A key to understanding why a text is difficult to process
Lexical uniqueness of academic English texts

Natalia Borza

ABSTRACT:
While the register of English language tertiary textbooks has been investigated substantially, moderately little is explored about the register analytical features of secondary textbooks. The purpose of the present pedagogically-driven study is to analyse the register of biology textbooks for secondary students from the point of view of English as a second language (ESL) teaching by describing the lexical uniqueness of the register of the biology corpus (BIOCOR) 10th-grade students need to process during their studies at a bilingual secondary school. The BIOCOR (consisting of 7,021 words) was compared to a reference corpus (REFCOR) of general English texts at a CEFR B2 level (comprising 7,098 words) by exploring its high-value positive and negative keyness lexical items. The results of the investigation disclose that the lack of specialised uniqueness is prevalent in the BIOCOR with regard to academic English and specific biology terminology. The lexical plainness of the biology textbook can be regarded as one of the linguistic features revealing the non-academic but popularizing nature of the secondary textbook register.

KEYWORDS:
academic English, English for specific purposes (ESP), keyness, lexical uniqueness, register analysis

1. RATIONALE AND THE RESEARCH QUESTION

Students at an English-Hungarian bilingual secondary school in Budapest tend to face an academically challenging situation in the second year of their studies when they start to master what is required in the 10th grade nationwide. The current pedagogically driven research to investigate one of the possible linguistic sources of the problem is motivated by my experience as a practicing English language teacher having observed the regular reappearance of the same hardships among the 10th graders.

The present study analyses the written register of English-language biology textbooks for secondary students from the viewpoint of English as a second language (ESL) teaching by describing the register of the biology corpus students need to process during their studies. The register analysis is expected to result in a pool of data relevant for gaining pedagogical insights applicable by teachers instructing in the intensive English language preparatory year of the bilingual secondary school as to what extent the language foci of the preparatory year enable students to handle the language use of the biology texts 10th-graders are assigned to process. Besides gaining a deeper understanding of the 10th-grade bilingual students' needs in terms of English language and thus supporting my own and my colleagues' professional development as general English teachers, this exploratory and descriptive corpus-based study can provide insights for future biology English for specific purposes (ESP) teachers, once biology ESP has been included in the 'zero year' language programme of the secondary bilingual school. Although the present research launches a close investigation into describing the language use of two types of texts at a particular bilingual secondary school in Hungary, the results of the enquiry are not restricted to the secondary school at hand, they can be meaningfully transferred and applied by educators working in any English-language international school where some of the students are non-native students.

Keeping the 10th-graders' difficulty of tackling academic subjects in English in the foreground, the present pedagogically motivated study aims to investigate the following problem from a linguistic point of view: To what extent do the general English reading texts (referred to in the study in its acronym form as the REFCOR for short) assigned in the intensive language preparatory course in the 9th grade at an English-Hungarian bilingual secondary school enable students to handle the biology texts used in the subsequent term (hereafter referred to as the BIOCOR)? Since the most outstanding linguistic features along which registers differ from one another is considered to be vocabulary (Atkins, Clear & Ostler, 1992; Biber, 1989, 1993; Sinclair, 1991), the present research reports on the characteristic linguistic features of the BIOCOR with regard to its lexical uniqueness. Accordingly, the paper attempts to answer the research question: What lexical uniqueness is characteristic of the BIOCOR in comparison with that of the REFCOR?

2. REVIEW OF THE LITERATURE

To ensure that the current analysis yields data practically valuable for ESL and ESP teachers, the register analytical approach rather than the genre analytical one was adopted in the present research (for the underlying reasons see Borza, 2015). The use of computer technology renders register analysis more reliable (Biber & Conrad, 2009), thus computerized corpus-linguistical methods were applied in the current project. Within the frame of corpus linguistics, Biber (1988) introduced a comprehensive methodological approach to describing patterns in register variations, the computerized method of multidimensional analysis (MDA). This method aims at finding underlying linguistic parameters, or dimensions, as well as specifying linguistic parallels and dissimilarities among registers along the dimensions identified. MDA relies on multivariate statistical techniques, especially factor analysis, to investigate the co-occurrences of linguistic features when discovering systematic patterns of variation among registers. As is characteristic in the register approach, the complexity of linguistic features is emphasised in the process of obtaining adequate descriptions of registers. In line with the early recognition of the importance of linguistic co-occurrences (Brown & Fraser, 1979), MDA follows Biber's (1988) observation that statistically significant linguistic features tend to cluster in texts as they share communicative functions. Consequently, the method finds it misleading to focus on specific, isolated linguistic features and does not investigate single parameters individually.[1] The MDA perspective aims at finding groups of linguistic features that co-occur in registers. To map registers onto the groups of linguistic markers or dimensions, texts in the corpora are automatically analysed for linguistic features representing numerous major grammatical and functional characteristics. After carrying out the quantitative, numerical analyses, the frequent (positive) and rare (negative) features in the dimensions detected through factor analysis are interpreted in terms of communicative functions. The qualitative analysis specifies how the language features with statistically significant values are well-suited to the communicative purposes of the text.

Using the numerical and functionally interpretive method of MDA developed in Biber's (1988) seminal work, numerous registers have been explored, among them are letters (Biber & Finegan, 1989), medical academic prose (Atkinson, 1992), 18th-century authors across different registers (Biber & Finegan, 1994b), spoken and written registers in a variety of languages (Biber & Finegan, 1994a; Biber, 1995), research articles and textbooks (Conrad, 1996), internet-based and computer-mediated communication (Herring, 1996), newspapers (Biber & Finegan, 1997), scientific prose (Atkinson, 1999), newspapers, magazine articles and medical writing (Vilha, 1999), elementary school writing (Reppen, 2001), disciplinary texts (Conrad, 2001), historical and contemporary registers (Conrad & Biber, 2001), speech and writing at university (Biber et al., 2002; Biber, 2006), radio and TV sports commentary (Reaser, 2003), biology research articles (Biber & Jones, 2005), university classroom talk (Csomay, 2005), biochemistry research articles (2007), blogs (Grieve et al., 2011), academic registers and sub-registers (Nesi & Gardner, 2012), movie language (Forchini, 2012).

Applying Biber's (1988) rather complex MDA for capturing register specific features has been challenged by Tribble's (1999) proposition claiming that the application of the keyword function of WordSmith (Scott, 2008) could reveal similar patterns as MDA. Xia and McEnery (2005) investigated whether this assertion proves to be correct. Contrary to the most straightforward implication of the term 'key words,' they are not the most frequently used words in the register, neither are they the ones that carry the most important propositions in the text; however, key words make the text characteristically different compared to a large reference or benchmark corpus. Key words can be identified through statistical comparison carried out by the keyness function of keyword programs. The test of keyness is predicated on a log-likelihood test, Dunning's procedure (1993) most typically, which is not based on the presupposition that data have a normal distribution in the text (McEnery et al., 2006). Showing the lexical uniqueness of a text, keyword lists reveal register specificity by containing words that are either significantly frequent or on the other end of the spectrum, significantly infrequent in the collection of texts. In the first case the list allows to investigate positive keyness, that is, words and structures that make the target corpus different from a larger reference corpus, while the second list provides information about negative keyness, about the words, expressions and structures that are dramatically missing from the corpus under scrutiny compared to a benchmark corpus.

[1] Biber's (1988) work was ground-breaking in examining 67 linguistic features in 481 texts and 23 registers.

Through investigating the effectiveness of the keyword function, Xia and McEnery (2005) endeavoured to find a labour-effective method that could substitute the rather complex and "extremely time-consuming" (McEnery et al., 2006, p. 308) MDA procedure, which resists any simple characterisation. Although MDA is a powerful tool in register analysis, which has been used to uncover various registers as demonstrated above, it is undoubtedly demanding to carry out. The reason for its laborious nature is the fact that it requires the sophisticated statistical analysis of a large number of linguistic features to identify the groups of features that co-occur in the text with high frequency. To show that MDA fails to be irreplaceable with a less arduous tool for register analysis, Xiao and McEnery (2005) undertook a keyword analysis to compare three registers (conversation, speech, and academic prose) by producing wordlists of corpus files extracted from large American corpora (the Santa Barbara Corpus of Spoken American English, the Corpus of Professional Spoken American English, and the Freiburg-Brown corpus of American English), which were compared to a reference corpus, the British National Corpus, to detect and compile those words whose frequency differed from the reference corpus either by being unusually high (positive keywords) or extremely low (negative keywords). The results of their study confirmed that applying the keyword approach is capable of producing comparable results to the MDA approach and can identify important register patterns, despite creating a less nuanced comparative contrast of registers, which is "not as likely to work for finer distinctions among texts" (Conrad, 2015, p. 318).

3. METHODS

In order to make the study replicable and the results transferable, the following Methods Section comprises four sections. First, the steps of the keyness analysis are explicated (Section 3.1.1), then the methods of compiling the biology corpus under investigation are accounted (Section 3.1.2), which is followed by those of the reference corpus (Section 3.1.3). Next, it is explained why a mini-corpus was adopted in the research (Section 3.1.4).

3.1 THE PROCESS OF THE KEYNESS ANALYSIS OF THE CORPUS

3.1.1 THE STEPS OF KEYNESS ANALYSIS

Although Biber's (1988) multidimensional analysis (MDA) has a long record of reliably uncovering linguistic patterns of registers, the present research follows a more recent analytical method which is considered to be a replacement of MDA (Tribble, 1999). The reason for choosing the keyword application of the WordSmith program (Scott, 2008) instead of carrying out a MD analysis on the BIOCOR is not simply due to the novelty of the software. In pragmatic terms, the decision was based on considering Xia and McEnery's (2005) empirical research results. Their study proves that revealing keyness with the WordSmith program is a method that provides comparable results to MDA since the new application can identify similar linguistic patterns among registers. In theoretical terms, the models of already identified dimensions to explore the characteristic linguistic features of texts by the application of MDA (Biber, 2001;[2] Biber et al., 2014;[3] Staples et al., 2018[4]) fail to appear to be utterly relevant considering the focus of the present research. Neither the ESL teachers instructing 9th-grade students in the bilingual programme, nor ESP teachers in general would benefit directly from the linguistic data of these dimensions in their teaching practice. The reason why the above-mentioned dimensions do not fully address the key aspects relevant in the present educational setting might lie in the fact that these dimensions were identified with the aim of finding generalizable parameters of linguistic variations for a different discourse domain: 1) general parameters of variations among spoken and written registers in English and 2) patterns of grammaticality and lexico-grammatical characteristics in university student writing/speaking tasks, that is, productive skills were in the focus, which are greatly different from the receptive skill of processing reading tasks. Thirdly, the multivariate statistical technique on which MDA is based is factor analysis, which is a sophisticated method that can be applied effectively to large corpora. In practical terms, factor analysis does not work effectively on a mini-corpus, thus the current corpus of 7,000 running words cannot be investigated fruitfully along the Biberian lines of factor analysis.

Keyness describes the distinguishing lexical characteristics of a register by comparing its language use to that of another register (Xia & McEnery, 2005). The keyword application of WordSmith version 5 (Scott, 2008) was used in the present research to extract lexical items that are present in the BIOCOR, ones which are, however, not typically used in the REFCOR. That is, keyness results show the lexical uniqueness of the BIOCOR by compiling lexical items that make the register markedly different from the REFCOR. Inversely, the keyness application was also applied to collect lexical items that are underrepresented in the BIOCOR compared to the REFCOR, which set of lexical tokens is labelled as displaying negative keyness values.

[2] Seven dimensions of the Biberian multidimensional analysis (2001)
1) Involved versus informational
2) Narrative versus nonnarrative
3) Elaborated reference versus situation dependent reference
4) Overt expression of argumentation
5) Abstract style versus nonabstract style
6) Online informational elaboration marking stance
7) Academic hedging

[3] Four dimensions of the multidimensional analysis by Biber et al. (2014)
1) Literate versus oral response
2) Information source: Text versus personal experience
3) Abstract opinion versus concrete description/summary
4) Personal narration

[4] Four dimensions of the multidimensional analysis by Staples et al. (2018)
1) Compressed procedural information versus stance toward the work of others
2) Personal stance
3) Possible versus completed events
4) Informational density

It needs to be underlined that keyness, either positive or negative, does not reveal lexical items that frequently or infrequently occur in the BIOCOR but ones which are characteristically different with respect to their frequencies when the register is compared to the REFCOR. Using this method ensured that lexical items which are not register specific, ones which occur with similar frequencies in both corpora, such as *the*, *of*, or *and*, are not compiled.

Keyness is determined by statistical comparison carried out by keyword programs. A word is considered to be key if its frequency in the corpus when compared with its frequency in the reference corpus is such that the statistical probability as computed by the appropriate procedures described below is smaller than or equal to a *p* value of 1E–6 (Scott, 2008).[5] To compute the keyness of an item, WordSmith version 5 (Scott, 2008) calculates four values, which are consequently cross-tabulated. The four values include the raw frequency of the item in the corpus, the number of running words in the corpus, the raw frequency of the item in the reference corpus, and the number of running words in the reference corpus. The statistical procedure of finding key words includes the chi-square test of significance with Yates' correction for continuity to reduce the error in approximation. The test of keyness in the case of the WordSmith program (Scott, 2008) relies on a log-likelihood test, Dunning's procedure (1993). The fact that Dunning's procedure is not based on the presupposition that data have a normal distribution in the text (McEnery et al., 2006) increases the instrument's reliability. The application of a log-likelihood test, disfavouring normal distribution, was especially important in the present research environment, where the REFCOR does not contain considerably more running words than the target corpus but was compiled to be approximately of the same size as the BIOCOR. WordSmith version 5 (Scott, 2008) treats words which are not represented in the REFCOR as if they occurred 5.0E–324 times (that is 5.0×10^{-324}) in the baseline corpus. To apply a keyword program that assigns such a small value to non-represented lexical items in the corpora was a decisive factor in the choice of the software. Without this slight modification, uncovering stark contrasts between the two registers would have been impossible since cross-tabulation with values of zero does not produce any meaningful result. An infinitesimally small number, however, allows for the handling of lexical items that do not occur in either of the two corpora, and due to the number's close-to-zero value, it does not affect the calculation materially. To ensure reliability, WordSmith version 5 (Scott, 2008) defines those items as key whose *p* value is smaller than or equal to 1E–6, that is 0.000001. The *p* value shows the danger of being ungrounded when claiming relationships. Consequently, an extremely low p-value threshold increases reliability. In the present case the chance of erroneously listing words with similar frequency in the two corpora as key words is 0.00001%.

In order to arrive at data which are practically useable for ESL teachers instructing in the bilingual programme of the school and for biology ESP teachers alike, words of the same root were lemmatized by the keyword program before running the keyword application. That is, keyness values were determined for word families rather than

[5] 1E–6 is a standard scientific notation for the value of one times 10 to the power of –6, which equals one over 1 million, or 0.000001.

for individual word forms. Lemmatization was treated as fundamentally essential since the investigation of word families produces more useful data for ESP teachers than that of conjugated verb forms and various word formations in the process of working out the lexical dimension of ESP syllabi.[6] The same argument supports the practical reason why word lists for learners of English also tend to group words into families (West, 1953; Xue & Nation, 1984). Besides, compiling words in word families instead of listing isolated elements of different word forms was chosen for theoretical reasons too, namely, word families form a unit in the mental lexicon (Bauer & Nation, 1993; Nagy et al., 1989). Lemmatization rendered the following different word forms as one group:

— singular and plural forms, e.g., *cell — cells*, *parasite — parasites*, *segment — segments*;
— nominative and genitive forms, e.g., *mosquito — mosquito's*;
— regular inflections of the verb (verbs in different tenses), e.g., *cause — caused*, *reproduce — reproduces*;
— verbs and gerunds, e.g., *spread — spreading*;
— base, comparative and superlative adjectives, e.g., *small — smaller — smallest*;
— derivations of the word: *amoeba — amoebic*, *blood — bleeding*, *chemicals — chemically*, *class — classify*, *contract — contractile — contraction*, *dead — death — die*, *digestive — digestion*, *granules — granular*, *saliva — salivary*, *slime — slimy*.

Yet compound words were not joined in one batch, thus *flat* and *flatworm*, *stream* and *streamlined* for instance were computer-counted separately. The reason for not lemmatizing compound words lies in the strong possibility that the parts of the compounds cover relatively distant meanings, for instance *cow* and *cowslip* or *Mary* and *marigold*.

After running the appropriate statistical procedures of the keyness software, the key words of the BIOCOR were listed by the software in rank order. The computer-counted keyness values of the lemmatized items on the list reveal to what extent the frequency of the particular item is different when compared to that in the REFCOR. Subsequently, the key words were manually correlated to the most frequently occurring lexical items in the BIOCOR (for the most frequently occurring lexical items in the BIOCOR see Borza, 2014). Such a correlation was considered to be important in order to find out more in depth about the nature of the biology register. The most prevalent words in the BIOCOR were recorded in rank order, and arranged in frequency bands. Band 1 contains the most pervasive, most frequent words in the BIOCOR, the ones which are used no fewer than 30 times, while Band 10 comprises more rare items, word families which appear four times. Table 1 shows the frequency of items in particular bands, expressed both in the number of their raw occurrences and in percentages.

[6] For insights regarding the working out of the grammar dimension of the ESP syllabi, e.g., tenses, modal auxiliaries, active-passive voice, sentence complexity, see other research such as Borza 2013, 2016.

Rank order	Raw frequency of lemmas	Frequency of lemmas
Band 1	30 or more	0.42% or more
Band 2	20–29	0.28% – 0.41%
Band 3	15–19	0.21% – 0.27%
Band 4	12–14	0.17% – 0.20%
Band 5	10–11	0.14% – 0.15%
Band 6	8–9	0.12% – 0.13%
Band 7	7	0.10%
Band 8	6	0.08%
Band 9	5	0.07%
Band 10	4	0.06%

TABLE 1: The frequency bands in the BIOCOR.

Individual lexical items and lemmatized tokens which occur fewer than four times in the BIOCOR were not compiled in this investigation. The reason for disregarding low-frequency lexical items is the assumption that in an informational, educational register, such as that of the biology textbook for secondary school students, essential lexical items appear repeatedly in order to fulfil the textbook's instructional function.

Next, in order to gain a better understanding of the degree of the use of specific lexis in the register, the key words were classified into three categories: biology terms, academic English and general English. The category of *biology terms* contains lexical items which have a specific meaning within the context of biology, a meaning or a shade of meaning which differs from the everyday use of the word. A current dictionary of biology (Thain & Hickman, 2004) served as a reference point in determining if a lexical item is to be categorized as a biology term or if the biology related word falls into the category of general English. Thain & Hickman's biology dictionary (2004) was applied as the baseline of categorization since its entries, according to the dictionary's editors, venture to explain the most indispensable notions in biology for teachers and students alike, that is, its selection of information is perfectly relevant in the present educational setting. Words and expressions which appeared as separate entries in the biology dictionary were grouped as biology terms. Every single member of a lemmatized word family was checked in the dictionary in order to ensure that word classes did not affect the labelling of biology terms. For instance, the noun *reproduction* appears as an entry in the biology dictionary; however, the verb *reproduce* does not. In this case the lemmatized word family including the items *reproduce*, *reproduction*, *reproductive* was labelled as a biology term. Yet multi-word dictionary entries, where a lexical item was the head of the entry in conjunction with other words, were not grouped as biology terms unless they were present in the BIOCOR with the exact same word combinations. For example, the word *body* is not a separate entry in the biology dictionary, while the lexical item *carotid body* is. Accordingly, the word *body* was not categorized as a biology term in the present research unless it was used in the BIOCOR in conjunction with the word *carotid*.

The label of *academic vocabulary* was given to those lexical items that appeared on Coxhead's (2000) extensive list of academic vocabulary comprising 570 word families.

Coxhead's academic word list (AWL) was selected to be used in the present research since it is a systematic collection of academic English, a set of wide-ranging lexis typically used in the register of academic English. Furthermore, the list is applied with a high rate of validity in the present research environment as the collection of lexical items was particularly compiled for pedagogical purposes. The AWL was gathered in order to provide insights for English teachers preparing students for their tertiary studies in English as Coxhead aimed at showing what specific lexis was prevalent in academic text. Thus, the AWL accords well with the educational context of the current research as it is concerned with the teaching applications to improve second language students' success in an academic environment when studying disciplines in English. The AWL has been proven to pinpoint the collection of lexical items that makes academic registers markedly different from other registers (Coxhead, 2000), thus it is a reliable instrument to find academic vocabulary in texts in English. The corpus in which the frequency of words was run by Coxhead (2000) embraces four sub-corpora of the following faculty sections: arts, commerce, law, and science. Each of these faculty sections are further divided into seven subject areas. Biology is one of the subject areas of the science sub-corpus, which allows its use as a baseline in the present research environment with a high rate of construct validity. The AWL contains word families that appeared in over half of the twenty-eight subject areas. Words that occurred in fewer than fifteen of the subject areas were labelled as narrow range words and were excluded. This principle ensured that the list could be used for any academic subject area, its coverage is not restricted to specific subjects. In the development of the list, frequency played a key role, word families that were used more than 100 times in the 3,500,000-word-long corpus were shortlisted. Basic vocabulary, words that are among the first 2,000 most frequently occurring words of English as compiled by West in his General Service List (1953), were not involved in the short list, since academic reading presupposes the learner's familiarity with basic vocabulary at tertiary level. From this respect, AWL is advantageous to be used in the current research environment since 10th-grade students are also expected to be familiar with the most widely used words in general English. This similarity ensures a high rate of criterion related validity for the present research. Besides basic lexis, proper nouns, for example names of places and people, as well as Latin forms, such as *etc.*, *i.e.*, were also removed from the AWL short list. Finally, the list was organized into ten sublists based on the frequency of the particular word family. The sublists were numbered consecutively, where sublist one contains the most common academic words in the corpus, while sublist ten comprises less frequent academic lexis. The present research uses Coxhead's (2000) findings in order to see whether the biology texts assigned to 10th-grade students in the bilingual secondary school are difficult to read due to the fact that they contain a large number of academic lexical items among their key words.

Thirdly, lexical items which failed to fit either in the category of biology terms or in the group of academic vocabulary were assigned the label *general English*. High-keyness lexical items within the general English category were collected and listed in order to help general English teachers and biology ESP teachers gain knowledge about the nature of the general English lexis used in the biology textbook for secondary school students.

Following Hoey's (2005, p. 8) notion that our "knowledge of a word includes the fact that it co-occurs with certain other words", the depiction of the lexical environment of the most frequently used key biology terms was considered to be vital. Since knowing a word involves being familiar with its lexical environment, it is indispensable from a pedagogical point of view to underline the importance of such descriptions. Thus, the lexical environments of the biology key words were also described by running the KWIC concordancing application of the software, where the range of investigation was the sentence boundary. All the words that co-occurred with the biology key words in the BIOCOR were compiled and organized according to their part of speech. To make the list straightforward, the alphabetic principle was applied when ordering collocations in the list. The lexical items which describe the environment of the key biology terms were listed in their dictionary forms, which resulted in several changes of form and some of meaning. Particular tenses in which the verbs that collocate with the biology terms were altered, for example the dictionary form *be treated with certain drugs* appears rather than *John was treated with certain drugs*. Similarly, modal verbs which are present in the BIOCOR were not registered, therefore *put bacteria on the surface of the agar* is listed instead of *you should put bacteria on the surface of the agar*. Lastly, to provide a user-friendly list for ESL and ESP teachers, relative clauses applied in the BIOCOR were also neglected, even if it led to certain changes of meaning. Slightly altering the content information present in the BIOCOR was not treated as central in the analysis as the description of the environment of high-keyness biology terms is fundamentally of lexical nature. The main aim of the lexical accounts was to map potential collocations; the descriptions did not endeavour to gather information in the field of biology. For this reason, the phrase *rings divide the body up into segments* was registered rather than the defining relative clause *rings which divide the body up into segments*. Ignoring authentic tenses, modal verbs and relative clauses used in the BIOCOR might appear as running the risk of losing the complexity of the descriptions of the lexical environment, yet, this omission was a conscious decision in order to keep the list as much lexis-focused as possible. The grammatical aspect of the BIOCOR as a potential source of difficulty in processing the corpus was unveiled in another study (Borza, 2013).

Finally, items with negative keyness in the BIOCOR were also collected, and their role in shaping the register of the biology textbook was investigated.

3.1.2 COMPILING THE CORPUS OF THE BIOLOGY TEXTS FOR SECONDARY STUDENTS (BIOCOR)

A register description is of high validity if the corpus is composed of texts which represent the greater part of the register in an appropriate manner. The situational characteristics of the biology textbook from which the texts originate (Roberts, 1981) were described in detail at an earlier phase of the research (see Borza, 2016). Correspondingly, in the process of the compilation of the BIOCOR, careful attention was paid to selecting properly representative texts.

First the exact biology texts which 10th-grade bilingual students are expected to process in the first term were detected. Five high-achieving 10th-graders in English were asked in a structured interview to write down the topics which they covered in the autumn term. The students received their biology textbook (Roberts, 1981) in order to raise the level of accuracy of their academic memories. The question was formulated for high-achievers in English since low-achiever students are more likely to be hesitant when reflecting on their studies. Furthermore, low-achievers tend to be unsuccessful in remembering with precision what has been dealt with in class. Every single interviewee picked the same chapters, which are collected in Table 2. To ensure the highest rate of representativeness possible, the themes of the biology classes from September to mid-January were traced in the electronic register of the school, which is an official documentation written by the biology teacher. It was then confirmed that the chapters of the biology textbook listed by the high-achiever students was exhaustive. Subsequently, the corpus was typed to make the texts computer analysable, and a word count was run. The number of words in the BIOCOR was affirmed to be 7,021.

Order of topics	Title of the chapter	Number of words in the chapter
1	The characteristics of living things	1613
2	Classifying, naming and identifying	875
3	Amoeba and other protists	767
4	Bacteria	689
5	Viruses	777
6	The earthworm	517
7	Harmful protists	1085
8	Parasitic worms	698

TABLE 2: The BIOCOR: the eight chapters of the biology textbook (Roberts, 1981) and their lengths given in words.

3.1.2 COMPILING THE REFERENCE CORPUS (REFCOR)

In the following step, the general English texts which can fulfil the function of a valid base of comparison were chosen. The reference corpus (REFCOR) was compiled from English texts processed in the 9th-grade general English classroom. The texts were collected from the course book students use the last month prior to taking the end-term exam testing their level of English (Prodromou, 1998). The reference texts were selected so that they were representative of all the four task types of the reading component of the exam, the First Certificate in English Cambridge Examination (FCE). Despite the fact that the data of the present research were gathered after 2008, the parts of the reading paper embody a former version of FCE. The reason for not applying the latest version of the exam lies in the fact that the 9th-grade course book (Prodromou, 1998) prepares for the earlier one.[7] The entirety of the reading paper

[7] At the time of data collection, the FCE exams administered by the school were still structured according to the composition of the examination in practice before the 2008 mod-

of the FCE exam contains approximately 2,000–2,500 words in general. To make the comparison of the 7,000-word-long BIOCOR viable, the same-length REFCOR was set out to be built, which readily implied the compilation of more than one single exam. Although the use of the keyness program necessitates the application of a large reference corpus, its compilation was beyond the bounds of possibility in the present pedagogical setting. Namely, the 9th-grade course book, whose 22 units contain five to seven practise texts of each part of the reading exam, merely includes a near 15,000-word-long set of reading tasks. The final guiding principle in compiling the REFCOR was that each part of the reading exam should be present in the corpus with equal importance regarding the number of task types and the number of words in each task type. Consequently, the REFCOR contains twelve general English texts, whose length measures 7,098 words. The precise distribution is shown in Table 3.

Part 1	**Part 2**	**Part 3**	**Part 4**
Unit 6: 557 words	Unit 1: 638 words	Unit 3: 706 words	Unit 4: 588 words
Unit 12: 620 words	Unit 9: 569 words	Unit 13: 567 words	Unit 14: 592 words
Unit 21: 605 words	Unit 19: 579 words	Unit 20: 504 words	Unit 17: 573 words
1,782 in total	**1,786 in total**	**1,777 in total**	**1,753 in total**

TABLE 3: The REFCOR: the general English texts chosen from the 9th-graders' FCE course book (Prodromou, 1998) and the lengths of the texts given in words.

3.1.3 THE SIZE OF THE CORPUS

Biber and Conrad (2009) define a large corpus as a set of texts or excerpts whose length in total approximates a million words. Since the length of the BIOCOR, 7,021 words in sum, does not come close to this benchmark, the present corpus is considered to be a mini-corpus in Biber and Conrad's (2009) terminology. A mini-corpus was adopted in the present research since its application has several advantages in the particular educational environment. It is commonly believed that the larger the corpus, the more representative register features can be unveiled. However, this notion is valid only in the case of describing general language use (Sinclair, 1991). While examining the language use of a specific area, the use of a mini-corpus is advised. O'Keffee and McCarthy emphasize that a carefully compiled mini-corpus, whose representativeness is high in the particular register, serves as "a powerful tool for the investigation of special uses of language, where the linguist can 'drill down' into the

ifications. The immersion programme's principle behind not updating the mock exams served a practical reason: the majority of the resources (practice books and test samples) available at the school were published before 2008 and it would have been far too costly to replace them. The updating of the educational resources was impossible in the state-run school. (The same holds for the biology textbook (Roberts, 1981) and the English course book (Prodromou, 1998), which were published more than three and two decades ago, respectively.)

data in immense detail" (2010, p. 6). They also pinpoint the manageableness of a mini-corpus as a further advantage. Besides, a mini-corpus is proven to show a higher rate of pedagogical relevance (Ma, 1993) and as a consequence, it is appreciated for bringing insights which can be used for specific learning purposes (Flowerdew, 2002). Additionally, a mini-corpus is more useful for collecting linguistic data for non-native learners (Howarth, 1998). Other scholars find the compilation of a mini-corpus suitable from the point of view of the student, as it offers an easier grasp and tends to be more learnable (de Beaugrande, 2001). Low-frequency items can also be investigated in the study of a mini-corpus, which are unlikely to be explored in the case of a large corpus (O'Keffee & McCarthy, 2010). Finally, the analysis of a mini-corpus renders a close link between the corpus and the context possible (Biber & Conrad, 2009), as the texts are not de-contextualized.

4. RESULTS AND DISCUSSION

4.1 KEYNESS VALUES

The lexical uniqueness of a corpus can be described with a high rate of effectiveness by keyness values, which compare the frequency of lexical items in the corpus with that in the reference corpus (Xia & McEnery, 2005). The across-register nature of the method allows for the comparison of two registers, for pinpointing lexical characteristics that distinguish one register from another. From the point of view of the ESL teacher, the statistical comparability of the uniqueness of the language use of two registers provides directly applicable data since it is not only frequently occurring words that characterise a register (and thus urge the need to be covered in a biology ESP course) but high-keyness tokens too. In the present research, gaining information about the markedly different lexis of the BIOCOR was considered to be beneficial since the collection of key words can indicate what kind of lexical challenges 10th-grade students (who read through the reference corpus when pursuing their studies in the 9th grade) meet when processing the biology texts. Lexical items which are not register specific, ones which occur with similar frequencies in both corpora, are not compiled in the process of keyness comparison. For this reason, the high-frequency lexical item *animal,* for example, does not occur among the key words since the BIOCOR tends to use this item nearly as often as the REFCOR. In contrast, low-frequency words with a high keyness value, ones which are register specific compared to the reference corpus, are entered in the list. For instance, the token *host*, which appears no more than eight times in the biology corpus, but whose keyness is still outstandingly high (k=38), was compiled in the study. It is important to note that the more common lemma *call* has a similar keyness value (k=45) to that of the token *host* despite the fact that it appears nearly eight times more often in the BIOCOR. The obvious reason behind the stark difference in frequency is the second token's fairly common appearance in the REFCOR, against which keyness was computed. A similar pattern can be seen in the case of the more ordinary word *food,* which is applied 46 times in the BIOCOR, and has a similarly significant key value (k=30) as the biology term *intestine* (k=29), which is used seven times less frequently in the BIOCOR.

4.2 POSITIVE KEYNESS

The biology register is described here through listing lemmatized key words in their order of outstandingness. The key words are organized in three categories: biology terms, academic English and general English lexis (see Section 3.1.1). The correlation between lexical items with significantly high keyness values and their level of frequency in the BIOCOR was examined and displayed in Table 4, where the particular frequency bands are also indicated (for the methods of developing the ten frequency bands see Section 3.1.1). Finally, the lexical environments of the biology key words are also uncovered.

The overwhelming majority of the lexical items that differentiate the BIOCOR from the REFCOR are general English tokens. More than half of the lemmas that have a significantly high keyness value belong to general English lexis. There is one single token with significantly high keyness value that belongs to academic English, the lemma *process*. Besides the 60% general English tokens, a great bulk of biology terms (38% of all the key words) also appears as register-distinguishing. A larger part of the 15 key words that belong to the category of biology terms appears with dominantly high frequency in the biology corpus. Eight of them are in the range of the most frequently occurring word families in the BIOCOR, and correspondingly fit into the first three bands of frequency. The high frequency of the register-distinguishing biology key words indicates that the BIOCOR uses its register-specific lexis lavishly. Only seven of the biology key words are used less commonly in the BIOCOR, whose frequency bands range from four to eight. In their order of keyness value, the less frequently used key words are *host*, *segment*, *genus*, *intestine*, *drug*, *gut*, and *agar.* Two among these key lemmas, *segment* and *gut*, appear relatively more recurrently than the others, which indicates that these two word families are more often used in the REFCOR than the other five.

The lexical environments of the eight biology key words that belong to the first three frequency bands (*bacteria*, *virus*, *tapeworm*, *parasite*, *amoeba*, *cell*, *malaria*, and *blood*) were described in an earlier study examining the prevalent lexis of the same corpus (Borza, 2014), thus they are not repeated here. The highest keyness value token among the less frequently appearing lemmas, *host* (k = 38), shows a narrow range of word combinations (see Table 5). It tends to form noun phrases, such as *intermediate host,* or more typically genitive constructions, *the host's digestive food*, or *the host's digestive juices*, and *the host's faeces.* Even greater scarcity is displayed by the verbs it combines with, the single example of the verb with which it appears together is *carry*.

The second highest keyness value word among the less frequent biology terms, *segment* (k = 34), combines in a rich manner (see Table 6). It appears in various noun phrases, such as *gut segments,* or *mature segments* and shows an even more diverse set of verbs it collocates with (e.g., *pass a segment*, *produce segments*, *segments drop off*, or *segments mate*). The token does not disagree with the passive voice either, even if the BIOCOR displays no more than one single example of it (*the body is divided up*

Key word	Keyness (k value)	Type	Raw frequency	Band
bacteria	136,6856537	biology term	41	1
virus	83,84221649	biology term	34	1
tapeworm	72,29248047	biology term	18	3
parasite	68,68048096	biology term	57	1
body	68,47974396	general English	35	1
mosquito	60,57646179	general English	17	3
amoeba	60,50664139	biology term	20	2
plant	54,08612823	general English	40	1
organism	53,18547821	general English	55	1
name	48,05667114	general English	32	1
call	45,21485138	general English	62	1
substance	43,9251976	general English	19	3
cell	41,82590866	biology term	51	1
malaria	40,34447479	biology term	19	3
host	38,4834137	biology term	8	6
live	36,06760406	general English	55	1
group	35,42422485	general English	11	5
figure	33,97449493	general English	39	1
segment	33,66395569	biology term	13	4
worm	33,66395569	general English	20	2
key	33,58996582	general English	11	5
thing	32,68690109	general English	25	2
water	32,53123093	general English	15	3
genus	30,98904419	biology term	8	6
process	30,39152718	academic English	11	5
food	30,23670197	general English	46	1
blood	30,08574486	biology term	19	3
intestine	28,84708214	biology term	7	7
drug	28,22444725	biology term	7	7
energy	27,01064491	general English	8	6
gut	26,9503727	biology term	11	5
earthworm	26,9503727	general English	7	7
cavity	26,9503727	general English	5	9
John	26,2686615	general English	8	6
agar	24,25426292	biology term	6	8
sickness	24,18762207	general English	6	8
sleep	24,18762207	general English	8	6
disease	24,18762207	general English	17	3
bloodstream	24,18762207	general English	6	8
egg	24,03279495	general English	14	4

TABLE 4: Key words and their frequency in the BIOCOR.

In a noun phrase	*the host's digested food*
	the host's digestive juices
	the host's faeces
	intermediate host
Verb it collocates with	*an intermediate host carries it*

TABLE 5: Lexical environment of the biology term *host* in the BIOCOR.

In a noun phrase	*gut segments*
	mature segments
	new segments
	the youngest segment
Verb it collocates with	*to pass a segment*
	produce segments
	rings divide the body up into segments
	segments drop off
	segments mate
	segments reach the rear end of the worm
With a verb in the passive voice	*the body is divided up into a series of segments*
Prepositional phrase	*in each segment*

TABLE 6: Lexical environment of the biology term *segment* in the BIOCOR.

into a series of segments). The lemma also shows the possibility of being combined in a prepositional phrase (e.g., *in each segment*).

The lemma *genus*, with the third highest keyness value (k=31), shows a rather scarce variety of collocations. It combines only in noun phrases and verb phrases (see Table 7). There is one single noun with which it goes together in the BIOCOR (*name*). In a similar fashion, neither is the number of verbs it collocates with more numerous, since it is used in no more than one verb collocation, with the verb *belong*.

In a noun phrase	*genus name*
	name of the genus
Verb it collocates with	*belong to a genus*

TABLE 7: Lexical environment of the biology term *genus* in the BIOCOR.

The token *intestine*, with an outstandingly high keyness value (k=29), forms word combinations within a narrow range (see Table 8). It appears in noun phrases which refer either to its type, *large intestine* or *small intestine*, or to its structure, *wall of the intestine*. The number of verbs it combines with is even less manifold; the lemma appears only within one verb phrase, *live in the intestine*.

In a noun phrase	*large intestine*
	small intestine
	wall of the intestine
Verb and a prepositional phrase	*live in the intestine*

TABLE 8: Lexical environment of the biology term *intestine* in the BIOCOR.

The next significantly high keyness value item (k=28), *drug*, is applied in the BIOCOR in a slightly more versatile way (see Table 9). It forms verb combinations both in the active voice (*drugs save lives*) and in the passive voice (*drugs are taken* and *be treated with certain drugs*). Also, the lemma is capable of forming an adjective phrase with *resistant*.

Verb it collocates with	*drugs save lives*
Adjective it collocates with	*resistant to drugs*
With a verb in the passive voice	*drugs are taken*
	be treated with certain drugs

TABLE 9: Lexical environment of the biology term *drugs* in the BIOCOR.

The lemma *gut*, with a high keyness value (k=27), shows a diverse set of lexical collocations (see Table 10) in the BIOCOR. It appears in various noun phrases (e.g., *human gut* and *gut wall*) and verb phrases alike (*the gut has a special region*, or *gut segments contain*). Besides, the token is also used as a reference to location in prepositional phrases, such as *above the gut* or *beneath the gut.*

In a noun phrase	*animal's gut*
	human gut
	gut wall
Verb it collocates with	*the gut has a special region*
	gut segments contain
Prepositional phrase	*above the gut*
	beneath the gut
	in the gut

TABLE 10: Lexical environment of the biology term *gut* in the BIOCOR.

The last lemma with a significantly high keyness value (k=24), *agar*, appears in relatively few combinations (see Table 11) in the BIOCOR.

There is one single noun phrase it forms (*agar jelly*), and its verb collocations is no more miscellaneous, there being only one verb with which it collocates in the passive voice (*the agar is put in petri dish*).

In a noun phrase	*agar jelly*
With a verb in the passive voice	*the agar is put in petri dish*
Verb and a prepositional phrase	*grow bacteria on the agar*
	put bacteria on the surface of the agar

TABLE 11: Lexical environment of the biology term *agar* in the BIOCOR.

Besides the above listed biology terms, the BIOCOR contains no other subject specific terms with significantly high keyness values. Among the general English items with high keyness value, however, there are two lemmas worthy of attention. The token *figure* (k=34) is notable from the point of view of the ESL and biology ESP teacher, since its meaning in the BIOCOR (*data* or *number*) is different to a great degree from its similarly-formed Hungarian version (*figura*, which only conveys the meaning of bodily shape). The other conspicuous high-keyness word (k=26) is the proper noun *John*, which is hardly expected to be an item distinguishing the biology register from the register of general English reading tasks. The reason behind the high rate of appearance of the proper noun in the BIOCOR compared to its use in the REFCOR is the fact that the biology texts incline to use vivid sample situations instead of providing theoretical explanations for the teenage target readers. The exemplified imaginary person in these situations is called John, which makes the lemma's frequency of occurrence in the BIOCOR extremely high.

4.3 NEGATIVE KEYNESS

Lemmas with high negative keyness value reveal words which are systematically untypical in a particular register compared to a reference corpus. In the present study, lemmas with high negative keyness value show the set of lexical items which occur in the REFCOR but are significantly less often used in the BIOCOR. In other words, tokens with high negative keyness value shed light on a special group of words which 9th-grade students process during their general English studies: it is the collection of word families which are underrepresented (or not present at all) in the biology texts the students read the following term. The findings of running the keyword application of WordSmith version 5 (Scott, 2008) strikingly show that the BIOCOR contains no such item. Notably, there is not one single lemma in the BIOCOR with significantly high negative keyness value when compared to the REFCOR.

4.4 HIGH-FREQUENCY LOW-KEYNESS WORDS

It is not insignificant to take note of the fact that the BIOCOR encompasses a great many frequently occurring lemmas that do not appear among the word families with high-keyness value (for an extensive list of words frequently applied in the BIOCOR see Borza, 2014). This group of words, the set of high-frequency low-keyness items, show that the majority of the frequently used lexis of the BIOCOR is present in the REFCOR with a similar rate of frequency. Table 12 displays the collection of all these words, shows each item's frequency expressed in frequency bands, as well as the type of the lexical item (biology term, academic English or general English). It can clearly be seen that the group of high-frequency low-keyness words embraces nearly exclusively general English terms; only two instances of biology terms occur (*growth* and *reproduce*) and there are no academic terms at all.

Key word	Type	Band
growth	biology term	1
animal	general English	1
get	general English	1
reproduce	biology term	2
do	general English	2
make	general English	2
person	general English	2
small	general English	2
way	general English	2
see	general English	3
use	general English	3
cause	general English	3
contain	general English	3
move	general English	3
place	general English	3
shape	general English	3
water	general English	3

TABLE 12: High-frequency low-keyness words in the BIOCOR.

5. CONCLUSION

The present study investigated the lexical uniqueness of biology texts for secondary students used in a bilingual secondary school in the 10th grade (BIOCOR) from the point of view of English language teaching. The pedagogically-driven research analysed the lexical characteristics of the BIOCOR by unveiling its high-keyness value lexical items. The aim of discovering and describing the key lexical features of the BIOCOR was twofold: i) to gain insights into the lexical difficulties which might pose obstacles to 10th-grade students in smoothly processing the corpus and ii) to collect information about the lexical uniqueness of the corpus which is applicable for ESL (English as a second language) and ESP (English for specific purposes) teachers in the process of building the lexical dimension of the syllabus of an intensive language preparatory course and that of a biology ESP course, respectively.

The keyness characteristics of the BIOCOR provide revealing information about the register of the biology textbook for secondary school students.

1) The keyness results uncover that there is a nearly absolute scarcity of academic words among the key lexis. That is, the lexis of the BIOCOR can hardly be distinguished from that of the REFCOR on account of the use of academic English terms.

This finding goes contrary to the expectations of the biology teachers and the students of the bilingual programme alike, who expressed their certainty about the biology texts being abundant in academic vocabulary, which makes the register of biology texts starkly different from other registers in their perception (Cserép, 1997). Based on the findings, the difficulty of processing the BIOCOR cannot originate from the texts' extensive use of academic English.

2) The great majority of biology key words appear with high frequency in the BIOCOR. This indicates that 10th-grade students are expected to read a string of texts which contains recurrently repeated biology key words. In other words, the students' difficulty of processing the biology texts is hard to be accounted for by the students' unfamiliarity with the biology lexis due to the sporadic appearance of the specific lexis.

3) The noun *John* is used so extensively in the BIOCOR that it appears among the key words. The abundance of the proper noun demonstrates that the register intends to convey and clarify its subject information more through practical examples than through highbrow, scholarly theoretical lines of thought. This tendency is in line with Shapiro's (2012) findings highlighting the fact that the register of science textbooks written for secondary students is more popularizing than academic. As a result, the language of pre-college textbooks for secondary students, who are non-experts in sciences, is less technical than that of tertiary textbooks, which are used in the discourse community of would-be scientists.

4) The BIOCOR contains no lemmas with high negative keyness value at all, in other words, there are no lexical items which occur significantly less often in the BIOCOR than in the REFCOR. That is, the register of the biology texts cannot be distinguished from the REFCOR in this respect, there are no significantly underrepresented general English lexical items. From the point of view of the ESL teacher, this similarity signifies that the vocabulary of the general English reading tasks assigned in the 9th grade cannot be characterized by a superfluously expanded vocabulary in comparison with the lexis used in the biology texts.

5) Finally, the BIOCOR can be characterized by a bounteous use of not register specific frequently occurring words, which are also frequently present in the REFCOR. This indicates that by the time students start pursuing their biology studies in the 10th grade they have already become familiar with a great part of the lexis of the BIOCOR through reading the texts of the REFCOR in the 9th grade. Thus, processing the REFCOR in the 'zero year' provides a firm linguistic grounding for the students. Considering the lexical dimension of the language preparatory programme, reading the texts assigned in the 9th grade prepares bilingual students substantially for their academic studies in English the following year. To draw pedagogical implications, it is important to point out, however, that the level of difficulty of the general lexis in the BIOCOR does not go beyond the CEFR B2 level. Since the CEFR level of the lexis of the BIOCOR ranges from A1 to B2,[8] language preparatory courses should not necessarily aim at more advanced levels.

[8] The online software developed by the Lifelong Learning Programme of two departments of the University of Cambridge (Cambridge University Press and Cambridge English Language Assessment, http://vocabulary.englishprofile.org) was applied to define the CEFR levels of the particular lexical items.

Considering all the aspects of the keyness results above, the lexis of the BIOCOR can hardly be described as challenging for 10th-grade bilingual students. The BIOCOR fails to show a more intriguing complexity in its key vocabulary than the REFCOR. The results of the research reveal that the lack of specialised uniqueness is prevalent in the BIOCOR with regard to academic English and specific biology terminology. The lexical plainness of the biology textbook can be regarded as one of the linguistic features typical of the register of non-academic but popularizing secondary textbooks. The prevailing lexically straightforward character of the BIOCOR, however, suggests that the perceived challenges 10th-grade bilingual students face during their studies in English are not explicable in terms of the lexis of their textbook; that is, they should stem from a different source.

REFERENCES:

Atkinson, D. (1992). The evolution of medical research writing from 1735 to 1985: The case of the Edinburgh Medical Journal. *Applied Linguistics*, *13*, 337–374.

Atkinson, D. (1999). The philosophical transactions of the Royal Society of London, 1675–1975: A sociohistorical discourse analysis. *Language in Society, 25*, 333–371.

Atkins, S., Clear, J., & Ostler, N. (1992). Corpus design criteria. *Literary and Linguistic Computing, 7*, 1–16.

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

de Beaugrande, R. (2001). Large corpora, small corpora, and the learning of language. In M. Ghadessy (Ed.), *Small Corpus Studies and ELT. Theory and Practice* (pp. 3–28). Philadelphia, PS: John Benjamins.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D. (1989). A typology of English texts. *Linguistics*, *27*, 3–43.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8*, 243–257.

Biber, D. (1995). *Dimension of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. (2001). Multi-dimensional methodology and the dimension of register variation in English. In S. Conrand & D. Biber (Eds.), *Variations in English: Multi-dimensional Studies* (pp. 13–42). London: Longman.

Biber, D. (2006). *University Language: A Corpus-based Study of Spoken and Written Registers*. Amsterdam: John Benjamins.

Biber, D., & Jones, J. (2005). Merging corpus linguistics and discourse analytic research goals: Discourse units in biology research articles. *Corpus Linguistics and Linguistic Theory, 1,* 151–182.

Biber, D., & Conrad, S. (2009). *Register, Genre, and Style*. Cambridge: Cambridge University Press.

Biber, D., & Finegan, E. (1989). Drift and the evolution of English style. *Language*, *65*, 487–517.

Biber, D., & Finegan, E. (1994a). *Sociolinguistic perspectives on register*. New York: Oxford University Press.

Biber, D., & Finegan, E. (1994b). Multidimensional analyses of authors' styles: Some case studies from the eighteenth century. In D. Ross & D. Brink (Eds.), *Research in Humanities Computing* (pp. 3–17). Oxford: Oxford University Press.

Biber, D., & Finegan, E. (1997). Diachronic relations among speech-based and written registers in English. In T. Nevalainen & L. Kahlas-Tarkka (Eds.), *To Explain the Present: Studies in the Changing English Language* (pp. 253–275). Helsinki: Societe Neophilologique.

Biber, D., Gray, B., & Staples, S. (2014). Predicting patterns of grammatical complexity across language exam task types and proficiency levels. *Applied Linguistics, 37*(5), 1–31.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, *36*, 9–48.

Borza, N. (2013). Register analysis of biology texts: A corpus-based exploratory study of grammar. *Working Papers in Language Pedagogy, 7*, 29–47.

Borza, N. (2014). Does specific lexis make biology texts difficult? A corpus-based lexical analysis of the register of biology texts. *Practice and Theory in Systems of Education, 9*(2), 181–196.

Borza, N. (2015). Analysing ESP texts, but how? *Practice and Theory in Systems of Education, 10*(1), 1–15.

Borza, N. (2016). The bewildering complexity of the biology register: An investigation into the syntactic complexity of secondary biology texts. *Research in Corpus Linguistics*, *4*, 9–24.

Brown, P., & Fraser, C. (1979). Speech as a marker of situation. In K. R. Scherer & H. Giles (Eds.), *Social Markers in Speech* (pp. 33–62). Cambridge: Cambridge University Press.

CEFR: Council of Europe. Language Policy Unit, Modern Languages Division. (1996). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Strasbourg: Cambridge University Press.

Conrad, S. (1996). Investigating academic texts with corpus-based techniques: an example from biology. *Linguistics and Education*, *8*, 299–326.

Conrad, S. (2001). Variation among disciplinary texts: A comparison of textbooks and journal articles in biology and history. In S. Conrad & D. Biber (Eds.), *Variations in English: Multi-dimensional Studies* (pp. 94–107). London: Longman.

Conrad, S. (2015). Register variation. In D. Biber & R. Reppen (Eds.), *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press.

Conrad, S., & Biber, D. (2001). *Variations in English: Multi-dimensional Studies*. London: Longman.

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

Cserép, S. (1997). *Technical terms in biology. An investigation into scientific English.* Unpublished master's thesis, Budapest: University of Economic Sciences.

Csomay, E. (2005). Linguistic variation within university classroom talk: A corpus-based perspective. *Linguistics and Education*, *15*, 243–274.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*, 61–74.

Flowerdew, J. (Ed.). (2002). *Academic Discourse*. Harlow: Longman.

Forchini, P. (2012). *Movie Language Revisited: Evidence form Multi-dimensional Analysis and Corpora.* Bern: Peter Lang.

Grieve, J., Biber, D., Friginal, E., & Nekrasova, T. (2011). Variation among blogs: A multi-dimensional analysis. In A. Mehler, S. Sharoff & M. Santini (Eds.), *Genres on the Web* (pp. 303–322). Dordrecht: Springer.

Herring, S. C. (1996). *Computer-mediated Communication: Linguistics, Social and Cross-cultural Perspectives*. Amsterdam: John Benjamins.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Kanokshilapatham, B. (2007). Rhetorical moves in biochemistry research articles. In D. Biber, U. Connor & T. Upton (Eds.), *Discourse on the Move* (pp. 73–103). Amsterdam: John Benjamins.

McEnery, A., Xiao, R., & Tono, Y. (2006). *Corpus-based Language Studies*. London: Routledge.

Ma, K. C. (1993). Small-corpora concordancing in ESL teaching and learning. *Hong Kong Papers in Linguistics and Language Teaching*, *16*, 11–30.

Nagy, W., Anderson, R., Schommer, M., Scott, J., & Stallman, A. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, *24*, 262–281.

Nesi, H., & Basturkmen, H. (2006). Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics*, *11*(3), 283–304.

O'Keffee, A., & McCarthy, M. (2010). *The Routledge Handbook of Corpus Linguistics*. London: Routledge.

Prodromou, L. (1998). *First Certificate Star*. Oxford: Macmillan Publishers Limited.

Reaser, J. (2003). A quantitative approach to (sub)registers: The case of sports announcer talk. *Discourse Studies, 5*(3), 303–321.

Reppen, R. (2001). Register variation in student and adult speech and writing. In S. Conrad & D. Biber (Eds.), *Multi-dimensional Studies of Register Variation in English*, 187–199. Harlow: Pearson Education.

Roberts, M. B. V. (1981). *Biology for Life*. Surrey: Thomas Nelson and Sons.

Scott, M. (2008). WordSmith Tools (Version 5) [Computer software]. Liverpool: Lexical Analysis Software.

Shapiro, A. R. (2012). Between training and popularization: Regulating science textbooks in secondary education. *Isis*, *103*(1), 99–110.

Sinclair, J. (2004). *Trust the Text: Language, Corpus and Discourse*. London: Routledge.

Staples, S., Biber, D., & Reppen, R. (2018). Using corpus-based register analysis to explore the authenticity of high-stakes language exams: A register comparison of TOEFL iBT and disciplinary writing tasks. *The Moderns Language Journal, 102*(2), 310–332.

Thain, M., & Hickman, M. (2004). *The Penguin Dictionary of Biology.* London: Penguin Books.

Tribble, C. (1999). *Writing Difficult Texts*. Unpublished doctoral dissertation, Lancaster: Lancaster University.

Vilha, M. (1999). *Medical Writing: Modality in Focus*. Amsterdam: Rodopi.

West, M. (1953). *A General Service List of English Words*. London: Longman.

Xia, Z., & McEnery, A. (2005). Two approaches to genre analysis. *Journal of English Linguistics*, *33*(1), 62–82.

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

Natalia Borza | Pázmány Péter Catholic University, Budapest, Hungary
<nataliaborza@gmail.com>