DOES SPECIFIC LEXIS MAKE BIOLOGY TEXTS DIFFICULT? A CORPUS-BASED LEXICAL ANALYSIS OF THE REGISTER OF BIOLOGY TEXTS

© Natalia Borza
(Eötvös Loránd University, Budapest, Hungary)

nataliaborza@gmail.com

Received: 01.08.2013; Accepted: 19.12.2013; Published online: 16.08.2014

While considerable research has been conducted on analysing biology related research articles, relatively little is known about the register of biology textbooks written for secondary school students. To fill this lacuna, the present study aims to explore biology textbooks from a lexical point of view. This research sets out to describe the nature of the typical lexis prevalent in English language biology texts written for secondary school students. The purpose of the study is to gain insights into one of the possible reasons for which 10th grade bilingual students might find studying biology in English difficult. Within the register perspective, data was collected through a representative mini-corpus, and the frequency of lexical items was computer counted by using text analysing software program WordSmith version 5. In the process of analysis, words of the same root were lemmatized in order to find the most common specific word families of the register. Individual words and lemmas were manually sorted out into any of the following three categories: biology terms, academic English vocabulary, and general English lexis. Using the KWIC (key word in context) application of the same software, collocations of the specific lexis were searched. By describing the lexical environment of specific lexis, it is intended that the study will serve the purpose of giving assistance to biology ESP teachers. The findings of the research reveal that the biology corpus does not abound in biology terms; the use of academic English is rare; and it is general English lexis that is massively present in the register. The results imply that the use of specific lexis does not count for the difficulty bilingual students face when processing biology text. Implications for further investigation are discussed.

Keywords: academic English, bilingual education, English for specific purposes (ESP), register analysis, specific lexis, textbook analysis

Rationale and the research question

The aim of the present study is to describe the typical lexis prevalent in English language biology texts written for secondary school students. The analytical framework conceived in a pedagogical perspective is unique of its kind as to my knowledge no lexical investigation has been carried out so far on the register of biology textbooks from the point of view of the EFL teacher in particular.

This pedagogically motivated empirical research is the result of having observed a discrepancy students face at a bilingual secondary school in Budapest, Hungary. Pupils at the end of their first year at the school take the intermediate level Cambridge examination (level B2 in the Common European Framework of Reference for Languages), the First Certificate in English (FCE). Students who pass this exam are expected to be able to study academic core subjects in English, such as mathematics, history, geography, physics and biology. However, when it comes to studying various subjects in the 10th grade in English, students face considerable difficulties. Although at this point they generally find most subjects difficult to follow in English as a foreign language, biology was chosen to be investigated here in particular, as its status in the school differs from that of the other subjects. During the language preparatory year, even complete beginner students have the chance to master English as a foreign language in no less than twenty English classes a week. This highly intensive language course contains sixteen classes of general English besides four specialized classes: one history ESP, one mathematics ESP, one physics ESP and one geography ESP a week. However, there is no biology ESP provided for the students in the 9th grade. The reason behind this practice is that the special terminology of biology is considered to be too diverse and difficult to grasp for 9th graders by the biology teachers working at the school. This means that in the 10th grade students attending biology classes delivered in English rely on the knowledge they gained in their *general* English studies and the *other* four English classes for specific purposes. Besides this, an interview study conducted at the school (Cserép, 1997) revealed that it is the language of biology bilingual students find the most challenging among all the subjects taught in English. As a teacher of general English in the 9th grade, and having been informed of the lexical problems students encounter when learning core subjects in English, I have become increasingly interested in what exactly the students need to know in order for them to be able to handle biology texts successfully in the 10th grade in terms of lexis. The analysis intends to gain insights into one of the possible reasons, namely lexical challenges, for which 10th graders might find studying academic subjects, such as biology in English difficult. With such a pedagogical aim, the present case study attempts to answer the following research question:

What lexical items characterise the biology textbook used by 10th grade students in a bilingual secondary school?

Theoretical background

This case study investigates the discourse characteristics of a string of biology texts using the register approach. The research follows the de *Beaugrandian tradition* by treating a text as a communicative event and by using the term discourse synonymously with text (de Beaugrande, 1997).

Genre and register

Biology textbooks are written for a specific audience with a specific purpose. As a result, its language use and structures vary considerably, due to which it can be treated as being a distinct type of text. The *Swalesian* definition of genre (Swales, 1990:46), distinguishes biology textbooks from other textbooks since the previous one is clearly used in different communicative events. Consequently, it can be regarded as belonging to a distinct genre.

In a similar manner, the concept of register used by Biber and colleagues (1998), as a "*cover term for varieties defined by their situational characteristics*" considering the "*purpose, topic, setting, interactiveness, mode, etc.*" of the situation (1998:135), can also be applied in order to differentiate biology textbooks from other textbooks. As biology textbooks and other textbooks fail to share the same purpose, topic, setting and interactiveness, biology textbooks represent different registers in the Biberian sense. In more general terms Biber et al. claim that discourse analysts working in the field of ESP uncover "*specialized registers in English*" (Biber, 1998:157), revealing their stance of ESP being a different register from general English. Biber and Finegan (1994) remark that a distinction can be made between two registers based on the fact that their "*identifying markers of language structures and language use differ from the language of other communicative situations*" (1994:20). According to Halliday (1988), the systematic study of registers is possible since their clusters of "*associated features have a greater than random tendency to co-occur*" (1988:162).

It was Biber and Conrad (2009) who distinguished between the overlapping concepts of register and genre by treating them as two different approaches for text analysis. In their terminology, the genre approach examines rhetorical organisation and linguistic characteristics that structure whole texts. Such structural features might occur only once in the text, for instance, the abstract of a research article, or the title of a chapter in a textbook. Given that a certain structural element might as well occur not more frequently than once, studies in the genre approach investigate complete texts. On the other hand, the register approach to text analysis has a different focal point. It centres on words and grammatical features that are frequent in the representative text excerpts. As a result, an analysis with the register approach can be based on the collection of excerpts of texts instead of investigating entire texts.

The present study explores eight complete texts of a biology textbook. The fact that full texts are analysed might suggest that the genre approach was chosen in the research. Still, the current study takes the register perspective, as the research focuses on uncovering frequent lexis prevalent in biology texts, instead of discovering their overall structure or the rhetorical structures of certain parts of the texts.

Academic English

A set of wide-ranging lexis typically used in academic English was systematically collected by Coxhead (2000) for pedagogical purposes. The list of 570 word families was compiled in order to provide insights for teachers preparing students for their tertiary studies in English as to what special lexis appears frequently in academic texts. The corpus in which the frequency of words was run by Coxhead embraced twenty-eight subject areas. Only those word families were involved in the list that appeared in over half of the areas, which ensured that the list could be used for any academic subject area. In developing the list, frequency played a key role, word families that were used more than 100 times in 3,500,000 words were short listed. Basic vocabulary, words that are among the 2,000 words of English, were taken off from the short list, as academic reading presupposes the learner's familiarity with such vocabulary. Besides, proper nouns, for example names of places and people, as well as Latin forms, such as *etc, i.e.,* were also removed from the short list of English academic vocabulary. Finally, the list was organized into ten sublists according to the frequency of the given word family, that is, sublist one contains the most common academic words in the corpus. The present case study uses Coxhead's (2000) findings in order to see whether the biology texts assigned to 10th grade students in a bilingual secondary school are difficult to read for the fact that they contain a large number of academic lexical items. This study will collect academic English vocabulary that is used in the biology texts 10th graders process in order to be of assistance to biology ESP teachers.

Earlier research on biology texts

There has been extensive research in the field of register analysis; numerous written registers, including biology related texts, have been described with various aims applying different frameworks. Following the genre approach, the rhetorical structure of biology research articles has already been analysed. The difference in discourse units in biology research articles (Biber & Jones, 2005) and the variations among moves within biochemistry research articles (Kanoksilapatham, 2005) have been explored. However, to my knowledge, no register study has yet been conducted to reveal the lexical characteristics of the register of biology textbooks written for secondary school students. Regarding the level of abstraction, lexical characteristics of science textbooks were studied by Wellington (1983), while it was Kukemelk and Mikk (1993) who measured the frequency of specific lexis in biology textbooks. However, no comprehensive register study has been done to describe the genre of textbooks for secondary school students, in particular, that of biology textbooks from a lexical point of view.

Methods

The corpus

The size of the collection of biology texts under investigation may only be considered as a mini-corpus since the number of words it contains does not come close to a million. This mini-corpus was selected for the current investigation as a carefully targeted corpus that represents a particular register (O'Keffee & McCarthy, 2010:6). In contrast to describing general language use, where the bigger the size of the corpus, the more representative patterns can be revealed, in the case of examining a specific area of the language a small corpus is advised to be compiled for several reasons. On the one hand, a mini-corpus is more manageable to handle (O'Keffee & McCarthy, 2010), it also has a higher rate of pedagogical usefulness (Ma, 1993), besides, it tends to yield insights which can be used for specific learning purposes (Flowerdew, 2002), moreover, it can be used for teaching non-native learners (Howarth, 1998), additionally, it is more 'learnable' (de Beaugrande, 2001), furthermore, all occurrences, including low-frequency items, can be examined, and a close link between the corpus and the context can be established as the language use is kept intact in the sense that the texts are not de-contextualised (O'Keffee & McCarthy, 2010).

In order to make the corpus representative in terms of what 10th grade students might have lexical difficulties with, it was first checked which biology texts bilingual students in the 10th grade are expected to read and process in their first academic term. In a structured group interview with five high-achieving 10th graders in English, students were given their biology textbooks (Roberts, 1981) to pick the topics covered in the autumn term. From the class of 10th graders, high-achievers in English were chosen as low-achievers tend to be more reluctant to share information about their studies; besides, they also have a tendency not to remember precisely what has been covered in class. Each interviewee chose the same eight chapters, see Appendix A. To affirm the students' choices of the texts, the topics of the biology classes were followed in the electronic register of the school written by the students' biology teacher from September to mid-January, and it was observed that the list compiled by the students was exhaustive. Next, the eight chapters were typed in order to make them computer analysable.

Procedures of data collection and analysis

After having compiled the corpus, the hard copies of the texts were digitalised by use of keyboarding to carry out the following analyses in order to tap the special language use of the register of biology texts as far as its lexis is concerned. The frequency of words in the corpus of the biology texts was computer counted by using text analysing software program WordSmith version 5 (Scott, 2008) in order to find the most common lexical elements of the register. Words of the same root were lemmatized so that it was the frequency of word families determined, not that of individual word forms. Lemmatization was considered to be crucial as it is more valuable for ESP teachers to possess knowledge about the frequency of word families than that about conjugated verb forms or different word formations when it comes to working out the lexis part of ESP syllabi. Lemmatization allowed the following different word forms to be considered as one batch:

- ➢ singular and plural forms, e.g. bacterium – bacteria, flagellum – flagella, phylum – phyla, mosquito - mosquitoes;
- ➢ nominative and genitive forms, e.g. female – female's;
- ➢ verbs in different tenses, e.g. attach – attached, kill – killed, know – known;
- ➢ verbs and gerunds, e.g. borrow – borrowing;
- ➢ base, comparative and superlative adjectives, e.g. large – larger – largest;
- ➢ word formation: amoeba – amoebic, blood – bleeding, chemicals – chemically, class – classify, contract – contractile – contraction, dead – death – die, digestive – digested – digestion, granules – granular, saliva – salivary, slime – slimy.

Compound words, however, were not joined in one batch. Respectively, *cow* and *cowslip*, *flat* and *flatworm*, *Mary* and *marigold*, *stream* and *streamlined* for example were computer counted separately. After lemmatization, the most common words in the biology corpus were listed in rank order, arranged and displayed in bands of frequency. Band 1 contains the most ubiquitous, most typical words in the text, the ones that appear minimum 30 times in the investigated chapters of the biology book, while Band 10 involves more unusual items, word families that occur only four times in the corpus. Table 1 shows how often items of particular bands appear in the register expressed both in the number of their raw occurrences and in percentages.

Table 1. The frequency bands of the corpus of the biology texts

Rank order	Raw frequency of individual tokens and that of lemmas	Frequency of individual tokens and that 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%

Individual lexical items and tokens that are used fewer than four times in the biology corpus were not compiled in this study. The reason behind this is the presupposition that in an informational, educational register, such as biology textbooks for secondary school students, lexical items of importance occur repeatedly to serve an instructional function.

In each band the individual words and lemmas were manually sorted out into any of the following three categories. First, the category of *biology terms* contains lexical items that carry a specific meaning within the context of biology, a meaning or a shade of meaning which is different from the everyday use of the word. A dictionary of biology (Thain & Hickman, 2004) was applied as the baseline when deciding whether a word should be labelled as biology term or if it is simply a general English word that happens to be related to a certain biology topic. Thain & Hickman's biology dictionary (2004) was chosen since it was written for students and teachers alike, and the dictionary claims to clarify the most essential concepts in biology, including the core vocabulary of several subfields. In this study, lexical items that appear as entries in the biology dictionary were labelled as biology terms. Within a word family, all the members of the lemmatized batch were checked in the dictionary, thus it was ensured that a lexical item was labelled as biology term irrespective of its word class. For example 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 marked as biology term. On the other hand, dictionary entries where a lexical item appears in conjunction with other words, that is, biology terms that contain more words, were not labelled as biology terms unless they appeared in the biology texts with the same word combinations. For instance, the lexical item 'body' is not a separate entry in the biology dictionary, while 'carotid body' is one. Consequently, the word 'body' was not labelled as biology term in the present analysis unless it was used in the biology texts in conjunction with the word 'carotid'. The second category, *academic vocabulary* was assigned to those lexical items that appear on Coxhead's (2000) list of academic vocabulary, a collection of 570 word families particularly compiled for pedagogical purposes. Finally, words that belong neither to the category of biology terms nor to that of academic vocabulary were labelled as *general English* words.

From a pedagogical point of view, it was treated as essential to describe the lexical environments of the most frequent biology terms as our "*knowledge of a word includes the fact that it co-occurs with certain other words*" (Hoey, 2005:8). The lexical environments of the biology terms appearing in the first three bands in the corpus were described by compiling the words that go together with them. In order to look more deeply behind the quantitative results collected through frequency analysis, collocations were searched using the KWIC (key word in context) application of the same software within the range of the boundary of the sentence. Compiling all the word combinations with which the frequent biology terms are used gives the possibility to gain pedagogical implications for biology ESP teachers working out biology ESP syllabi. The words that collocate with the frequent biology terms were sorted out according to their part of speech. To produce an easy-to-follow list, collocations were recorded in an alphabetic order, in their dictionary forms. That is, tenses in which the given verbs that go together with the biology terms were not kept, one can find for instance 'parasites make for John's liver' instead of 'parasites made for John's liver'. In a similar manner, modals that appear in the biology texts were not accounted here, thus 'viruses are released' appears in the description and not 'viruses may be released'. Finally, to keep the descriptive list of lexical environment of biology terms as easy-to-grasp as possible, relative clauses used in the biology texts were also omitted, even if it resulted in a slight change of meaning. Minor changes of information were not considered crucial in the present analysis since the description is of lexical nature. In other words, the main focus of the lexical accounts is to tap the possible collocations used with the frequently applied biology terms, and the descriptions do not attempt to collect information in the field of biology. That is the reason why 'animals transmit parasites' is listed in the study instead of recording 'animals which transmit parasites'.

Results and discussion

In the following section the corpus of the biology texts is described as to the nature of its prevalent lexical items. The lexical environments of the most repeatedly occurring biology terms are also recorded here.

Band 1

After carrying out lemmatization of the words that share the same root, and computer counting the frequency of the word families in the corpus of the biology texts, the results were arranged in frequency bands. The lexical items that appear most recurrently, minimum thirty times in the biology corpus are listed in Band 1. Table 2 contains these items, showing the number of their raw occurrences and also their frequency expressed in percentages. For instance, the very most frequent biology term 'parasite' appears fifty-seven times in the biology texts, which constitutes 0.8% of the corpus. There are five biology terms among the most frequently used lexical items, 'parasite', 'cell', 'bacteria', 'virus', and 'growth'. However, the majority of the typically applied items is general English lexis, not biology terms. Although most of these items are related to the topic of biology, e.g. 'animal', 'plant', 'body', they still do not form specific biology vocabulary. Contrary to the expectations that biology texts are full of academic vocabulary (Cserép, 1997), the band of most frequently used lexical items contains no academic English vocabulary at all.

Table 2. Band 1: the most frequent lexical items in the biology corpus

Biology terms	Academic English	General English
parasite (57; 0.8)		call (61; 0.85)
cell (51; 0.71)		animal (57; 0.8)
bacteria (41; 0.57)		live (55; 0.77)
virus (34; 047)		plant (53; 0.74)
growth (30; 0.42)		food (47; 0.66)
		get (44; 0.61)
		organism (44; 0.61)
		figure (39; 0.54)
		name (36; 0.5)
		body (35; 0.49)

In order to make pedagogical implications for biology ESP teachers, the following subsection describes the lexical environment of the most frequently used biology terms. Table 3 shows all the collocations the lexical item ‘parasite’ takes in the biology corpus. It can be seen that the token appears in various noun phrases, such as ‘life cycle of the parasites’, ‘malarial parasite’, or ‘worm-like parasite’. The term ‘parasite’ is even richer with regard to the verbs it takes, there are fifteen different verbs used with it in the biology corpus. A bit more sparingly, however, it also appears as an object of verbs, for instance ‘kill the parasites’, or ‘transmit parasites’ and with verbs in the passive voice, e.g. ‘parasites are carried to humans’.

Table 3. Lexical environment of the token ‘parasite’

In a noun phrase	life cycle of the parasites
	malarial parasite
	new batch of parasites
	sleeping sickness parasite
	worm-like parasite
Verb it collocates with	parasites attack the blood cell
	parasites become resistant to drugs
	parasite bores its way into a red blood cell
	parasites cause serious diseases
	parasites grow
	parasites leave the liver
	parasites live in wild animals
	parasites make for John’s liver
	parasites move around by flapping a membrane
	parasites multiply
	parasites pass out with the person’s faeces
	parasites reproduce
	parasites split
	parasites undergo multiple fission
	parasites weaken people
As an object of a verb	animals transmit parasites
	get rid of the parasite
	kill the parasites
	the mosquito carries the malarial parasite
With a verb in the passive voice	carry: parasites are carried to humans
	know: known as parasites
	pass: the parasite is passed

The second most frequent biology term, ‘cell’, has plentiful word combinations in the biology texts, see Table 4. It forms numerous noun phrases, including 'cell membrane', 'cell wall' and 'red blood cell' among the more than dozen combinations. However, the variety of verbs it takes in the biology texts is not that vast, including 'burst' and 'contain'. Nevertheless, it has a tendency to function as the object of verbs, for instance 'attack', 'fill', and 'rob'. It is also typically applied with verbs in the passive voice, such as 'bound', 'release', and 'surround'.

Table 4. Lexical environment of the token 'cell'

In a noun phrase	bacterial cell
	cell membrane
	cell wall
	contents of a cell
	leaf cell
	life of the cell
	living cells
	normal cell
	one-celled organisms
	plant cells
	protective cell wall
	red blood cells
	rest of the cell
	single cell
	source of cells
	surface of the cell
	thin cell membrane
	typical cell
Verb it collocates with	the cell becomes dormant
	the cell bursts
	the cell bursts open
	the cell contains
As an object of a verb	attack more cells
	call them cells
	fill the cell
	rob the cell
	see cells
	take a few cells out of an animal
With a verb in the passive voice	bound: the cell is bounded by
	make: living organisms are made of cells
	release: the cell is released
	surround: the cell is surrounded by

The third most frequent token, 'bacteria', has a modest number of collocations in the biology corpus, see Table 5. It appears in noun phrases both as an adjective, e.g. 'bacterial cell' and 'bacterial colonies' and it also functions as the head of the noun phrase, for example 'streptococcal bacteria'. The selection of verbs it takes is wide-ranging, including 'clump', 'multiply', 'survive', and 'vary'. Neither is its appearance as an object of a verb scarce, it is applied for instance with 'grow', 'hold back', and 'remove' among others. It being used with a verb in the passive voice is not typical, however. There are no more than two such combinations, namely 'give' and 'surround'.

Table 5. Lexical environment of the token 'bacteria'

In a noun phrase	bacterial cell
	bacterial colonies
	disease-causing bacteria
	growth of the bacteria
	individual bacteria
	streptococcal bacteria
	type of bacteria
Verb it collocates with	bacteria appear in the microscope
	bacteria clump together
	bacteria make organic food
	bacteria multiply into colonies
	bacteria occur almost everywhere
	bacteria reproduce quickly
	bacteria survive bad conditions
	bacteria vary in their shape
As an object of a verb	get rid of bacteria
	grow bacteria
	hold back the bacteria
	put bacteria on the surface of agar
	remove the bacteria
With a verb in the passive voice	give: bacteria is given moisture
	surround: bacteria are surrounded by

The fourth most repeatedly applied biology term, 'virus', has a humble set of collocations in the biology texts, see Table 6. There is hardly any noun phrase where it is the head, as in 'new virus' and 'structure of the virus'. However, it combines with a fair number of verbs, such as 'attach', 'attack', and 'reproduce'. It is no more than two verbs that take the token 'virus' as an object, namely 'cultivate' and 'grow'. The most numerous collocations of the lexical item are verbs in the passive voice, for instance 'discover', 'form', 'release' and 'set free'.

Table 6. Lexical environment of the token 'virus'

In a noun phrase	new virus
	structure of the virus
Verb it collocates with	the virus attaches itself
	the virus attacks different cells
	the virus comes from inside the cell
	the virus has a simple shape
	the virus reproduces
As an object of a verb	cultivate the virus
	grow viruses
With a verb in the passive voice	discover: viruses were discovered
	form: a new virus is formed
	release: viruses are released
	see: viruses are seen
	set free: viruses are set free

The fifth most recurrent word family in the biology corpus, 'growth', takes a fair number of collocations, see Table 7. In the form of a past participle modifier, it appears in one single noun phrase, 'full-grown earthworm'. The verbs it combines with are related to the time span of growth, for example 'speed up', 'stop' and 'go on'. Signifying an action, it appears both as a transitive verb, for instance increasing the number of 'bacteria' and 'viruses', and an intransitive verb, such as 'living things', 'moulds' and 'worms' become larger. The token as a verb is also typically applied with prepositional phrases, either showing directions, e.g. 'in a particular direction', and 'towards light', or indicating a place e.g. 'on the agar' or showing dimensions, 'to their full size'.

Table 7. Lexical environment of the token 'grow'

In a noun phrase	full-grown earthworm
Verb it collocates with	go on growing
	growth takes place
	growth stops
	speed up their growth
	stop growing
Nouns it collocates with	amoebas grow
	grow bacteria
	grow viruses
	living things grow
	moulds grow
	parasites grow
	worms grow
Verb and a prepositional phrase	grow in a particular direction
	grow on the agar
	grow to their full size
	grow towards light

Band 2

The second most repeatedly applied lexical items in the biology corpus belong to Band 2, which contains word families that appear at least twenty times in the corpus, see Table 8. While Band 1 includes five biology terms, Band 2 comprises no more than two, such as 'amoeba' and 'reproduce'. Similarly to the previous band, the word families of Band 2 are characterized by the abundance of general English lexis, which is four times more prevalent among the lemmas of this band than biology terms. While general English lexis in Band 1 is mostly biology related, general English vocabulary in Band 2 is not closely connected to biology topics. Such items as 'do', 'make', 'take', 'person', 'thing' and 'small' belong to common, basic vocabulary, they are not associated with biology areas at all. It is only the item 'worm' that is related to the field of biology. In the same way as in Band 1, the complete lack of appearance of academic English vocabulary goes contrary to assumptions.

Table 8. Band 2: the second most frequent lexical items in the biology corpus

Biology terms	Academic English	General English
amoeba (20; 0.28)		do (29; 0.41)
reproduce (20; 0.28)		make (28; 0.39)
		take (26; 0.36)
		person (24; 0.34)
		thing (22; 0.31)
		small (21; 0.29)
		way (21; 0.29)
		worm (20; 0.28)

The sixth most frequent biology term, 'amoeba', is used with a small number of collocations in the biology corpus, see Table 9. It appears with no more than three modifiers in a noun phrase, taking the adjectives ‘dysentery’, ‘live’ and ‘ordinary’. The variety of verbs the token combines with is not rich either; what is more, most of the collocating actions denote basic verbs, such as ‘change’, ‘eat’, and ‘live’. In a similar manner, the biology term is narrowly used as an object of verbs; its appearance is limited to ‘examine’ and ‘see’.

Table 9. Lexical environment of the token ‘amoeba’

In a noun phrase	dysentery amoeba
	live amoeba
	ordinary amoeba
Verb it collocates with	amoebas change shape
	amoebas eat organisms
	amoebas grow
	amoebas live in ponds
	amoebas reproduce
As an object of a verb	examine a live amoeba
	see an amoeba

The seventh most recurrent biology term, ‘reproduce’ appears in a twofold way in the biology corpus, see Table 10. The token either combines with a noun phrase as its subject, namely, the living thing that reproduces, e.g. ‘amoeba’, ‘bacteria’, ‘euglena’, ‘parasite’, ‘virus’ or in more general terms ‘offspring’ and ‘organism’; or it collocates with an adverb of manner or a prepositional phrase describing how the reproduction takes place, for instance ‘quickly’, ‘sexually’, ‘by splitting in two’ or ‘on their own’.

Table 10. Lexical environment of the token ‘grow’

Noun it collocates with	amoeba reproduce
	bacteria reproduce
	euglena reproduce
	malarial parasites reproduce
	offspring reproduce
	organisms reproduce
	viruses reproduce
Adverb it collocates with	reproduce quickly
	reproduce sexually
Verb and a prepositional phrase	reproduce by splitting in two
	reproduce by splitting into new individuals
	reproduce on their own

Band 3

The third most frequently used word families, which appear at least fifteen times in the biology corpus, constitute Band 3, see Table 11. Similarly to Band 2, this band scarcely contains biology terms, there being only three of them, such as ‘malaria’, ‘blood’, and ‘tapeworm’. Four times as abundant as the use of biology terms is, however, the appearance of general English vocabulary. Most of the general English lexis in Band 3 is part of common, basic vocabulary, such as ‘see’, ‘use’, ‘cause’, ‘move’, ‘place’, ‘shape’ and ‘water’. It is only a small part of the general English lexical items here that are related to biology topics, for instance ‘substance’, ‘disease’, and ‘mosquito’. Not differing from Bands 1 and 2, this band does not contain a single academic English lexical item. This feature is highly unexpected of the register, as biology textbooks are supposed to use a large number of academic English vocabulary (Cserép, 1997).

Table 11. Band 3: the third most frequent lexical items in the biology corpus

Biology terms	Academic English	General English
malaria (19; 0.27)		see (19; 0.27)
blood (18; 0.25)		substance (19; 0.27)
tapeworm (18; 0.25)		use (18; 0.25)
		disease (17; 0.24)
		mosquito (17; 0.24)
		cause (16; 0.22)
		contain (15; 0.21)
		move (15; 0.21)
		place (15; 0.21)
		shape (15; 0.21)
		water (15; 0.21)

The application of the eighth most recurring biology term, ‘malaria’, is varied to a limited extent in the biology texts, see Table 12. In an adjective form, ‘malarial’, it combines with nouns, such as ‘area’ and ‘parasite’, in addition, it also takes the negative prefix ‘anti’ to form the collocation ‘anti-malarial tablet’. The range of verbs the token collocates with is extremely narrow; there is no more than one single combination in the corpus, apparently with the verb ‘occur’. The scope of the token to function as an object of a verb is wider, there are four such instances, namely, ‘conquer’, ‘get’, ‘have’ and ‘carry’. The passive voice is also typical with the lexical item, it is used in combination with ‘control’, ‘cure’, and ‘spread’ in the biology texts.

Table 12. Lexical environment of the token ‘malaria’

In a noun phrase	anti-malarial tablet
	malarial area
	malarial parasite
Verb it collocates with	malaria occurs
As an object of a verb	conquer malaria
	get malaria
	have malaria
	the mosquito carries the malarial parasite
With a verb in the passive voice	control: malaria is controlled
	cure: be cured of malaria
	spread: malaria is spread by mosquitoes

The ninth most frequently applied biology term, ‘blood’, appears to be bounded in its use in the biology corpus, see Table 13. The token is used variedly in noun phrases; it makes its appearance in ‘blood-sucking tsetse’, ‘blood system’, ‘dorsal blood vessel’ and ‘red blood cell’ among others. However, the lexical item combines in a limited way with verbs. There is no instance of it taking a verb in the corpus at all. Besides, no more than one single verb takes it as an object, namely the phrasal verb ‘suck up’. Scant is the choice of verbs in the passive voice it collocates with, there is no other such instances but ‘pump’.

Table 13. Lexical environment of the token ‘blood’

In a noun phrase	blood-sucking tsetse
	blood system
	dorsal blood vessel
	fluid part of the blood
	main blood vessel
	red blood cell
	system of blood vessels
As an object of a verb	sucks up your blood
With a verb in the passive voice	pump: blood is pumped by the heart

The tenth most typical biology term in the texts, ‘tapeworm’ shows a similarly restricted selection of collocations, see Table 14. The area where it forms collocations multifariously is the noun phrase. It demonstrates a wide range of combinations, for instance ‘beef tapeworm’, ‘life cycle of the tapeworm’ and ‘tapeworm bladder’. Its tendency to combine with verbs is not that diverse, however. There are no more than two examples of it taking a verb; it collocates with the phrasal verbs ‘pop out’ and ‘get round’. Its use as an object of a verb is even more restricted, no other verb but ‘get’ takes it.

Table 14. Lexical environment of the token ‘tapeworm’

In a noun phrase	beef tapeworm
	life cycle of the tapeworm
	pork tapeworm
	structure of the tapeworm
	tapeworm bladder
	tapeworm’s eggs
	young tapeworm
Verb it collocates with	a tapeworm pops out
	the tapeworm gets round this
As an object of a verb	get rid of tapeworms
	get tapeworms

Biology terms and academic English vocabulary that appear fewer than fifteen times in the corpus were also collected in the present study. However, due to the lack of space, they are not recorded here, and thus their lexical environments are not presented either. For a list of specific lexical items, biology terms and academic English vocabulary, which appear minimum four times in the biology corpus, see Appendix B and C respectively. It is worth noting, however, that only thirty-four biology terms and no more than thirteen academic English individual tokens or lemmas were found in the set of texts altogether.

Conclusion

The aim of the present study was to depict what kind of lexical items characterize biology texts written for secondary school students. The goal of the analysis was to gain insights into whether bilingual 10th graders find studying academic subjects in English difficult for lexical reasons, more specifically, for the biology texts being full of biology terms and academic English vocabulary.

The findings of the research clearly reveal that the above lexical reasons do not count for the difficulty bilingual students face when processing the biology texts assigned to them in the tenth grade. The biology corpus does not abound in biology terms; it is more the general English lexis that is massively present in the biology texts. The most frequently used biology terms, the ones that appear more than thirty times in the corpus, show a wide range of collocations. However, biology terms that are repeated less frequently than thirty times in the corpus demonstrate a much more limited, less diverse scope of lexical combination. Although the textbook is written for secondary school students, academic English is infrequently rare, at a more recurrent level even absent in its language use.

Since the examined corpus can hardly be characterized by profusely applying specific lexis, the difficulties tenth grade bilingual students face when processing them cannot be attributed to the abundance of unfamiliar biology terms or academic English vocabulary. Moreover, the main reason for English – Hungarian bilingual students' finding the biology texts difficult can barely be recognized in the texts' specific terminology as many of the anyway small number of specific vocabulary items are similar in the students' mother tongue, in Hungarian. Consequently, further investigation is needed to find out what makes biology texts hard for tenth grade students to understand. It is worthwhile to explore the texts' readability level, their lexical density and grammar use, as well as sentence complexity and text organisation.

References

BIBER, D., & FINEGAN, E. (1994). *Sociolinguistic Perspectives on Register.* New York: Oxford University Press.

BIBER, D., CONRAD, S., & REPPEN, R. (1998). *Corpus Linguistics. Investigating language structure and use.* Cambridge: Cambridge University Press.

BIBER, D., & JONES, J. K. (2005). Merging corpus linguistics and discourse analytic research goals: Discourse units in biology research articles. *Corpus Linguistics and Linguistics Theory, 1,* 151-182.

BIBER, D., & CONRAD, S. (2009). *Register, genre, style*. Cambridge: Cambridge University Press.

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

COXHEAD, A. (2000). A New Academic Word List. *TESOL Quarterly, 34,* 213-238.

DE BEAUGRANDE, R. (1997). *New Foundations for a Science of Text and Discourse: Cognition, Communication, and the Freedom of Access to Knowledge and Society*. Norwood, N.J.: Ablex.

FLOWERDEW, L. (2002). Corpus-based Analysis in EAP. In Flowerde, J. (Ed.), *Academic Discourse* (pp. 95-114). London: Pearson.

HALLIDAY, M. A. K. (1988). On the language of physical science. In Ghadessy, M. (Ed.), *Registers of written English* (pp. 162-178). London: Pinter Publishers

HOEY, M. (2005). *Lexical priming: A new theory of words and language*. London: Routledge.

HOWARTH, P. (1998). Phraseology and Second Language Proficiency. *Applied Linguistics, 19* (1), 24-44.

KANOKSILAPATHAM, B. (2005). Rhetorical structure of biochemistry research articles. *English for Specific Purposes, 24,* 269-292.

KUKEMELK, H., & MIKK, J. (1993). The prognosticating effectivity of learning a text in physics. *Glottmetrica, 14,* 82-96.

MA, K. C. (1993). Small-corpora Concordancing in ESL Teaching and Learning. *Hong Kong Papers in Linguistics and Language Teaching, 16,* 11-30.

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

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

SCOTT, M. (2008). *WordSmith Tools version 5*. Liverpool: Lexical Analysis Software.

SWALES, J. M. (1990). *Genre Analysis. English in academic and research settings.* The University of Michigan: Cambridge University Press.

THAIN, M., & HICKMAN, M. (2004). *The penguin dictionary of biology.* London: Penguin Books.

WELLINGTON, J. J. (1983). A taxonomy of scientific words. *School Science Review, 64,* 767-773.

Appendices

Appendix A The corpus of the biology texts

1. The characteristics of living things
2. Classifying, naming and identifying
3. Amoeba and other protists
4. Bacteria
5. Viruses
6. The earthworm
7. Harmful protists
8. Parasitic worms

Appendix B List of biology terms and their frequency in Bands 4-10

microscope (14; 0.2), gut (11; 0.15), genus (10, 0.14), cytoplasm (9; 0.13), muscle (9; 0.13), nucleus (9; 0.13), poison (9; 0.13), class (8; 0.12), host (8; 0.12), protists (8; 0.12), system (8; 0.12), develop (7; 0.1), digest (7; 0.1), drug (7; 0.1), intestine (7; 0.1), nerve (7; 0.1), stimulus (7; 0.1), agar (6; 0.08), diffuse (6; 0.08), excretion (6; 0.08), flagellum (6; 0.08), photosynthesis (6; 0.08), species (6; 0.08), eye (5; 0.07), liver (5; 0.07), membrane (5; 0.07), phylum (5; 0.07), sperm (5; 0.07), spore (5; 0.07), chlorophyll (4; 0.06), endoplasm (4; 0.06), faeces (4; 0.06), saliva (4; 0.06), vacuole (4; 0.06)

Appendix C List of academic English vocabulary and their frequency in Bands 4-10

investigate (12; 0.17), process (12; 0.17), respond (10, 0.14), vary (9; 0.13), identify (6; 0.08), constant (5; 0.07), release (5; 0.07), feature (4; 0.06), intermediate (4; 0.06), method (4; 0.06), series (4; 0.06), similar (4; 0.06), survive (4; 0.06)