (2008) through a corpus-based study of the most recurrently medical academic vocabulary observed in relevant research articles written by English native speakers. Results revealed a list of 623-word families to be called the Medical Academic Word List (MAWL). That the study has been inspired by Coxhead's *Academic Word List* (2000) pursued strict procedures on revising and refining its data collection and processing the MAWL not only by following criteria carefully but also through consulting discipline specialists (pp. 445-448). Worth mentioning, only 54.9% (nearly half) of the medical list are found in Coxhead's Academic Word List.

Lei (2016) claimed that "all of the existing discipline-specific word lists have been developed using Coxhead's (2000) method that excluded general high-frequency words". Because of the unreasonable account, he, therefore, developed a new medical academic vocabulary list that includes some high-frequent words of the GWL (i.e., General Word List, 1953). In both works, though, medical academic word lists overlap with the 570-word families of the AWL, whereas almost halves in either list contain discipline-specific vocabulary items.

Using language effectively holds a variety of roles that determine the distinctive ways it is presented by the user (Hyland, 2008; 2016; Hyland & Wong, 2019). Based on *Longman Dictionary of Language Teaching &*

Applied Linguistics (Richards, 2013) definitions, English for Specific Purposes (ESP) provides language contents that are "fixed by the specific needs of a particular group of learners, for example, courses in *English for Academic Purposes*, *English for Science and Technology*, and *English for Nursing*." English for Academic Purposes (EAP), on the other hand, refers to English language courses that are designed to "help learners to study, conduct research or teach in English, usually in universities or other post-secondary settings." (Richards & Schmidt, 2010). EAP is marked by Biber (2006) as the English language that can be found in research articles and other related documents. Besides, Hyland and Shaw stressed the *communicative* role of EAP as "it goes beyond preparing learners for study in English to understanding the kinds of literacy found in the academy" (2016, p. 1).

The increasing interest in identifying linguistic features of Academic English over the last 40 years (Hyland, 2006; Hyland & Shaw, 2016; Hyland & Wong, 2019) led researchers (such as Biber, 2006; Coxhead, 2000) to apply quantitative and qualitative methods in describing this specific type of English which marks English as the *Lingua Franca* of ongoing research articles (Coxhead, 2008; Hyland & Shaw, 2016). Despite the existence of having three catalogs of English studies in the sense of *English as a Lingua Franca* (ELF), John Flowerdew has brought a fourth term, namely *English for Research Publication Purposes* (ERPP) (Flowerdew, 2015). He justified the need for the emerging approach as a result of the undergoing interest in "internationally scholarly publication as a field of research" (2015). He exhibited a list of advantages of establishing such a distinct research approach that applies problem-driven methods that tackle EAL writers' issues and, ultimately, serve "ESP researchers and practitioners" (Flowerdew, 2015). To him, ERPP sets rules and practices of "discourse analysis and social constructivism/situated learning" mainly for research articles (RA), whether they are L1 or L2 writers. Subsequently, a distinction between a native-speaker publication and a non-native-speaker one is replaced with the question of having a junior/senior scholar.

Flowerdew's focus through his ERPP theory is to support L1 and L2 writers in fulfilling their research publication, as both show difficulties in this respect. Further, he pinpointed the current and forthcoming roles of "corpora" (see previous section, 1.3) in the ERPP approach. He said "... large corpora for the various discipline written by EAL writers might help to identify what is acceptable in terms of intelligibility in written academic

English and what is not." (Flowerdew 2008). We can, thus, envision the relationship between ESP, EAP & ERPP as in the following diagram.

Had the suggested English of Research Publication Purposes been widely acknowledged so far, the title of this paper could have included the acronym, ERPP. Even though, However, Flowerdew (Hyland & Wong, 2019) argued about the scarcity of ELF academic compositions when he said: "what has been missing […] are studies which examine academic writing of a disciplinary nature" - that is what the current study is trying to do. In his argument, he has supported Tribble's call (2017) for establishing a new EAP paradigm that converges "both worlds" (Hyland & Wong, 2019) where EFL & EAP is brought together under one umbrella of ELFA, the acronym of English as Lingua Franca for Academic Purposes. Their suggested paradigm depends on dichotomies of "Native Speaker (NS) vs. Non-Native Speaker (NNS)" (Tribble, 2017, p. 30) for which either learner corpus research (LLC) or ELF corpus research in the academic disciplinary writing (i.e., EAP). Flowerdew referred to possible benefits to gain from the CAR (Corpus of Research Articles) he argued earlier in his ERPP model since our interest has been further crystallized to involve more 'worlds' henceforth.

Learner research studies agree that non-native writers usually suffer from language challenges, particularly on the phraseological level. Ferguson (2013) and Hyland & Shaw (2016) clearly stated that English research writers need to acquire academic literacy as well as epistemologically deep language use which they are usually poor at either during their previous school study or at their tertiary levels where EAP is taught incompetently through non-specialists – a close description of the current status of Egyptian researchers' case in the medical fields based on an interview held with the MMJ managing editor, Ragab (2018).

According to Ackermann & Chen (2013), the rendered 2,468 most-frequently-used collocations in academia "serve as a new tool in EAP for teaching and learning collocations". It mainly concerns with lexical combinations appearing cross-disciplinarily in 28 academic domains, including health sciences which consist of 1,429,679 words out of almost 37 million in the total corpus (3.8%) as derived from the Pearson International Corpus of Academic English – a corpus exclusively produced by native English speakers (Ackermann & Chen, 2013, p. 5). Two major approaches were adopted to generate the list: corpus-driven and expert-judged. The rigorous procedures of corpus extraction, quantitative and qualitative before the refinement and systematization phases were further validated to ensure its representativeness to the academic formulaic language. However, Ackermann & Chen recommended conducting future research to generate tailored academic collocations exclusively targeting specific fields of study. The final list of academic collocations is classified into five lexical categories (see Table 1 below).

Table 1: Categories in Academic Collocation List (Ackermann & Chen, 2013, p. 13)

Lexical Combination	**No. of Entries**	**%**
adjective + noun	1773	71.8
noun + noun	62	2.5
verb + noun	310	12.6
verb + adjective	30	1.2
verb + adverb	170	6.9
adverb + adjective	124	5.0
Total	**2469**	**100**

There is no research study that investigated linguistic features in the medical discourse of the Egyptian authors (at the time of writing this paper). The current work, accordingly, presents a novel inspection of the non-native academic writings that leads to further scrutiny in other linguistic and non-linguistic phenomena. If we consider medical authors in Egypt as advanced learners of English, they fall in the categorization of learner research which deals with possible non-nativity issues concerning the production of native writers/speakers of English. Further, this effort contributes to enhancing the academic compositions of non-native speakers of English.

Corpus Data: Design, Compilation, and Processing:

Taking into account the standards set by Ackermann & Chen in developing the *Academic Collocation List* (2013), which in turn consider the methods developed by Coxhead (2000), Biber et al. (2004), Chen & Baker (2010) along with the medical corpus development methods set by Wang, Liang, and Ge (2008) and the successive enhanced method proposed by Lei & Liu (2016), the researcher has constructed a corpus deploying the method of the *web as a corpus* in terms of raw extraction, but with manual marking up though. The process resulted in *Menoufia Medical Abstract Corpus* (MMAC) that comprises over 216 thousand words. However, some technical steps followed in establishing the *Medical Academic Word List* (Wang, Liang, & Ge, 2008) subject the outcome NNC collocations for the refining data processing and list development (Lei & Liu, 2016, pp. 42–53; Wang et al., 2008, pp. 446–448).

The *American National Standard for Writing Abstracts* (1979) signified the role of a well-written abstract as it is an “accurate representation of a document […] informative as the nature of the document will permit.”. The literature, furthermore, has appraised research abstracts as a genre detached from the main academic work (Gläser, 1991; Gledhill, 2000, pp. 49–52). Yet most significantly, Bondi & Sanz argued that the standalone abstracts genre is deemed “one of the most important in present-day research communication” (2014). Notably, the MMJ editors stress, in both their official website and in the journal’s instruction page, following structured abstract composition to support readers.

What the researcher finds worth inspecting closely in the medical abstracts is the careful language the writer ought to use to best represent his/her work. Therefore, the researcher manually marked up each of the FIVE moves in each abstract available after grabbing them from the MMJ website using the *WebBootCat* tool (780 in total) (Baroni et al., 2006). The procedure aims at scrutinizing the revealed collocational patterns deployed and comparing them with correspondent medical research work of native speakers of English. Thereby, the researcher delimited each move according to structured abstracts, version 3: 1. Introduction, 2. Aims/Objectives, 3. Method, 4. Results, and 5. Conclusion (Hartley & Cabanac, 2017). Then, attributes of the author's name(s), gender, department, publication year, issue, and corresponding months were added. The following steps summarize the data collection & design process.

1. Corpus Creation: He used the *WebBootCat* tool to electronically collect all applicable RAs from the *Menoufia Medical Journal* website: http://www.mmj.net.eg. Then, performed the compiling, annotating, and adding metadata to the resulting corpus.

2. Vocabulary List Extraction: Instead of following the conventional method of excluding the General Service List (West, 1953) and working on a long chain of filtration and refinement (e.g., Coxhead, 2000; Wang et al., 2008), the researcher deployed the *New Academic Word List* (Gardner & Davies, 2014) in combination with Coxhead et al.'s *Science-Specific WordList* (2007) in one file. Afterward, the file was accessed to whitelist our created corpus and shortlist its academic vocabulary collectively before utilizing the content lexemes as nodes in investigating correlated collocates.

3. Collocation Development: The outcome list of the previous step was processed onto a search for collocates primed with. Statistically, the minimum retrieval score is set to ≥4 based on the LongDice measure deployed by Sketch Engine. For accuracy reasons, the entire procedure was repeated using LancsBox, v.4.0 (Brezina et al., 2018) which uses more sophisticated statistical measures (as suggested by Brezina, 2018; Gries, 2013, 2016) along with other AMs.

4. Filtration & Refinement: This stage was conducted in cooperation with professors and experts in the medical field after the researcher cleared possible function vocabulary and proper nouns from the source collocates.

5. Comparisons to the *Academic Collocation List* (Lei & Liu, 2018) where both NS and NNS adoption of academic collocations were evaluated in their RAs.

Menoufia Medical Abstract Corpus (MMAC):

Referring to the needs of establishing a corpus of RA abstracts with moves marked up aforementioned in the previous section/stages (Swales, 1990), extracting RA abstracts was followed by cleansing and refining procedures before fulfilling the annotation and compilation phases. Incorporated efforts from THREE medical experts have helped in determining certain data attributes for the markup and compilation phases. Also, regarding technical details, the researcher relied on his programming background, as said earlier, and the support endorsed by the Sketch Engine team via email communication. In other words, the MMAC design and compilation wasn't run by a few clicks on a web software only, but through taxing and time-consuming efforts of a single researcher to produce **795** abstracts with **11** corpus attributes suitable for sub-corpora creation whenever desired (see table below). Corpora files can be reached at the Open Science Framework website (Mandour, 2019).

Despite being an advanced technical and linguistic tool, the *WebBootCat* tool does not target the articles' language solely. Extractions would include unneeded excerpts of the Journal's content pages, editorials, font and layout formats, etc. Thereby, a *Cleansing* phase had to manually take place to validate possible results of our corpus research. Moreover, further steps of encoding collected HTML files in other eXtensible Markup Language (XML) by which the researcher, and other future researchers, should be able to look into corpus files using any corpus software (see McEnery & Hardie, 2011, pp. 29–48). Though, another important step to do while encoding was to add *structures and attributes* (i.e., year, issue, the author's gender, etc.) to enable the researcher to explore this medical discourse more deeply - e.g., we can compare lexical items produced by male and female researchers. Accordingly, the researcher followed these procedures:

1. Using the Sketch Engine's *WebBootCat* tool to extract all available abstract files from the MMJ website (http://mmj.eg.net) with all related metadata. The outcome was **20** files for **20** journal issues that cover **6** years of publications. All files were coded in HTML format.
2. The outcome files were refined by: i) excluding editorial files from the results, ii) downloading the 20 files and having them processed in XML format (coded in UTF-8) using the open-source Notepad++ software (Ho, 2017), and ii) cleansing each file's contents by removing page layout codes and unneeded tags and codes,

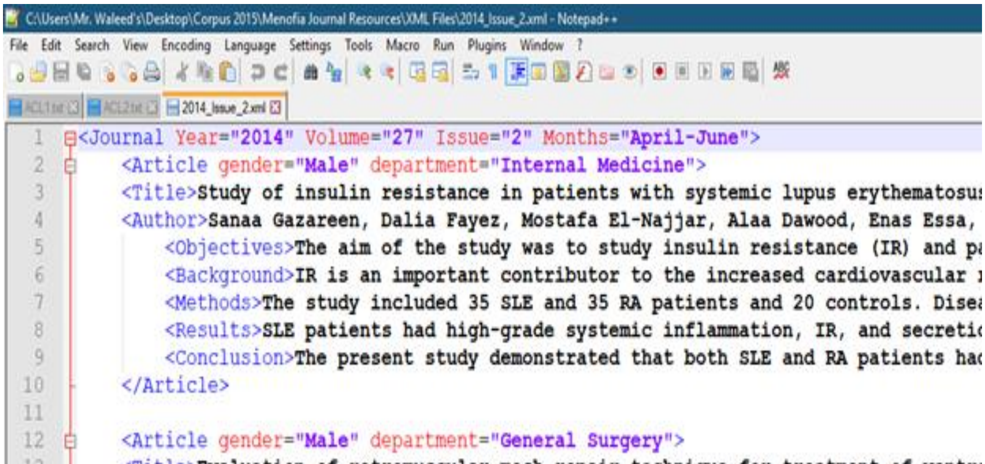
3. Each file that resembles an issue comprising its abstracts' data was then annotated with relevant metadata (i.e., publication year, issue number, and related months). Thereafter, another layer of metadata was inserted to mark the authors' genders and the medical departments they belong to,
4. Removing irrelevant abstracts attributed to non-Egyptian researchers found in the collection, as well as the broken (unstructured) abstracts if found.
5. Multi-time reviewing was accounted for guaranteeing consistency and error-free compilation administered by the researcher with the support of medical professionals in order to ensure that all attributes are consistently matching the data revealed in published abstracts,
6. Auto-tagging/compilation stage was performed upon uploading the corpus files on the Sketch Engine database. For grammatical tagging, the researcher used *the English TreeTagger pipeline*, version 2 (the latest version at the time of this corpus production), and *TreeTagger Extraction* 2.3 for term definitions (the top choice). Below is a table that shows the structure and attributes. And there is a screenshot of an MMAC file with proper annotations as set in the table.

Table 2.

Corpus Structures & Attributes Created in MMAC Files

Structures or Attribute	**XML Tags Used**
Journal Year	*e.g. Year=”2013”*
Issue number	*e.g Issue=”1”*
Articles sequence	*e.g <Article no=”1”></Article>*
Articles' Titles	<title></title>
Authors' names	<author></author>
Author(s) gender	<gender></gender>

Department	*e.g. department= "Cardiology"*
Objectives / Aims of Research / Research Aims / Purpose/ Introduction + Objectives	<Objectives></Objectives>
Background / Data Source	<Background></Background>
Methods / Materials and Methods / Patients and Methods / Participants + Materials and Methods / Subjects & Methods / Data & Summary / (Data Selection, Data Extraction) / Settings and Design	<Methods></Methods>
Results / Findings / Data Synthesis	<Results></Results>
Conclusion	<Conclusion></Conclusion>


C:\Users\Mr. Waleed's\Desktop\Corpus 2015\Menofia Journal Resources\XML Files\2014_Issue_2.xml - Notepad++
File Edit Search View Encoding Language Settings Tools Macro Run Plugins Window ?
ACL1.txt
ACL2.txt
2014_Issue_2.xml

```
<Journal Year="2014" Volume="27" Issue="2" Months="April-June">
    <Article gender="Male" department="Internal Medicine">
    <Title>Study of insulin resistance in patients with systemic lupus erythematosus
    <Author>Sanaa Gazareen, Dalia Fayez, Mostafa El-Najjar, Alaa Dawood, Enas Essa,
        <Objectives>The aim of the study was to study insulin resistance (IR) and pa
        <Background>IR is an important contributor to the increased cardiovascular r
        <Methods>The study included 35 SLE and 35 RA patients and 20 controls. Disea
        <Results>SLE patients had high-grade systemic inflammation, IR, and secretio
        <Conclusion>The present study demonstrated that both SLE and RA patients had
    </Article>

    <Article gender="Male" department="General Surgery">
```

Figure 1. Screenshot of an MMAC file with Proper Annotations

The screenshot above, along with the table that precedes, exhibits certain modifications that were decided by the researcher and the professional consultancies to retain consistency within the created corpus data following the journal’s instructions and Hartly & Cabanac’s third modal (2017). First, due to some sort of inconsistent structures employed in the published abstracts, sections have been swapped differently from the norm (e.g., Objectives, Background). Furthermore, some authors combined two sections in one, particularly *Objectives & Background*, which ought to be corrected manually by the researcher to retain unity in those conventional parts, with proper delimitation. Another issue underlies in the *Background* section, which was neglected by a number of medical researchers. Some other authors wrote their research tools and librarian sources deployed in segments at the *Background* section. For those two observations, the researcher kept them as are.

A further issue the researcher has encountered was departments’ names; some authors used different names for certain departments (e.g., *Public Health/Community Health*, *Cardiothoracic Surgery/Cardiac Diseases*). Thus, a single name was chosen by specialists to attribute RA departments. Subsequently, **35** medical subject categories were resulted in the compiled corpus, compared to 25 categories of Wang et al.’s corpus (2008). See the table below. It contains 35 subject categories with **795** total abstracts displayed in descending order (from 2013 to 2018). The top 6 departments (internal medicine, pediatrics, family medicine, general surgery, clinical Pathology, ophthalmology) resemble almost half the number of collected abstracts.

Table 3.

Medical subject categories & correspondent numbers of Abstracts

#	Medical Subject Category	No. of Abstracts	#	Medical Subject Category	No. of Abstracts
1	Internal Medicine	69	19	Pharmacology	11
2	Pediatrics	68	21	Neurosurgery	11
4	General Surgery	68	20	Chest Diseases	10
3	Family Medicine	61	22	Forensic and Clinical Toxicology	9
6	Ophthalmology	57	24	Clinical Oncology and Nuclear Medicine	9
5	Clinical Pathology	53	23	Histology	8
7	Cardiology	42	26	Cardiothoracic Surgery	8
8	Radiology	36	28	Urology	8
9	Medical Microbiology	33	25	Parasitology	7
10	Obstetrics and Gynecology	33	27	Physiology	6
11	Public Health	27	29	Neuropsychiatry	6
12	Dermatology and	25	30	Physical Medicine and	5

	Andrology and STDs			Rehabilitation	
13	Otorhinolaryngology	23	31	Anatomy	4
14	Clinical Biochemistry	22	32	Orthopedic Surgery	2
16	Orthopedics	22	33	Liver Surgery	1
15	Anesthesiology and Intensive Care	18	34	Artificial Kidney	1
17	Tropical Medicine	16	35	Rheumatology	1
18	Plastic Surgery	15			

There are some further important figures to highlight: Regarding gender distribution, female medical researchers slightly outnumber the male participants (405 to 390). In terms of tokens produced, male researchers produced 133,669 tokens compared to 129,911 for female authors. Notwithstanding, the ratio can be still viewed as balanced gender participation in medical research work. In addition, the abstracts' section sizes are structurally bona fide in terms of words and tokens distribution. See the graph below.

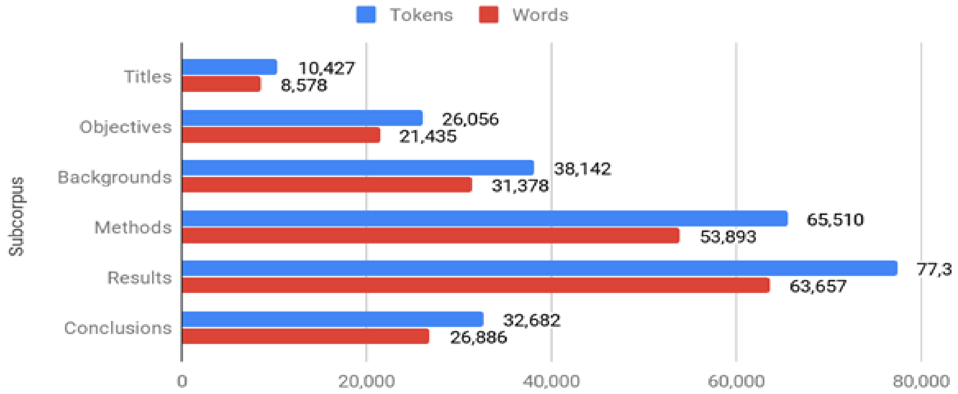


Figure 2. Tokens & Word Distributions in MMAC

The bar graph also indicates that methods and results sections constitute collectively more than half of the corpus sizes. Interestingly, the conclusion and objective sections (26,886 and 21,435 respectively) follow the medical abstracts' background parts which include almost 30% more words than their rhetorical precedent section (objectives). The Egyptian researchers' composition, therefore, conforms with the guidelines of the American Medical Association (AMA) style as well as adheres to the local journal's instructions displayed on its official website.

Results and Discussion:

Outcomes substantiate enough allegations of the near-nativity composition in the Egyptian medical discourse. Based on the Ackermann & Chen's *Academic Collocation List* (2013) which is compiled from "the written curricular component of the Pearson International Corpus of Academic English (PICAE)", 1,672 academic collocations were identified and generated in an *n*-gram file before whitelisting with the medical abstract corpus (the MMAC). The following table exhibits the common ACL in the Egyptian medical abstracts which roughly resemble 6% (150 items in 779 total frequency) (see table 4). In other terms, we can predict at least one academic collocation in each published Egyptian medical abstract with a probability of 97.9% - a percentage which is, according to

Mauranen (2012), relatively lower than the expected range of academic production in a specific domain.

Table 4.
Used Academic Collocations in the MMAC

#	**Academic Collocation**	**Freq. in MMAC**	#	**Academic Collocation**	**Freq. in MMAC**	#	**Academic Collocation**	**Freq. in MMAC**
1	significant difference	174	51	sexual violence	3	101	profound effect	1
2	positive correlation	53	52	accurate assessment	2	102	considerable evidence	1
3	significant increase	53	53	risk assessment	2	103	scientific evidence	1
4	significant correlation	50	54	limited capacity	2	104	detailed examination	1
5	comparative study	44	55	significant change	2	105	main focus	1
6	negative correlation	32	56	normal development	2	106	dominant form	1
7	significant improvement	30	57	adverse effect	2	107	ethnic group	1
8	socioeconomic status	18	58	negative effect	2	108	potential harm	1
9	statistical	14	59	positive effect	2	109	great impact	1

	analysis							
10	major cause	12	60	high frequency	2	110	positive impact	1
11	significant relationship	12	61	mental illness	2	111	significant influence	1
12	medical treatment	12	62	emotional impact	2	112	little information	1
13	rural area	11	63	profound impact	2	113	new insight	1
14	mental health	10	64	significant impact	2	114	high intensity	1
15	significant reduction	10	65	low intensity	2	115	social interaction	1
16	significant role	10	66	brief overview	2	116	considerable interest	1
17	significant effect	8	67	adverse reaction	2	117	renewed interest	1
18	high incidence	8	68	direct relationship	2	118	controversial issue	1
19	mean score	8	69	current research	2	119	daily living	1
20	negative impact	7	70	recent research	2	120	geographic location	1
21	wide range	7	71	main source	2	121	vast majority	1

22	high rate	7	72	broad spectrum	2	122	effective management	1
23	informed consent	6	73	current status	2	123	alternative method	1
24	statistical significance	6	74	diagnostic test	2	124	integral part	1
25	alternative approach	5	75	useful tool	2	125	initial period	1
26	positive attitude	5	76	increasing trend	2	126	high probability	1
27	long duration	5	77	academic year	2	127	evolutionary process	1
28	effective method	5	78	sexual abuse	1	128	high proportion	1
29	critical role	5	79	human activity	1	129	broad range	1
30	essential role	5	80	social activity	1	130	causal relation	1
31	major role	5	81	final analysis	1	131	strong relationship	1
32	random sample	5	82	qualitative analysis	1	132	potential risk	1
33	physical activity	4	83	quantitative analysis	1	133	central role	1

34	primary care	4	84	comprehensive approach	1	134	key role	1
35	strong evidence	4	85	standard approach	1	135	minor role	1
36	natural history	4	86	geographic area	1	136	vital role	1
37	high percentage	4	87	urban area	1	137	common source	1
38	comparative analysis	3	88	considerable attention	1	138	economic status	1
39	essential component	3	89	first author	1	139	social structure	1
40	physical development	3	90	historical background	1	140	longitudinal study	1
41	beneficial effect	3	91	underlying cause	1	141	pilot study	1
42	special emphasis	3	92	major challenge	1	142	previous study	1
43	major factor	3	93	major component	1	143	recent study	1
44	standard method	3	94	careful consideration	1	144	financial support	1
45	positive relationship	3	95	social context	1	145	mutual trust	1

46	crucial role	3	96	strong correlation	1	146	continued use	1
47	direct role	3	97	next decade	1	147	independent variable	1
48	appropriate treatment	3	98	cognitive development	1	148	modified version	1
49	effective treatment	3	99	future development	1	149	domestic violence	1
50	widespread use	3	100	subsequent development	1			

Despite the somehow promising figures of ACL use in the non-native abstract production, most collocation items occasionally occur; i.e., only 32 academic collocations are habitually occurring (>4) according to the traditional frequency-based approach. Thus, the incidence of academic collocations in the Egyptian medical discourse barely reaches 1.3%. Such percentage indicates an inevitable need for medical research quality reinforcement in terms of the phraseological patterns used when it is related to the most crucial part of research articles.

The top list illustrates the NNS’s strong tendency of including collocations of statistical connotations rather than the purely medical ones, such as *statistically significant, positive correlation*, and *significant increase*. *Medical treatment* and *mental health* are two examples of highly salient academic collocations though. From the Lexico-grammatical perspective, research results agree with the literature regarding the heavy use of forms of nominalization (primarily *adjective+noun* and *noun+noun* collocations). Notwithstanding, the Egyptian medical researchers’ interest in using nominal combinations conforms with the fact that they hold the lion's share (74.3%) in the entire ACL. In other words, regardless of the low frequencies in academic collocations revealed in NNS medical abstracts, they remain consistent with NS choices. The published abstracts retain the balance of manifesting academic language despite the innate nature of focusing primarily on medical and statistical patterns.

Concluding Remarks:

The current research work contributes to the advanced learner research in general, and to the (non-linguistic) Egyptian medical research in particular. It investigates the academic collocations used in medical research abstracts which are written by Egyptian researchers as non-native speakers of English. This phraseological inspection compares their formulaic expressions to those provided by native-English-speaking writers. Outcomes assure the traditional phraseological aspect of the medical discourse where the focus is predominantly on nominals. The NNS's composition succeeded in using a large spectrum of academic collocations, in spite of its relatively non-significant usage. On the other hand, there is an intriguing indication of excessive use of collocations of statistical references at the expense of medical and other academic expressions, in a way that may question the qualitative value of the rendered research work.

To recommend, medical authors in Egypt are to take into account including 1. Qualitative interpretations to their submitted abstracts, 2. More academic collocations to incorporate in medical abstract moves (introduction, methods, results, and discussion), and 3. As a significant pedagogical implication, they ought to involve young researchers to identify and evaluate top-rated medical abstracts' formulaic language in terms of their conformity with academic standards, and possibly other linguistic features. Thus, they get trained on scrutinizing their prospective research medium prior to publication. Ultimately, the researcher suggests future studies on comparing medical abstracts written by NNS vs NS's formulaic production. Further, with the MMAC sectionally structured in hand, further research work is to recommend investigating other lexico-grammatical features, e.g., compounds.

Acknowledgment

I wish to acknowledge the help provided by the Menoufia Medical Journal's managing director, prof. Ahmed Ragab. My gratitude is extended to the Journal's support team who facilitated the transfer of files and gave access to the published research papers in the faculty of medicine, Menofia University. I would also like to show my deep appreciation to the professional medical experts, as well as my colleagues, who helped me finalize my research project.

References/ Kaynakça

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*. https://doi.org/10.1016/j.jeap.2013.08.002

American National Standard for Writing Abstracts. (1979). https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&q=American+National+Standard+for+Writing+Abstracts&btnG=

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: A Web Tool for Instant Corpora. *Proceeding of the EuraLex Conference*, 123–132.

Benson, M., Benson, E., & Ilson, R. (2010). *The BBI Dictionary of English word Combinations* (3rd ed.). John Benjamins Amsterdam.

Biber, D. (2006). Stance in Spoken and Written University Registers. *Journal of English for Academic Purposes*. https://doi.org/10.1016/j.jeap.2006.05.001

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. In *Applied Linguistics*. https://doi.org/10.1093/applin/25.3.371

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. http://corpora.lancs.ac.uk/stats.

Brezina, V., Timperley, M., & McEnery, T. (2018). *LancsBox v. 4.2 [software].*

Chen, Y. H., & Baker, P. (2010). Lexical bundles in l1 and l2 academic writing. *Language Learning and Technology*.

Cowie, A. (1998). *Phraseology: Theory, analysis, and applications*. https://books.google.com/books?hl=en&lr=&id=Df-iQpNMLcgC&oi=fnd&pg=PR10&dq=cowie+phraseology+theor

y+analysis+and+applications&ots=fQYMFRk5nj&sig=0Ssq3OMPR1d9EbW2yESmywzkONo

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 1–2. https://doi.org/10.2307/3587951

Coxhead, A. (2008). Phraseology and English for academic purposes. In *Phraseology in Foreign Language Learning and Teaching*. https://doi.org/10.1075/z.138

Coxhead, A., & Hirsch, D. (2007). A Pilot Science-Specific Word List. *Revue Française de Linguistique Appliquée*, *12*(2), 65–78. https://www.cairn-int.info/load_pdf.php?download=1&ID_ARTICLE=E_RFLA_122_0065

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

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL - International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral.2009.007

Ferguson, G. (2013). English for Medical Purposes. *The Handbook of English for Specific Purposes*, 243.

Firth, J. (1957). *Studies in Linguistic Analysis*. Oxford, Blackwell.

Flowerdew, J. (2015). Some Thoughts on English for Research Publication Purposes (ERPP) and Related Issues. *Language Teaching*, *48*(2), 250–262.

Gardner, D., & Davies, M. (2014). A New Academic Vocabulary List. *Applied Linguistics*. https://doi.org/10.1093/applin/amt015

Gläser, R. (1991). The LSP Genre Abstract Revisited. *ALSED LSP Newsletter*, *13*, 3–10.

Gledhill, C. (2000). *Collocations in Science Writing* (Vol. 22). Gunter Narr Verlag.

Granger, S. (1999). Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In Cowie, A. *Phraseology: Theory, Analysis and Applications*.

Gries, S. T. (2013). 50-something years of work on collocations: What is or should be next …. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.18.1.09gri

Gries, S. T. (2016). Quantitative corpus linguistics with R: A practical introduction. In *Quantitative Corpus Linguistics with R: A Practical Introduction*. https://doi.org/10.4324/9781315746210

Hartley, J., & Cabanac, G. (2017). Thirteen Ways To Write An Abstract. *Publications*, *5*(2). https://doi.org/10.3390/publications5020011

Ho, D. (2017). *Notepad++* (7.8.2). https://notepad-plus-plus.org/

Hogg, R. (1992). *Cambridge History of the English Language, vol. 1 Cambridge*.

Hornby, A. (2015). *Oxford Advanced Learner's Dictionary of Current English* (M. Deuter, J. Turnbull, & J. Bradbury, Eds.). Oxford University Press.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44.

Hyland, K. (2008). Genre and Academic Writing in The Disciplines. *Language Teaching*, *41*(4), 543–562.

Hyland, K., & Shaw, P. (2016). *The Routledge Handbook of English for Academic Purposes*. Routledge.

Hyland, K., & Wong, L. L. C. (2019). *Specialised English: New Directions in ESP and EAP Research and Practice*. Routledge.

Krishnamurthy, R. (2006). Collocations. In *Encyclopedia of Language & Linguistics*. https://doi.org/10.1016/B0-08-044854-2/00414-4

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

Lei, L., & Liu, D. (2018). The Academic English Collocation List. *International Journal of Corpus Linguistics*. https://doi.org/10.1075/ijcl.16135.lei

Mandour, W. (2019). *Egyptian Medical Corpora*. https://doi.org/10.17605/OSF.IO/WXZKD

Marco, M. J. L. (1998). Phraseological Patterns in Medical Discourse. *ESPecialist, 19*(1), 41–56.

Mauranen, A. (2012). *Exploring ELF: Academic English shaped by non-native speakers*. Cambridge University Press.

McEnery, T., & Hardie, A. (2011). *Corpus Linguistics: Method, Theory and Practice*. https://books.google.com/books?hl=en&lr=&id=3j3Wn_ZT1qwC&oi=fnd&pg=PR7&dq=corpus+linguistics+hardie+and+mcenery&ots=UGwwU9lVWw&sig=lnDO_WFTatEhgFB6vI82zi1-Gww

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*. https://doi.org/10.1093/applin/24.2.223

O'Dell, F., & McCarthy, M. (2008). *English collocations in use*. Cambridge University Pres.

Ordóñez-López, P., & Edo-Marzá, N. (2016). Medical discourse in professional, academic and popular settings. In *Medical discourse in professional, academic and popular settings*. https://doi.org/10.21832/9781783096268

Ragab, A. (2018). *Personal Communication*.

Richards, J. C. (2013). Longman Dictionary of Language Teaching and Applied Linguistics. In *Longman Dictionary of Language Teaching and Applied Linguistics*. https://doi.org/10.4324/9781315833835

Schmitt, N. (2012). Formulaic Language and Collocation. In *The Encyclopedia of Applied Linguistics*. https://doi.org/10.1002/9781405198431.wbeal0433

Sinclair, J. (1991). *Corpus, concordance, collocation*.

Sinclair, J., Jones, S., & Daley, R. (2004). *English collocation studies: The OSTI report*. Bloomsbury Publishing.

Staicu, S. N. (2017). MEDICAL JARGON: COLLOCATIONS, TYPOLOGIES, DISTRIBUTION. *Studii Şi Cercetări de Onomastică Şi Lexicologie*. http://cis01.central.ucv.ro/revista_scol/site_ro/2017/lexicologie/staicu_simona.pdf

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

Walter, E. (2008). Cambridge Advanced Learner's Dictionary. In *PONS-Worterbucher, Klett Ernst Verlag GmbH* (3rd ed., Vol. 3).

Wang, J., Liang, S. lan, & Ge, G. chun. (2008). Establishment of a Medical Academic Word List. *English for Specific Purposes*. https://doi.org/10.1016/j.esp.2008.05.003

West, M. (1953). *A general service list of English words, with semantic frequencies and a Supplementary Word-List for the Writing of Popular Science and Technology.*

Wray, A. (2005). *Formulaic language and the lexicon*. Cambridge University Press.