RiCL Research in Corpus Linguistics

The bewildering complexity of the biology register

Natalia Borza
Pázmány Péter Catholic University / Hungary

Abstract – While considerable research has been conducted on the register analysis of English language tertiary textbooks, relatively 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 register of the biology corpus (BIOCOR) that 10th grade students need to process during their studies at a bilingual secondary school. The research reports on the characteristic linguistic features of the BIOCOR with regard to the complexity of the texts syntactic structure. The BIOCOR (consisting of 7,021 words) is compared to a reference corpus (REFCOR) of general English texts at a CEFR B2 level (comprising 7,098 words) by exploring the frequency of ten types of syntactic structures (simple, compound and complex sentences of various number of dependent and independent clauses). The results of the investigation disclose that syntactic simplicity is prevalent in the BIOCOR: simple sentences abound, complex sentences are used in a modest manner, while complex-compound sentences are hardly present in the corpus. The syntactic simplicity 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 – register analysis, sentence complexity, bilingual education, biology textbook for secondary students, popularizing literature

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 observing 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 that 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 that 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 ESP (English for Specific Purposes) 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, but can be meaningfully

transferred and applied by educators working in any English-language international school where the alumni include 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 (hereafter referred to as the REFCOR) 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 (referred to in the study in its acronym form as the BIOCOR for short)? Accordingly, the paper attempts to answer the following research question: what syntactic complexity is characteristic of the BIOCOR in comparison with that of the REFCOR?

2. Review of the literature

The information packed in a written text is expressed through a series of sentences, some of them longer, others shorter. Sherman (1893), one of the forerunners of text analysts measuring the level of difficulty of written discourse, considered whether sentence length has a serious effect on readability. In a longitudinal study, he noticed that the average sentence length of English prose shortened dramatically over time. The length of sentences shrank from 50 words per sentence in Pre-Elizabethan times to 23 words per sentence in his days, the end of the 19th century, from which he drew the conclusion that we tend to prefer shorter sentences. Obviously, sentence length cannot be treated as directly proportional to sentence difficulty, since some exceptionally long sentences are easy to follow, while certain short sentences appear to be impenetrably difficult to the reader. Despite this clear doubt about the crucial priority attached to sentence length as an essential determinant of readability, it is still an important factor affecting sentence readability. This is expressed in writing manuals, which suggest in general that the average sentence should not exceed 20 words. A century later than Sherman's (1893) investigation, Harrison and Bakker (1998) tested their hypothesis that long sentences effectively broken up do not increase the level of the reading difficulty of texts. Their research considered the length of packets, a mechanically modelled unit of a group of words between any punctuation marks (full-stop, comma, colon, semi-colon, exclamation mark, question mark, long dash and parenthesis). Their findings revealed that the sample sentences containing even over 55 words in length were acceptably readable as long as their packets were clearly and unmistakeably shown to the reader. This result obviously highlights the fact that it is not the sentence length that essentially determines the level of difficulty of a text, but the extent to which its complexity is revealed through punctuation marks.

Regarding the level of readability of a piece of writing, the present research does not fail to recognize the fact that besides sentence length and the extent to which the sentences of a corpus are organized into easily recognizable packets, there is a further factor which has a crucial effect on the perceived difficulty of a text: the complexity of its clausal units. The underlying notion for focusing on sentence complexity is the fact that processing a string of simple sentences poses less serious challenges to the reader than comprehending a stretch of complex and compound sentences (Rogers 1962; Larsen et al. 1978; Carpenter et al. 1995). The cognitive demands a text poses varies considerably according to the complexity of the syntactic structure of its constituent sentences of various lengths (Perfetti et al. 2005; Crossley et al. 2008). That is, the readability of a text depends highly on the complexity of its sentences, which can be revealed by the complexity of its syntactic structures (Huddleston 1984). Consequently, the current research aims at exploring the variety and complexity of the relationship of clausal units prevalent in the biology corpus by unveiling the level of complexity of the syntactic structures of the register.

3. Method

In order to make the study replicable and the results transferable, this section comprises two major parts: the process of the syntactic analysis of the corpus (Section 3.1) and the fine-grained description of the setting (Section 3.2). First, the method of compiling the biology corpus under investigation is explained (Section 3.1.1), followed by that of the reference corpus (Section 3.1.2). Next the explanation of the decision about the size of the corpus is provided (Section 3.1.3) and the method of selecting the particular syntactic structures and the syntactic taxonomy used in the present study is discussed (Section 3.1.4). In order for the data to be transferrable, a thorough description is provided about the environment where the corpora are applied. The characteristic features of the bilingual immersion programme of the secondary school and those of the participants (students and teachers alike) are described (Section 3.2.1). The last part of the section sheds light on the textbook containing the selected texts (Section 3.2.2).

3.1. The corpus

3.1.1. Compiling the corpus of the biology texts for secondary students (BIOCOR)

A register description can only be considered to be of high validity if the corpus is composed of texts which appropriately represent the bulk of the register. For this reason, in the process of compiling the texts under investigation careful attention was paid to the issue of the representativeness of the corpus.

For the compilation of the biology corpus (the BIOCOR) representative of what the 10th grade bilingual students are expected to read and process in their first academic term, it was first checked which biology texts are assigned to them. In a structured group interview with five high-achieving 10th graders in English, students were given their biology textbooks (Roberts 1981) and were asked to choose and write down the topics covered in the autumn term. High-achievers in English were chosen from the 10th graders to answer this single question as low-achievers tend to be more reluctant to share information about their studies; besides, low-achievers also have a tendency to fail to remember precisely what has been covered in class. Each of the five interviewees named the same eight chapters, which are listed in Table 1. To confirm the students' choices, the topics of the biology classes were followed in the electronic register of the school written by the biology teacher of the class from September to mid-January. By observing the electronic register, it was confirmed that the biology chapters list compiled by the students was exhaustive. Next, the eight chapters were typed in order to make them computer analysable, and a word count was run. The number of words of the biology corpus, containing the eight biology chapters studied in the first academic term in the 10th grade, amounts to 7,021.

Order of topics	Title of the chapter	Number of words in the chapter
1	The characteristics of living things	1,613
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	1,085
8	Parasitic worms	698

Table 1: The BIOCOR: the eight chapters of the biology textbook (Roberts 1981) and their length

In the present study, the notion of the register of biology textbooks in English for secondary students (or for short, the biology textbook register) refers to this corpus of biology texts, to the BIOCOR. In the present study, the notion of 'the register of biology textbooks in English for secondary students' (or the shorter phrase 'the biology textbook register') denotes this corpus of biology texts, that is, the BIOCOR. Yet, at some specific points of the enquiry, where the values of the probability coefficients are smaller than 0.5 and thus the data are statistically generalizable, the broader sense of the word 'register' is used. In these cases the term 'register' refers to the wider-ranging idea of biology textbooks in English for secondary students. These instances are explicitly indicated in the descriptions.

3.1.2. Compiling the reference corpus (REFCOR)

After finding the relevant biology texts, the next step was to choose the general English texts that can serve as the basis of comparison in the register analysis. One of the guiding principles in choosing the reference corpus (the REFCOR), against which the results of the corpus of the biology texts are compared, was that the pool of general English texts used in the 9th-grade classroom should also contain approximately 7,000 words in total. The texts were selected from the general English course book the 9th graders use in the last month before they take their end-term exam (Prodromou 1998). The other principle that determined the choice of the reference texts was that the general English texts should be representative of all the task types of the reading component of the First Certificate in English Cambridge Examination (FCE), which the 9th graders take when completing their general English studies. Although the data for the present research were gathered after 2008, the four parts of the reading paper represent a former FCE version, the one before the 2008 modifications. The reason for not choosing the most up-to-date version of the exam is that the 9th grade students tackle to solve the previous version as their end-term exam.[1] A complete FCE reading exam consists of about 2,000–2,500 words, so that it was clear that more than one exam had to be chosen to build the reference corpus. The last guiding principle in choosing the general English texts was that each part of the exam should be represented by an equal

[1] At the time of data collection, the mock FCE exams administered by the school were still structured according to the composition of the examination in practice before the 2008 modifications. The mock exams were not updated for practical reasons: the majority of the resources (practice books and test samples) available at the school were published before 2008.

number of texts and, as much as possible, an equal number of words. The total length of the twelve general English texts measures 7,098 words. Table 2 displays the reference corpus of the general English texts and their length, as well as the size of each part of the exam in a separate row.

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 2: The REFCOR: the general English texts chosen from the 9th graders' FCE course book (Prodromou 1998) and the length of the texts

3.1.3. The size of the corpus

The number of words in the collection of the biology texts (7,021) in the present research project is clearly far from one million, the approximate benchmark of a large corpus; thus, considering its size, it can be considered a mini-corpus (Biber and Conrad 2009). To comply and rely on a mini-corpus instead of a large one for the current analysis was a decision based on the numerous benefits a mini-corpus offers in the particular educational environment of the texts under investigation. The generally accepted notion that the bigger the size of a corpus, the more representative patterns can be revealed holds only true for describing general language use (Sinclair 1991). However, for the examination of a specific area of the language various scholars recommend the compilation of small corpora. A carefully targeted corpus that represents a particular register proves "to be a powerful tool for the investigation of special uses of language, where the linguist can 'drill down' into the data in immense detail" (O'Keffee and McCarthy 2010: 6). It goes without saying that a mini-corpus is also more manageable to handle than a large one (O'Keffee and McCarthy 2010). Moreover, compared to a large corpus, a mini-corpus is believed to display a higher rate of pedagogical usefulness (Ma 1993) and it is praised for yielding insights which can be used for specific learning purposes (Flowerdew 2002). Moreover, it can also be used for teaching non-native learners (Howarth 1998). From the students' point of view, it is easier to grasp and more 'learnable' than a large corpus (de Beaugrande 2001). Additionally, all occurrences, including low-frequency items, can be examined, which is not possible in the case of a large corpus (O'Keffee and McCarthy 2010). The examination of items entails the possibility of establishing a close link between the corpus and the context (Biber and Conrad 2009), since the language use is kept intact in the sense that the texts in a mini-corpus are not de-contextualized.

3.1.4. The process of selecting the investigated syntactic structures and the method of applying the syntactic taxonomy in the analysis

From a syntactic point of view, sentences can be categorized according to the number and kind of clauses in their syntactic structure. A simple sentence consists of one single clause, while compound and complex sentences comprise two or more clauses. The difference between a compound and a complex sentence lies in the dependence of its clauses. A compound sentence involves clauses which are all independent, i.e., clauses that could stand on their own as separate sentences. The independent clauses of a compound sentence are joined by any of the following seven conjunctions: *for*, *and*, *nor*, *but*, *or*, *yet*, *so* (Miltsakaki et al. 2004). Diagram 1 shows an example of a two-clause compound sentence taken from the biology corpus. The independent clauses of the sample are connected by the conjunction *and*.

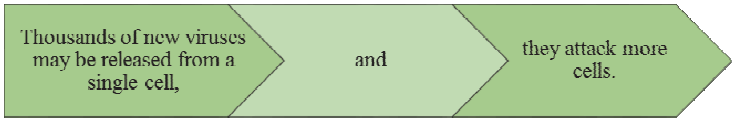

Diagram 1: A two-clause compound sentence from the BIOCOR

In contrast, a complex sentence contains at least one dependent clause, i.e. a clause that would not form a proper English sentence on its own. The dependent clause of a complex sentence is also called a subordinate clause, which is typically connected by one of the following subordinating conjunctions: *after*, *although*, *as*, *as if*, *because*, *before*, *even if*, *even though*, *if*, *if only*, *rather than*, *since*, *than*, *though*, *unless*, *until*, *when*, *where*, *whereas*, *whether*, *which*, *while* (Miltsakaki et al. 2004). Diagram 2 exemplifies a two-clause complex sentence from the biology corpus. The dependent subordinate clause of the sample sentence is connected by the subordinating conjunction *when*.

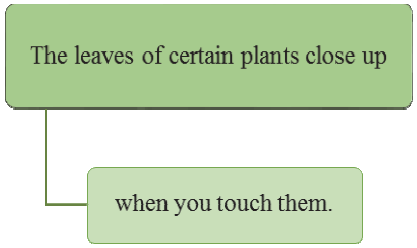

Diagram 2: A two-clause complex sentence from the BIOCOR

Finally, in the case of minimum three-clause-long sentences, the combination of compound and complex sentences is also possible. In a compound-complex sentence there are at least two independent clauses and a minimum of one dependent clause.

To gain data valuable for ESL and ESP teachers, the register analytical approach rather than the genre analytical method was adopted to investigate the corpus (for the underlying reasons, see Borza 2015). To describe the sentence complexity of the BIOCOR from the point of view of its characteristic syntactic structures, the frequency of ten types of sentence structures were tapped in the present research (see Table 3). These ten syntactic categories were set up and finalized as a result of having carried out a pilot study on two texts of the corpora. The piloted texts, each of approximately 500 words in length, were chosen from the books the bilingual students use in their studies: the FCE preparatory course book in the 9th grade and the biology textbook in the 10th grade, *First Certificate Star* by Prodromou (1998) and *Biology for Life* by Roberts (1981), respectively. In order to select texts for the pilot study from the above two sources, structured interviews were conducted with five low-achieving students in English in both grades. The aim of the interviews was to collect information on which text in particular students found exceedingly difficult to process during their studies. The question was articulated to low-achieving students in English with the assumption that the texts they find hard to process might abound in challenging linguistic features. Nearly unanimously, the students chose a newspaper article from the FCE course book (Unit 3). As regards the biology textbook, it was the chapter on viruses that all the low-achieving students in English found hardest to understand.

Code number	Syntactic structure	Number and type of clauses
1	simple sentence	one
2	compound sentence	two independent clauses
3	compound sentence	three independent clauses
4	complex sentence	one dependent clause
5	complex sentence	two dependent clauses
6	complex sentence	three dependent clauses
7	compound-complex sentence	two independent clauses and one dependent clause
8	compound-complex sentence	two independent clauses and two dependent clauses
9	compound-complex sentence	three independent clauses and one dependent clause
10	compound-complex sentence	three independent clauses and two dependent clauses

Table 3: The types of syntactic structures analysed in the corpora

The syntactic structures listed in Table 3 were chosen so that the possible combinations of syntactic categories covered all the variety of different sentence-types which appeared in the pilot texts. The researcher was fully open to extend the number and type of syntactic categories on the list in the process of analysing the pilot texts. As a result, the categories of the four-clause-long compound or complex sentences and that of the six-clause-long compound-complex sentence were added to the list. Further extension of the list was not necessary, since neither the BIOCOR nor the REFCOR contained any other syntactic structures different from the ones which were detected in the pilot study.

The syntactic analysis of the corpora was carried out manually according to the following steps. In order to tag the syntactic structures in the BIOCOR and the REFCOR, first clause boundaries were identified in both corpora. Next, the syntactic relationship of each clause was determined by conducting a clause type identification in accordance with the syntactic taxonomy described above. All the ten types of syntactic structures were tagged electronically in the texts by labelling them with the appropriate code number (see Table 3). Subsequently, the raw frequency of the code numbers was found by totalling their appearance in each individual text, and then adding them up in both registers separately. The relative frequency of each syntactic structure was counted against the basic unit of analysis, i.e. the sentence; this way the ratio of syntactic structures per number of sentences in the two registers was computed. To make the syntactic description of the BIOCOR meaningful, and thus increase the content validity of the analysis, the frequency of the various types of syntactic structures in the BIOCOR was compared with those in the REFCOR by means of computing t-tests. As the frequency ratios of the syntactic structures in the two registers are not influenced by each other at all, independent-sample t-tests were applied. In two cases, code numbers 9 and 10, only one of the two corpora contained

the given syntactic structure (Code 9 was absent in the BIOCOR, while Code 10 was not present in the REFCOR). Apparently, in these cases no statistical computation was possible, as t-testing tolerates no zero values. If either of the two registers contained a syntactic structure, its probability coefficient (Sig. 2-tailed) was tested in order to pinpoint if the difference in its frequency between the two registers was register-specific or sample-specific. The underlying reason for checking the probability coefficient was that frequency ratios with too high probability coefficients ($p>.05$) in the corpus do not show generalizable characteristics but sample-specific traits. For choosing the proper probability coefficient of a given frequency ratio of a syntactic structure, Levene's tests were also conducted. In those cases where Levene's tests revealed a significant difference ($p<.05$) equal variances were presumed, while the lack of significant difference when running Levene's test ($p>.05$) did not lead to presume equal variances. Examining the results of Levene's tests was a step in the analysis which guaranteed the interpretation of the results to be reliable in distinguishing register specific traits from sample specific ones. Conducting Levene's tests on the syntactic part of the analysis provided a statistical method of ensuring that register specific (generalizable) and sample specific results were differentiated, which increased the precision of transferability of the findings.

3.2. The setting

To increase the value of transferability of the data gained in the present study, the following section provides a thorough description both of the environment where the corpora are used (Section 3.2.1) and of the textbook from which the texts are drawn (Section 3.2.2).

3.2.1. The bilingual immersion programme of the secondary school and the participants

The bilingual education programme at the English-Hungarian bilingual secondary school was founded in 1987. The school was one of the 15 secondary schools countrywide that introduced a five-year-long bilingual programme in various languages, which meant an absolutely new type of education in Hungary at the time. Owing to its novelty, the Hungarian Ministry of Education launched the bilingual programme as an experimental one (Medgyes 2011). The implementation and the development of the English-Hungarian bilingual programme were ensured by a bilateral contract between the Department for Education of the United Kingdom and the Hungarian Ministry of Education. In practice, the British Council also gave its indispensable support to promote the progress of the English-Hungarian bilingual programme at school (Janni 2000). The aim of the bilingual education programme is to train future experts in various fields (economists, lawyers, doctors, engineers, IT specialists, etc.) whose linguistic abilities render them capable to study, to be engaged in research and to take professional responsibility in English. Besides, the programme also targets at educating students in a framework which cherishes cultural diversity and enhances understanding across cultural and linguistic differences (Bognár 2000). These aims are reached through the introduction of an educational programme that promotes content and language integrated learning (CLIL).

The secondary school was granted the name bilingual as curricular content is taught and learnt in two languages, at least five subjects in English and the others in Hungarian, and at least one of the teachers is a native English one. The language of instruction is English in the case of core academic subjects, such as mathematics, history, geography, physics and biology, while Hungarian is used as the medium of instruction in the case of Hungarian language and literature, IT, chemistry, music and physical education. Although at the time of the foundation of the programme, bilingual education meant the instruction of at least five academic subjects in the target language, a decade later the Ministry of Culture and National Education issued new principles of bilingual education, which reduced the compulsory number of subjects taught in the second language to three (Regulation no. 26/1997). Favourably, the new regulation did not affect the school adversely by reducing the number of classes delivered in English through curtailing the number of subjects taught in the target language. At present the subjects which are taught in English are those which were taught in English at the foundation of the programme. In line with the categorization of Swain and Johnson (1997), the bilingual education programme of the school can be best described as a bilingual immersion programme. Swain and Johnson (1997) claim that an immersion programme is multi-featured; it can be characterized by a bundle of traits (see Table 4). This educational model was given the name 'immersion' by Lambert and Tucker (1972), whose metaphor 'language bath' emphasized the intensive presence of the second language in the educational environment into which the students are immersed.

	Features characterizing immersion programmes (Swain and Johnson 1997)	Is the feature present in the bilingual education programme of the school?
1	use of the second language as a medium of instruction	Yes
2	a curriculum parallel to that used in the first language	Yes
3	overt support for the first language	Yes
4	additive bilingualism as programme aim	Yes
5	exposure to the second language being largely confined to the classroom	Yes
6	students entering the programme with similar, limited levels of second language proficiency	Yes
7	bilingually raised teachers	Not typical
8	the classroom culture being that of the local first language community	Yes

Table 4: Characteristic features of immersion programmes (Swain and Johnson 1997)

As can be seen in Table 4, nearly all the features of an immersion programme are present in the bilingual education programme of the school: (1) English as a second language is used as a medium of instruction in the majority of the academic subjects. (2) The curriculum runs parallel to that used in the first language in non-bilingual classes in all subjects. (3) The first language of the students is obviously overtly supported in the Hungarian language and literature classes, and at the same time, students are also provided with immediate first language aid in academic subjects taught in English if required, as all the subjects are taught by Hungarian teachers. (4) The educational programme aims to build additive bilingualism; by no means is the first language attempted to be suppressed or forced into the background, either linguistically or culturally. (5) Students in the bilingual programme use English as a second language mostly in the classroom, their exposure to English being confined to studying curricular content and conversing with the native teachers at times; although English is the language of instruction, it is not typically used outside the classroom. Both students and teachers tend to use their first language, Hungarian, in the breaks, during clubs, on class trips, at school assemblies or in any other extracurricular activities. (6) Students who enter the bilingual programme have a limited command of English; in the nearly three-decade-long history of the school no student raised in an English-Hungarian bilingual family has ever entered the bilingual programme. (7) The teaching staff of the school consists of monolingually raised Hungarians; some of the teachers are former students of the school, whom I consider as academic bilinguals (see below). The lack of bilingually raised teachers is the only trait where the school does not entirely meet the characterization of immersion programmes by Swain and Johnson (1997). (8) Finally, classroom culture is also similar to that of typical immersion programmes; in other words, it reflects that of the local Hungarian community.

Although immersion programmes represent an intensive form of bilingual education, it should be noted that it is the programme which is bilingual, not the students attending the school. The school does not offer academic language education for bilingual students, but rather bilingual academic education for students typically raised monolingually. The students entering the school are mostly monocultural Hungarians, whose parents communicate only in Hungarian at home. In their early age, they were not addressed regularly in two languages; thus the simultaneous acquisition of two languages, prerequisite for becoming bilingual in a classical, strict sense by the earliest scholars of bilingualism (Bloomfield 1933), does not take place in the micro-context of their homes. By the time they enter secondary school, they have mastered one single language, they do not have the native-like control of two languages, which is another characteristic feature of bilingualism in its severest sense (Bloomfield 1933). The macro-context in which the students were brought up is not different either. The English language is not significantly present in Hungarian society, it is not the language of wider communication (neither in administration nor in governance, it is not an official language in the country, nor the language of a minority). As a result, the students who embark on the programme have an acquired knowledge of Hungarian, but not of English. Language acquisition, as Krashen (1985) differentiated the two distinct types of mechanisms in language development, is a subconscious process that results in tacit knowledge of the language, while learning is a more conscious and laborious one. In their previous studies, most of the students learn English as a second language in the primary school for four to eight years. However, English is not a naturally acquired language for them, it is learnt to some extent after their first language has been acquired. The exceptions from monolingual Hungarian students are Vietnamese-Hungarian and Chinese-Hungarian bilinguals, whose number does not typically reach a handful in a year. Despite the name of the school, English-Hungarian bilingual secondary school, English-Hungarian bilinguals who were brought up in two languages in a bilingual speech community are not represented among the students at all. Distancing from the strictest sense of bilingualism proposed by Bloomfield in the early 1930s, the notion of bilingualism can be understood here in a much more allowing manner. On the other end of the spectrum of interpreting bilingualism, nearly everybody can be treated to be bilingual, at least anyone who knows “a few words in languages other than the maternal variety” (Edwards 2006: 7). To render the much-debated and polarized term bilingual as meaningful as possible in this particular educational context, I apply it in a dynamic sense. Baker and Jones (1998) suggest that bilingualism is a relative term, covering a spectrum of different degrees of bilingualism. In their wake, I endorse that the strength and the dominance of the first and second languages can change over time. Thus, individuals who were raised monolingually can become bilingual through constant exposition to a linguistic

environment different from their first language. When this process is induced through schooling, I use the term *academic bilingualism* to denominate the natural linguistic growth of distancing from monolingualism. In the environment under research, academic bilingualism signifies the process of Hungarian monolingual students gradually becoming bilingual through pursuing their studies in the English bilingual immersion programme. It should not go unnoticed, however, that academic bilingualism is an unbalanced form of bilingualism (in this sense radically different from early years bilingualism, either simultaneous or sequential), as equal competences in both languages are rare. Global language proficiency can be effectively described along two distinctively different dimensions, conversational and academic language use (Cummins 1999). In Cummins' terminology (1980), the former covers basic interpersonal communicative skills (BICS), such as accent, oral fluency and sociolinguistic competence, while the latter refers to cognitive and academic language proficiency (CALP), that is, to the intersection of language proficiency and cognitive and memory skills. The theoretical distinction between BICS and CALP was empirically supported by Biber's (1986) register analysis of a corpus containing one million running words. Training in the bilingual programme strengthens the second dimension, CALP, which is the major determinant of educational progress (Cummins 1999). Students educated in the bilingual programme perform at a native-like level in the CALP dimension of the second language. However, their BICS performance, particularly their sociolinguistic competence, lags somewhat behind. To conclude, when the term bilingual is used in the present research referring to the students of the immersion programme, it denotes academic bilingualism.

Since the completion of the bilingual immersion program requires the students to make continuous academic effort for five years, which might be more than demanding and strenuous for an average monolingual teenager, the school accepts highly performing students only. 8th graders are selected by the means of a rather competitive entrance exam. At the time of data collection, the entrance exam consisted of a national written exam testing students' skills of logical thinking in mathematics and their Hungarian vocabulary, linguistic flexibility along with reading comprehension and composition-writing skills. From 2013 onwards, after the data collection, a much debated school-based oral exam was introduced to check similar skills. Students have never been tested on their command of English; even complete beginners of English as a second language are accepted to the school.

In order to prepare monolingually raised Hungarian students for studying academic subjects in English, the school offers an intensive language course in the preparatory year, the so-called 'zero-year'. In other words, the five-year bilingual programme consists of a language preparatory year and four years of secondary studies leading to matriculation. The term 'zero-year' was officially in use until 1997, when the Ministry of Culture and National Education introduced new terminology in its principles of bilingual education (Regulation no. 26/1997). The regulations favoured numbering the academic years consecutively, thus the 'zero-year' became the 9th grade and the following first year of the national academic secondary school programme came to be known as the 10th grade. Consequently, students took their school-leaving exams in the 13th grade from 1997 on, which was previously taken in the 12th grade. Although the term 'zero-year' is not in official use at present, I use it synonymously with the intensive language preparatory 9th grade in my research, since it was widely applied at the time of data collection among the teachers and the students of the bilingual school alike. The intensive language course of the preparatory year comprises twenty hours of English a week, containing sixteen hours of general English classes and four English for specific purposes (ESP) classes. The 'zero-year' enables the students to continue their studies in English. In the following four years they pursue five core subjects in English, namely history, mathematics, physics, geography and biology. 9th graders are provided one history ESP, one mathematics ESP, one physics ESP and one geography ESP per week. Biology ESP is not part of the curriculum since the terminology of the subject is believed by the biology teachers working at the school to be far too diverse and difficult for 9th graders to grasp without learning the subject itself. Besides, an interview study conducted at the school (Cserép 1997) revealed that bilingual students find the language of biology most challenging among all the subjects taught in English. With regard to teaching of the English language, the aim of the preparatory year is to enhance the students' knowledge to reach a firm B2 level. In accordance with the Common European Framework of Reference for Languages, students passing the preparatory year are expected to "understand the main ideas of complex text on both concrete and abstract topics, including technical discussion in their field of specialization," "produce clear, detailed text on a wide range of subjects and explain," as well as "explain a viewpoint on a topical issue giving the advantages and disadvantages of various options" (CEFR 1996: 24). To ensure that students in the 'zero-year' develop these language skills to the appropriate level, only those students are allowed to continue their studies in the 10th grade who prove to be successful at passing an upper intermediate level mock Cambridge Exam, the First Certificate in English (FCE), administered by the school.

Although at this point 10th grade students generally find almost all subjects difficult to follow in English and complain about the level of difficulty of most of the textbooks in English, biology was chosen to be investigated in the present research as the status of this subject differs deeply from that of the other subjects taught at school: there is no biology ESP instruction provided for the students in the 9th grade. This means that students attending biology classes delivered in English in the 10th grade rely on the knowledge they gained in their *general* English studies and in the *other* four specialized English classes (history, mathematics, physics and geography ESP).

The seventy-two students who enter the bilingual immersion programme every year are divided into two classes, which are further split into three groups in the language preparatory 'zero-year'. The six groups of twelve students are

formed according to their level of proficiency in English. The groups are either mixed-level ones, containing complete beginners, false beginners and pre-intermediate students, or homogenous groups, where complete beginners are instructed separately from students with higher levels of English. The decisive factor whether to arrange students in either mixed or homogenous groups is the number of complete beginners in the year. If their number does not reach half a dozen, students with different levels of English are mixed, while larger sets of complete beginners tend to be grouped homogenously, as long as they attend the same class. At the time of data collection, students were grouped homogenously. Groups are headed by a group leader, an English teacher responsible for promoting and checking the linguistic development of each student in the group. This is attained by teaching a relatively high number of classes in the group: the group leader delivers a minimum of six classes a week in her group. The other general English classes are taught by non-native English teachers and one native English teacher, while the English for specific purposes (ESP) classes are given by non-native subject teachers.

3.2.2. The situational characteristics of the register of the biology textbook

Texts can be described from infinitely different viewpoints. In order to allow the comparison of the various research results in the field of text analysis, Biber and Conrad (2009) suggest a general framework. The advantage of their framework is that it can be employed in any analysis for describing the texts situational characteristics, that is, in what context and under what circumstances the texts are used and for what specific purposes. The comprehensive nature of the framework is due to the fact that it was developed as a compilation of previous theoretical models that describe registers. Table 5 shows the seven major situational characteristics considered in the framework and a brief description of the academic prose of the biology textbook under scrutiny here (Roberts 1981) along the given parameters.

Situational parameter	The biology textbook (Roberts 1981)
Participants	Addressor: single Addressees: un-enumerated
Relations among participants	Lack of interactiveness Inequality in social power Lack of personal relationship Shared knowledge is specialist
Channel	Writing Medium: printed
Production circumstances	Planned, revised, edited Controlled
Setting	Time and place of communication is not shared Public Contemporary
Communicative Purposes	Inform, explain, educate Factual information Certainty in epistemic stance
Topic	Education Biology

Table 5: The situational parameters of the biology textbook (Roberts 1981)
according to the framework of situational characteristics of a text (Biber and Conrad 2009)

The biology textbook was produced by a readily identifiable single author, Michael Roberts, who is the sole addressor of the texts. The intended readers, the addressees of the textbook, are 14-16 year old secondary school students preparing for their General Certificate of Secondary Education (GCSE) exam in biology and who study biology in English. The addressee is a group of individuals whose exact number cannot be specified, thus the audience of students forms a set of un-enumerated addressees.

The relationship among the participants does not bear interactive features. The addressees and the addressor are not directly involved with each other; the author is not easily accessible to address a response to. Addressees of the biology textbook tend to address their questions to their biology teacher, who is readily available for them in person, while the addressor of the textbook requires effort from the readers if they intend to contact him in writing. The participants do not share equal social roles: the addressor possesses considerably higher social power and more authority than the addressees. The relationship of the participants cannot be characterized as being personal, not so much because of the inequality of social power, but due to the complete lack of bidirectional encounters among the participants (e.g. meeting

in person, exchanging correspondence). The shared background knowledge of the participants covers a specialist field, since the addressor communicates information only in the field of biology. The addressees are not expected to have expert background in the field, their novice status being connected to their lower social power.

The physical channel of the register is writing, its specific medium of communication being the printed form. Although the textbook is also available in electronic form, the students at the bilingual school use its printed version, which is considered to be a permanent form by Biber and Conrad (2009).

The written mode of the texts immediately affects its production circumstances. The addressor carefully plans and revises the texts, and the level of unintendedness is extremely low, if any. The editor of the text is the single addressor himself. However, instances of revision of the original text are not evident for the readers, who are exposed to the final, published version only. From the point of view of the addressees, whose production involves comprehension of the text, the circumstances are also completely controlled. The addressees have a chance to determine their individual speed of reading according to their engagement in the comprehension process, and the sequencing of the bits of the text read is also their own choice. Communication is not produced in real-time by any of the participants.

The absolute lack of shared time and place of communication describes the setting of the biology textbook. The participants fail to share a physical context, unless one of the addresses strives to exchange information with the addressor, which has never happened in the history of the bilingual secondary school. The biology textbook offers a public way of communication, which occurs at present, so that its physical context is contemporary.

The communicative purpose of the biology textbook is manifold. The addressor intends to convey information about already established knowledge in the field of biology. With a metaphoric picture, Shapiro (2012: 100) underlines this function of science textbooks in general as forming "the papery strata between whose leaves the fossil traces of scientific practices are preserved". In less poetic terms, the biology textbook aims at training uninitiated learners, which involves disseminating established knowledge that has already been accepted by experts in the field. In other words, the textbook is not designed to impart newly tested hypotheses, but rather focuses on maintaining knowledge that has been widely accepted. Among other communicative purposes, the explanatory function of the biology textbook is essential; carefully chosen concepts are clarified in its chapters. Additionally, information is interpreted, practical investigations are displayed, and several states and processes are also described. The reason for writing a biology textbook is to convey factual information to the addressees. As a result, the epistemic stance of the biology textbook expresses a high rate of certainty, the information it imparts leaving little space for doubt. The claims in the textbook are generalizable, and the statements are verifiable.

The topic area of the biology textbook is education in general, while its specific topical domain is biology. Strictly focused informational purposes define its subfield, which covers the various topics GCSE students are tested in biology.

After the description of the register of the biology textbook in general, let us now turn our attention to the specific syntactic analysis of the biology corpus in particular.

4. Results and discussion

By exploring the sheer number of clauses in the sentences of the BIOCOR (see Diagram 3), it can be claimed that the majority of the sentences of the corpus (54%) are one-clause-long simple sentences. Another important part of the BIOCOR, one-third of the sentences of the corpus (33%), corresponds to two-clause-long sentences. From a syntactic point of view, the two shortest types of structures (one-clause long and two-clause long sentences) represent 87% of the whole corpus; that is, approximately nine out of ten sentences of the BIOCOR. The frequency of longer sentences, those which contain three clauses, drops drastically: only one every tenth sentence tends to be longer than two clauses. Longer sentences than these, those containing four clauses (3%) or five clauses (1 single instance, which amounts to less than 1%) are only sporadically found in the corpus. Therefore, at this point it can hardly be concluded that the syntactic structure of the BIOCOR might be challenging for the target readers.

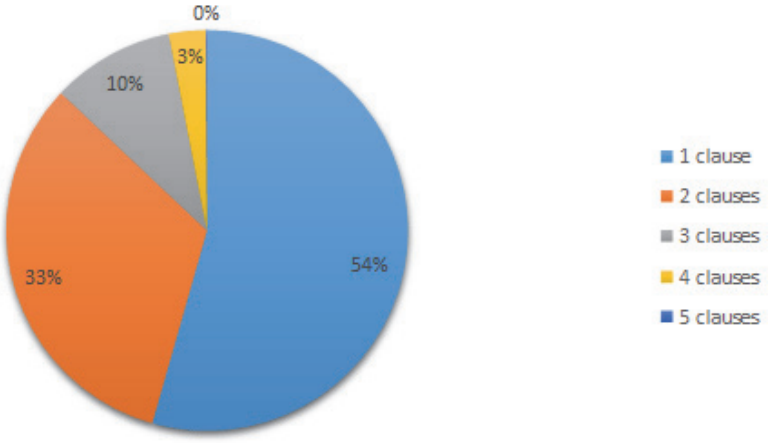

Diagram 3: The frequency of sentences with a different number of clauses in the BIOCOR

As a baseline of comparison (see Diagram 4), the REFCOR applies one-clause-long simple sentences to a smaller extent (39%) than the BIOCOR (54%). Two-clause-long sentences, however, are just as massively present in the REFCOR (32%) as in the BIOCOR (33%). Nevertheless, the sum of the two shortest types of syntactic structures (one-clause long and two-clause long sentences) reveals a conspicuously noteworthy difference between the two registers. While 87% of the BIOCOR is constructed of one- or two-clause-long sentences, the REFCOR relies on the simplest syntactic categories much more sparingly. One- or two-clause-long sentences in the REFCOR represent about two-thirds of the sentences in the corpus (71%). On the other hand, three-clause-long sentences are twice as heavily present in the REFCOR (20%) as in the BIOCOR (10%). These figures imply that processing the BIOCOR might be described as half as strenuous as that of the REFCOR from a syntactic point of view, since longer, syntactically more challenging sentences appear twice less recurrently. What is more, the frequency of even longer sentences discloses an even greater difference between the two registers. Four-clause-long sentences appear three times more often in the REFCOR (9%) than in the BIOCOR (3%). Similarly to the previous results, these figures also indicate that the level of difficulty of the BIOCOR is considerably lower than that of the REFCOR with regard to syntactic complexity. Finally, five-clause-long sentences are completely absent from the REFCOR. In view of the characteristic syntactic traits of the two corpora, more precisely, the number of clauses they include, the BIOCOR appears to be noticeably more easily readable than the REFCOR: its syntactic structures, compared to the REFCOR, display no challenging qualities at all.

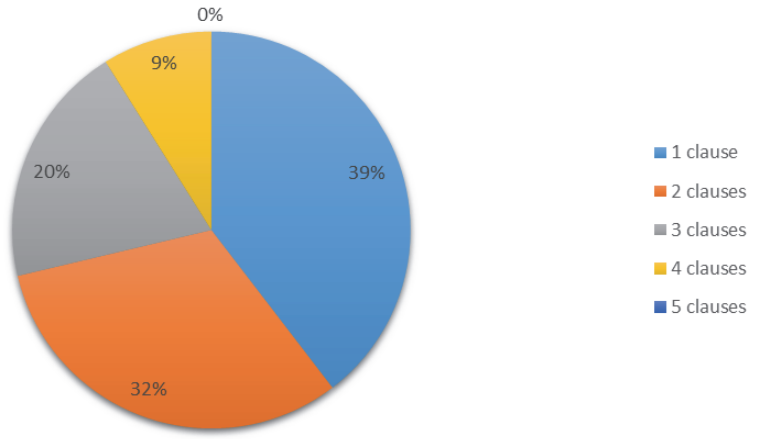

Diagram 4: The frequency of sentences with a different number of clauses in the REFCOR

For a better understanding of the syntactic nature of the register, Diagram 5 provides a more in-depth analysis considering the frequency of the different types of syntactic structures in the BIOCOR (for decoding the ten code numbers in the line graph, see Table 3).

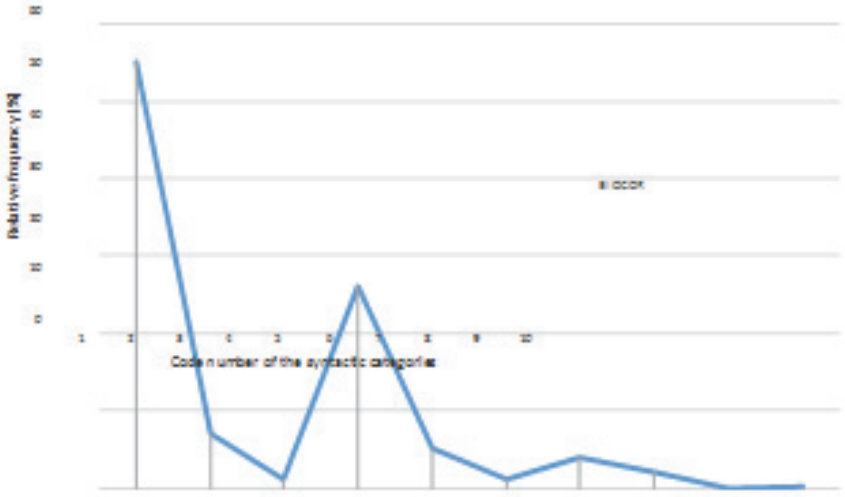

Diagram 5: The frequency of the ten types of syntactic structures in the BIOCOR

As already shown in Diagram 1 above, the BIOCOR makes extensive use of simple sentences (Code 1; 54%). The frequency of compound sentences with two independent clauses (Code 2) is dramatically lower (barely 7%), while compounding three independent clauses (Code 3) amounts merely to 1%. The presence of complex sentences with one single dependent clause (Code 4) is nearly half as numerous (26%) as that of simple sentences (54%). Longer complex sentences with two dependent clauses (Code 5), however, appear five times fewer (5%) than the shortest type of complex sentences (Code 4, 26%). The frequency of complex sentences with three dependent clauses (Code 6) is extremely low (1%) in the BIOCOR: such long complex sentences are not typically used in the corpus. Three-clause-long sentences of different syntactic types show a similarly small rate of appearance in the corpus. Both Code 5 (complex sentences with two dependent clauses) and Code 7 (compound-complex sentences with two independent clauses and one dependent clause) appear to a rather limited extent in the corpus, 5% and 4% respectively. Four-clause-long compound-complex sentences comprising two independent clauses and two dependent clauses (Code 8) are used half as rarely, their rate amounting to 2%, while four-clause-long compound-complex sentences with three independent clauses and one dependent clause (Code 9) are not present in the BIOCOR at all. The instance of the unusually complex five-clause-long compound-complex sentence (three independent clauses and two dependent clauses, Code 10) is a singular example in the BIOCOR

The distribution of various syntactic structures in the REFCOR displays a different pattern than that of the BIOCOR (see Diagram 6). Simple sentences (Code 1) are less extensively used in the REFCOR (40%) than in the BIOCOR (54%). By contrast, the frequency of compound sentences containing two independent clauses (Code 2) is higher in the REFCOR (10%) than in the BIOCOR (7%). Compound sentences with three independent clauses (Code 3) are not too frequent in the REFCOR (2%), but they appear twice as often in this corpus than in the BIOCOR. Among the two-clause-long sentences in the REFCOR, complex sentences with one dependent clause (Code 4) are twice as abundant (22%) as compound sentences with two clauses (10%). Three-clause-long complex sentences, those with two independent clauses (Code 5), are half as frequently present in the REFCOR (12%) as two-clause-long complex sentences; however, their appearance is more than double in this corpus than in the BIOCOR (5%). Longer complex sentences, those with three independent clauses (Code 6), are insignificantly used in the REFCOR (3%); nonetheless, the presence of this syntactic structure is three times less dominant in the BIOCOR (1%). Compound-complex sentences with two independent clauses and one dependent clause (Code 7) or with two independent clauses and two dependent clauses (Code 8) appear with a similarly modest frequency in the REFCOR (6% and 5%, respectively). The presence of four-clause-long compound-complex sentences, those with three independent clauses and one dependent clause (Code 9), is negligibly small (1%) in the REFCOR. Longer compound-complex sentences, those with three independent clauses and two dependent clauses (Code 10), are completely absent from the REFCOR.

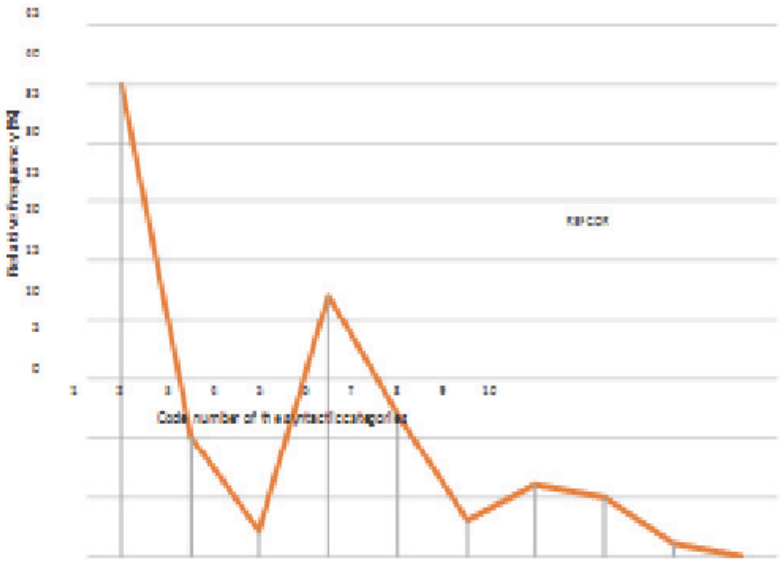

Diagram 6: The frequency of the ten types of syntactic structures in the REFCOR

Diagram 7 displays the comparison of the syntactic structures in the BIOCOR and in the REFCOR with the probability coefficient (p) of each type of syntactic structure, denoted by the structure code number. On the basis of the values of the probability coefficients, the difference between the two corpora is register-specific at two points, where the p values are smaller than 5 per cent ($p<.05$). These are the peak-points of the line graphs, at Code 1 ($p=.001$) and at Code 4 ($p=.0001$). At all the other points of the graphs the differences reveal corpus-specific dissimilarities, which cannot be generalized as register differentiating variations. The comparative line graphs disclose that one-clause-long simple sentences are more profusely used in the BIOCOR than in the REFCOR. This significant difference renders the implication evident that the BIOCOR is more accessible to process than the REFCOR. Two-clause-long compound sentences are outstandingly more abundant in the REFCOR than in the BIOCOR. This syntactic trait also implies that the BIOCOR is more straightforward to process than the REFCOR. Three-clause long compound sentences are not typical in either of the two corpora; nevertheless, their presence in the REFCOR is twice as high as in the BIOCOR. Again, this result reveals the syntactic simplicity of the BIOCOR compared to that of the REFCOR. The frequency of complex sentences containing one dependent clause is significantly higher in the BIOCOR than in the REFCOR, which may suggest that the BIOCOR requires more effort to be accessible on the part of its readers. However, the presence of complex sentences containing two dependent clauses is more than double in the REFCOR than in the BIOCOR. Furthermore, complex sentences with three dependent clauses are used three times more commonly in the REFCOR than in the BIOCOR. Considering all types of complex sentences (containing one, two or three dependent clauses), it is undoubtedly clear that the REFCOR is far more varied and poses more serious syntactic challenges than the BIOCOR.

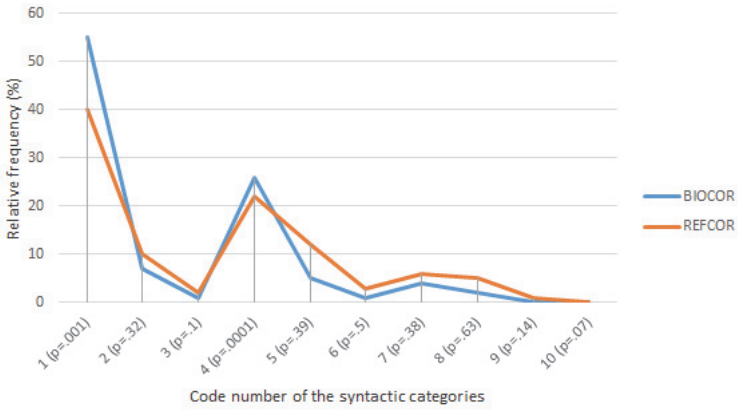

Diagram 7: The frequency of the ten types of syntactic structures in the BIOCOR and in the REFCOR

For this reason, 9th grade bilingual students trained on the REFCOR should hardly find the relatively modest use of complex sentences in the BIOCOR demanding to process. The extremely low presence of complex-compound sentences in the BIOCOR also makes the corpus more uncomplicated for its readers than the REFCOR. In conclusion,

the results of the examination of all the syntactic structures used in the BIOCOR do not explain fully the perceived difficulties of the 10th grade bilingual student when processing the relevant chapters of the biology textbook. These findings are in line with Shapiro's (2012) idea that the register of pre-college science textbooks written for non-experts is more popularizing than academic. Consequently, the language of secondary textbooks is less technical than the one used in the discourse community of scientists. His main argument for categorizing pre-college science textbooks under popularizing literature suggests that secondary students form an audience of young non-scientists (most of whom are not inclined to become scientists) to whom science is presented in a less or even non-technical language by authors who try to convey the product of their academic profession. The syntactic simplicity of the biology textbook register can be regarded as one of the linguistic features revealing the non-academic but popularizing nature of the register.

5. CONCLUSION

The present study investigated the level of complexity of the syntactic structure of the secondary school biology textbook register. The syntactic complexity of the BIOCOR compared to that of the REFCOR is summarized in Tables 6 and 7.

Number of clauses	BIOCOR	REFCOR	Does it render the BIOCOR more complex than the REFCOR?
1	54%	39%	No
2	33%	32%	No
3	10%	20%	No
4	3%	9%	No
5	<1%	0%	No

Table 6: The frequency of sentences with a different number of clauses in the BIOCOR and the REFCOR

Code number	BIOCOR	REFCOR	Does it render the BIOCOR more complex than the REFCOR?
1	54%	39%	No
2	7%	10%	No
3	1%	2%	No
4	26%	22%	No
5	5%	12%	No
6	1%	3%	No
7	4%	6%	No
8	2%	5%	No
9	0%	1%	No
10	<1%	0%	No

Table 7: The frequency of the ten types of syntactic structures in the BIOCOR and the REFCOR

As the data in Table 6 clearly reveal, the sheer number of clauses in the BIOCOR renders the biology textbook syntactically simpler than the REFCOR. None of the possible lengths of sentences (expressed in the number of clauses) are more challenging in the BIOCOR than in the REFCOR. On the contrary, the BIOCOR can be described as containing sentences with a high level of syntactic simplicity.

The frequency of the ten types of syntactic structures in the BIOCOR reinforces the syntactically straightforward nature of the register (see Table 7). All the categories in the taxonomy shed light on the fact that the BIOCOR, compared to the REFCOR, comprises sentences without syntactic challenges. The results of the investigation disclose that syntactic simplicity is prevalent in the BIOCOR: simple sentences abound, complex sentences are used in a modest manner, while complex-compound sentences are hardly present in the corpus.

In view of its low level of syntactic complexity, the BIOCOR obviously displays a lower degree of readability than the REFCOR. Processing the BIOCOR is expected to require a smaller degree of cognitive demands and to pose less severe problems to the reader than the REFCOR. The syntactic simplicity of the BIOCOR may stem from the fact that the register is written for non-experts, its language being therefore more popularizing than academic. These results are in agreement with Shapiro's (2012) findings about the language of pre-college textbooks. Yet, the constrained variety and complexity of syntactic structures in the BIOCOR fail to give satisfactory explanations for the difficulties 10th grade bilingual students face when they process the biology textbook. On the basis of the results of the present study, however, pedagogical implications can be drawn with regard to the choice of reading texts in the 9th grade general English classes. Since the 10th grade biology texts fail to reach the CEFR B2 level in terms of syntactic complexity, the 9th grade reading tasks are advised to be chosen from sources linguistically less demanding than the Cambridge First Certificate Examination.

REFERENCES

Baker, Colin and Sylvia Prys Jones. 1998. *Encyclopedia of bilingualism and bilingual education*. Clevedon: Multilingual Matters.

Beaugrande, Robert de. 2001. Large corpora, small corpora, and the learning of language. In Mohsen Ghadessy ed. *Small corpus studies and ELT. Theory and practice*. Philadelphia: John Benjamins, 3–28.

Biber, Douglas. 1986. Spoken and written textual dimensions in English: resolving the contradictory findings. *Language* 62: 384–414.

Biber, Douglas and Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Bloomfield, Leonard. 1933. *Language*. New York: Holt.

Bognár, Anikó. 2000. A két tanítási nyelvű oktatás 12 „nem tucat" éve a magyar közoktatásban. *Modern Nyelvoktatás* 6/1: 56–63.

Borza, Natalia. 2015. Analysing ESP texts, but how? *Practice and Theory in Systems of Education* 10/1: 1–15.

Carpenter, Patricia A., Miyake Akira and Marcel Adam Just. 1995. Language comprehension: sentence and discourse processing. *Annual Review of Psychology* 46: 91–120.

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.

Crossley, Scott, Jerry Greenfield and Danielle McNamara. 2008. Assessing text readability using cognitively based indices. *TESOL Quarterly* 42/3: 475–493.

Cserép, Szilvia. 1997. Technical terms in biology. An investigation into scientific English. MA thesis. University of Economic Sciences, Budapest.

Cummins, Jim. 1980. The cross-lingual dimensions of language proficiency: implications for bilingual education and the optimal age issue. *TESOL Quarterly* 14/2: 175–188.

Cummins, Jim. 1999. *BICS and CALP: clarifying the distinction*. Washington, DC: U.S. Department of Education. Office of Educational Research and Improvement, Educational Resources Information Centre (ERIC).

Edwards, John. 2006. Foundations of bilingualism. In Tej K. Bhatia and William C. Ritchie eds. *The handbook of bilingualism*. Oxford: Blackwell, 7–31.

Flowerdew, John. ed. 2002. *Academic discourse*. Harlow: Longman.

Harrison, Sandra and Paul Bakker. 1998. Two new readability predicators for the professional writer. *Journal of Research in Reading* 21/2: 121–138.

Howarth, Peter. 1998. Phraseology and second language proficiency. *Applied Linguistics* 19: 24–44.

Huddleston, Rodney. 1984. *Introduction to the grammar of English*. Cambridge: Cambridge University Press.

Janni, Gabriella. 2000. Mi szakmai elitet képzünk. *Educatio* 4: 799–810.

Krashen, Stephen. 1985. *The input hypothesis: issues and implications*. London: Longman.

Lambert, Wallace E. and G. Richard Tucker. 1972. *Bilingual education of children: the St. Lambert experiment*. Rowley, MA: Newbury House.

Larsen, Stephen C., Randall M. Parker and Barbara Trenholme. 1978. The effects of syntactic complexity upon arithmetic performance. *Learning Disability Quarterly* 1/4: 80–85.

Ma, Bruce Ka Cheung. 1993. Small-corpora concordancing in ESL teaching and learning. *Hong Kong Papers in Linguistics and Language Teaching* 16: 11–30.

Medgyes, Péter. 2011. *Aranykor. Nyelvoktatásunk két évtizede. 1989-2009*. Budapest: Nemzeti Tankönyvkiadó.

Miltsakaki, Eleni, Rashami Prasad, Aravind Joshi and Bonnie Webber. 2004. The Penn Discourse Treebank. In *Proceedings of the Language Resources Evaluation Conference (LREC)*. Lisbon: Universidade Nova de Lisboa. < http://www.lrec-conf.org/proceedings/lrec2004> (pdf doc. no. 618)

O'Keffee, Anna and Michael McCarthy. 2010. *The Routledge handbook of corpus linguistics*. London: Routledge.

Perfetti, Charles A., Nicole Landi and Jane Oakhill. 2005. The acquisition of reading comprehension skill. In Margaret J. Snowling and Charles Hulme eds. *The science of reading*. Oxford: Blackwell, 227–247.

Prodromou, Luke. 1998. *First Certificate star*. Oxford: Macmillan.

Roberts, Michael. 1981. *Biology for life*. Surrey: Thomas Nelson and Sons.

Rogers, John R. 1962. A formula for predicting the comprehension level of material to be presented orally. *The Journal of Educational Research* 56/4: 218–220.

Shapiro, Adam. 2012. Between training and popularization. Regulating science textbooks in secondary education. *Isis* 103/1: 99–110.

Sherman, Lucius Adelno. 1893. *Analytics of literature: a manual for the objective study of English prose and poetry*. Boston: Ginn & Co.

Sinclair, John. 1991. *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Swain, Merrill and Robert Keith Johnson. 1997. Immersion education: a category within bilingual education. In Robert Keith Johnson and Merrill Swain eds. *Immersion education: international perspectives*. Cambridge: Cambridge University Press, 11–16.

Corresponding author
Mikszáth tér 1.
Budapest 1088
Hungary
e-mail: nataliaborza@gmail.com

received: August 2016
accepted: November 2016