REGISTER ANALYSIS OF ENGLISH BIOLOGY TEXTS: A CORPUS-BASED EXPLORATORY STUDY OF GRAMMAR

Natália Borza
Language Pedagogy PhD Program, Eötvös Loránd University, Budapest
nataliaborza@gmail.com

Abstract: While considerable research has been conducted on describing the discourse features of academic writing in the field of biology, there is hardly any research available on the register specific features of textbooks written for secondary students, in particular on those of biology textbooks. The present quantitative investigation aims to fill this niche by means of developing an analytical instrument capable of providing data that describe the dominant grammatical features of English language biology texts used in the instruction of Hungarian students in a bilingual secondary school in Budapest. The results of the study reveal that biology texts for secondary students are less complex in terms of grammar than general English reading texts for intermediate (B2) learners. The findings are hoped to be of assistance to biology ESP teachers and general English teachers instructing in a bilingual secondary school alike.

Keywords: register analysis, biology textbook, biology ESP, bilingual secondary school

1 Introduction

The aim of the study is to test the applicability of an analytical tool designed for the comparative examination of grammar structures prevalent in biology texts and those in reading tasks of general English course books. The tool conceived in a pedagogical perspective is unique as no such tool has been devised so far that analyses the genre of biology textbooks from the point of view of EFL teaching, despite the fact that it is useful for both language pedagogical and theoretical reasons. By yielding data through applying the analytical tool to a carefully targeted corpus in order to identify the differences between biology and general English texts, pedagogical implications can be drawn as to ESP materials design, more precisely to the possible grammar foci as well as the sequencing of these in teaching of biology ESP. At the same time, with the help of the analytical tool, the genre of biology textbooks can also be characterized with regard to high-frequency grammatical features that distinguish them from general English texts.

2 Rationale and the research questions

This pedagogically and theoretically motivated empirical research is the result of having observed a discrepancy students face at a bilingual secondary school in Budapest, Hungary. Students at the end of their first year at the school take the intermediate level Cambridge examination, the First Certificate in English (FCE), which is level B2 in the Common European Framework of Reference for Languages. 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 as a foreign language, students have considerable difficulties. Although at this point they generally find most subjects difficult to follow in English, biology was chosen to be investigated in particular as its status differs from that of the other subjects in the school. During the language preparatory year, the so-called 'zero year', even complete beginner students have the chance to master English as a foreign language in no less than twenty hours a week. This highly intensive language course contains sixteen hours of general English classes 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 because the biology teachers working at the school think the special terminology of biology is far too diverse and difficult to grasp for 9th graders. 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 specialized English classes. Accordingly, as a teacher of general English in the 9th grade, I have become interested in what my students need to know in order for them to handle biology texts successfully in the 10th grade. The units and modules of all the three books used in the language preparatory year of the program (Falla & Davies, 2008, Cunningham & Moor, 2005, Prodromou, 1998) are built around grammar points; consequently, 9th graders are given a thorough training in grammar and are expected to master grammar proficiently at B2 level.

By applying the analytical tool to the texts the students read, I hoped to gain a deeper understanding of the students' needs in terms of grammar and thus to support my own and my colleagues' professional development as a general English teacher. Besides, this corpus-based exploratory study is intended to provide insights for biology ESP teachers, once biology ESP has been included in the zero year language programme of the secondary bilingual school.

With the above pedagogical and theoretical aims in mind, the study attempts to answer the following research questions:
1: Is the analytical tool designed for the comparative examination of grammar structures prevalent in biology texts and in reading tasks of general English course books used in the instruction of Hungarian students in a bilingual secondary school capable of providing reliable and valid data?
2: In contrast with general English, what distinctive features characterise biology texts in particular used by 10th grade students at an English-Hungarian bilingual secondary school in their first academic term with regard to grammar?

3 Theoretical background

3.1 Genres, registers and their study

Biology textbooks and general English course books are written for different audiences with different purposes. As a result, their language use and structure vary considerably and as such they can be treated as being two distinct genres. According to the Swalesian definition of *genre*, where "the principal critical feature that turns a collection of communicative events into a genre is some shared set of communicative events" (Swales, 1990, p. 46), the two types of texts under investigation are clearly distinguishable. In other words, they can be regarded as belonging to two distinct genres. In a somewhat similar manner, the concept *register* as defined by Biber et al. (1998), which is a "cover term for varieties defined by their situational characteristics" considering the "purpose, topic, setting, interactiveness, mode, etc." of the situation (1998, p. 135), can also be applied to differentiate between the two types of texts. That is to say, they are different registers in the Biberian sense in that their “identifying markers of language structure and language use differ from the language of other communicative situations” (Biber & Finegan, 1994, p. 20). In general terms, discourse analysts working in the field of ESP uncover “specialized registers in English” (Biber, 1998, p. 157). ESP and other registers can be analytically studied as their clusters of “associated features have a greater than random tendency to co-occur” (Halliday, 1988, p. 162).

The typical characteristics of a register can be described by a *comprehensive register study*. A comprehensive linguistic analysis of a register comprises three essential features (Biber, 1998, p. 136). First, the study should be based on a large number of texts investigated instead of a small corpus in order for the description to be accurate. Next, wide-ranging linguistic features should be examined as it is not typical for a register to be identified and well-described by the presence of one single distinctive linguistic feature. On the contrary, sets of several linguistic features tend to belong to different registers; that is, registers can be distinguished by their relative use. Finally, there should be a comparison across registers since average frequencies without comparison do not mean much. To carry out a reliable comprehensive register study, the corpus-based approach provides a suitable framework. The vital elements of a corpus-based analysis are as follows:

- it is empirical, analyzing the actual patterns of use in natural texts;
- it utilizes a large and principled collection of natural texts, known as a “corpus”, as the basis of analysis;
- it makes extensive use of computers for analysis, using both automatic and interactive techniques;
- it depends on both quantitative and qualitative analytical techniques (Biber et al., 1998, p. 4).

Accordingly, the present study examines a representatively selected collection of biology texts through a large number of grammar features in order to describe the specialized language use of the register by comparing it to a reference corpus of general English texts. The statistical analysis of the wide-ranging linguistic features reveals the typical grammatical characteristics of the register.

3.2 Earlier research in the area

There has been extensive research in the field of register analysis; numerous genres have been described within various frameworks. However, still not enough attention has been given to the genre of textbooks. A multidimensional analysis was carried out by Biber (1991) to describe particular linguistic features of primary school textbooks, but the register of secondary school textbooks has not been fully discovered. Keeping the target readers’ linguistic challenges in the focus of their studies, a handful of discourse analytical research has been done on grammar points relevant to language learners, for instance the different use of modals (Coates, 1983; Tottie, 1985) and participle clauses (Thompson, 1983) in written discourse have already been examined. Concerning 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. Taking a rhetorical point of view on university level biology texts, the difference in discourse units in biology research articles (Biber & Jones, 2005) and the variations among moves within biochemistry research articles (Kanoksilapatham, 2005) have also been investigated. 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 the point of view of grammar use. The aim of the present research is to discover the conventions of the register of biology texts for secondary students through developing a framework of analysis for the study of grammar from the point of view EFL teaching. This grammar-focused register study is based on a corpus consisting of more than 14,000 words and it comprises a large set of linguistic features comparing the register of biology textbooks to general English texts designed for intermediate (B2) level learners of English.

4 Methods

4.1 The analytical tool: procedures of design

In the process of developing the analytical tool for the study of distinctive grammar features of the genre of the biology textbook used in the 10th grade of the bilingual school, four steps were taken.

First, the literature was studied in order to see what kind of frameworks for linguistic features have been investigated in ESP discourse analysis. In order to give a comprehensive description of ESP registers, Biber (1998) investigated various linguistic items. The linguistic features he analysed range from word length through word level linguistic phenomena (such as various pronouns, several aspects of modals, etc.) to sentence structure (for instance different participle clauses, relative clauses, subordination). For the exhaustive list of linguistic items Biber (1998) used for analysis see Appendix A.

Having consulted the literature, I collected the grammar topics covered within the 34 grammar units in the FCE grammar book (Vince, 2003) used in the 9th grade in order to extend the Biberian framework with grammar items specifically relevant for EFL learners at the intermediate level. Relying on my own professional experience gained through teaching for six years at the bilingual secondary school, I chose those features of the grammar book that typically pose challenges to 9th graders by the end of the academic year. The analytical tool at this stage contained nine aspects of comparison with 74 items.

In the third stage, the grammar features of the tool developed so far were piloted through the analysis of two texts, each about 500 words in length. The texts analysed were chosen from the books the bilingual students use in their studies, that is, 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. The general English text served as a reference corpus, or the basis of comparison, providing baseline data against which the biology text can be compared and contrasted. Two registers were chosen as Biber (1998) warns that register analysis should never be done on one single register. Biber argues that average frequencies are not meaningful on their own; it is comparison across registers that makes them meaningful. In order to select texts for the small-scale study from the above sources, structured interviews were conducted with five low-achieving students in each grade. The aim of the interview with low-achieving students was to collect information on the texts that students had found difficult to understand and summarize during their studies. The presupposition was that the texts that low-achieving students find hard to process should abound in challenging grammar features, which are likely to contain register specific language items as well. Nearly unanimously, a newspaper article was chosen from the FCE course book used by 9th graders, and it was the chapter on viruses in the biology textbook that all the 10th grade students found hard to understand. As a result of the pilot analysis, the aspect of various infinitive forms and several additional grammar items that appeared in these texts were added to the tool, e.g. zero conditional, passive with an indirect object. The list of grammar items to be analysed swelled to 100 items along nine aspects of comparison.

Finally, in the last phase of the development of the analytical tool, two English teachers preparing 9th graders for the FCE exam and two biology teachers teaching in the 10th grade were interviewed in order to incorporate their insights and expertise in the current tool. The interviews with the teachers were carried out for the purpose of finalizing the grammar part of the analytical tool. As a result of the four interviews, only slight modifications were needed. Eight out of the nine aspects of comparison remained exactly the same, since all the teachers regarded the listed grammar items as essential to be familiar with when processing the given texts in their subjects, English or biology. Following the suggestions of my colleagues, it was the aspect of question tags, containing four items, which I decided to drop completely. Question tags were justly viewed by my colleagues as not being typical of written discourse in general; as a result, they were not included in the analytical tool. In order to follow the terminology of current English language course books which practising language teachers use, the various types of reported clauses were collected under the aspect of indirect speech. In the last phase of the development of the analytical tool, after the validation of the interviews with my colleagues, the aspect earlier termed as indirect speech was submerged into the aspect of relative clauses, since indirect speech utterances are nominal relative clauses in a linguistic sense. At this stage, the analytical tool comprising seven aspects of comparison with 96 items was finalized (for a complete list, see Appendix B).

4. 2 The corpora

In order to make the biology corpus representative of the texts students need to be able to process in biology after having taken their end-term FCE exams, it was first checked which biology texts bilingual students in the 10th grade are expected to study in their first academic term. In a structured group interview with five high-achieving 10th graders in English, students were given their biology book (Roberts, 1980) to pick the topics covered in the autumn term. Among the 10th graders, high-achievers in English were chosen this time as low-achievers in English tend to be more reluctant to share information about their studies, and they also seem to have a tendency not to remember precisely what has been covered in class. Each interviewee chose the same eight chapters, see Appendix C. To affirm the students' choices, the topics of the biology classes were tracked in the electronic register of the school written by the class's biology teacher from September to mid-January. It was observed that the list compiled by the students was exhaustive. Next, the eight chapters were transcribed electronically, in order to make them analysable, and a word count was run. As a result, it can be stated that the number of words in the eight biology chapters studied in the first academic term in the 10th grade amounts to 7,075.

After finding the relevant biology texts, I chose the general English texts that can serve as the basis of comparison in the register analysis. One of the guiding principles was that the general English texts should also contain approximately 7,000 words in total. The other principle that determined the choice of the texts was that the general English texts should be representative of the FCE reading exam, that is, all four parts of the exams should be present in the sample. As a complete FCE reading exam consists of about 2,000-2,500 words, it was clear that more than one exam had to be chosen. The last guiding principle in choosing the general English texts was that each part of the exam should be represented by an equal number of texts and, as much as possible, equal number of words. Matching all these criteria, twelve texts were chosen from the general English course book, their total length being 7,098 words. Table 1 shows the general English texts chosen, their lengths given in words, and the total length 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
1782 in total	**1786 in total**	**1777 in total**	**1753 in total**

Table 1. The reference corpus: the general English texts and their lengths in words

4.3 Procedures of data collection and analysis

After having chosen the texts, the analytical tool was applied in order to describe the register of biology texts written for secondary students. First the chunks of linguistic strings, the units of analysis were marked in the texts. As the most straightforward and also visibly indicated unit in a written text is the sentence, sentence boundaries were marked and counted in all the texts. Next, the grammar items of the analytical tool were tagged in each text, each grammar feature being indicated by a code number. Then the appearance of the code numbers was totalled in each text. The frequency of each grammar item was counted against the basic unit of analysis. In other words, the ratio of grammar items per number of sentences in the text was calculated. In order to describe the genre of biology texts, biology texts were compared with general English texts by means of computing t-tests. Since the frequency ratios of the two genres do not depend on each other, independent-sample t-tests were counted. If either of the two genres contained a given grammar item, its probability coefficient (Sig. 2-tailed) was tested to see if the difference in its frequency between the two registers was register-specific or sample-specific. The reason behind this was that frequency ratios with high probability coefficients (p>.05) in the sample do not show generalisable but only sample-specific traits. In other words, the number showing the percentage of the given pattern appearing in the sample by mere *coincidence* indicates a high likelihood of mere coincidence of the items' co-occurrence. In choosing the appropriate probability coefficient of a given frequency ratio, Levene's tests were consulted. In the case of the Levene's test showing a significant difference (p<.05) equal variances were assumed, while the lack of significant difference when running the Levene's test (p>.05) resulted in not assuming equal variances. This step in the procedure ensured the interpretation of the results to be reliable in distinguishing register specific traits from sample specific features. Finally, the mean value of each grammar item was compared and contrasted in the two registers.

5 Results and discussion

In the following section the selected Biology Corpus (BC) is compared and contrasted with the General English Reference Corpus (GERC) using independent sample t-test, to provide a thorough, comparative description in terms of the grammatical structures included in the analytical tool, namely:

1. tenses;
2. conditional structures;
3. passive voice and causative structures;
4. relative clauses;
5. nominal relative clauses;
6. infinitives;
7. prepositions at the end of sentences;
8. modals.

Tables with descriptive statistical data summarize the cases where the frequency of the use of a particular grammatical structure shows a significant difference between the two registers.

5.1 Tenses

In the tense aspect of comparison, the frequency of fourteen linguistic features was examined. Three of these proved to be significantly different in the two registers, see Table 2, seven of them showed no significant difference in the frequency of their presence, while four of them were not present at all in either of the registers.

Tense	**Probability coefficient**	**Mean Value**	
		Biology Corpus (BC)	**General English Reference Corpus (GERC)**
Present simple	p=.016	*M*=1.15	*M*=.83
Past simple	p=.03	*M*=.091	*M*=.364
Present perfect simple	p=.03	*M*=.016	*M*=.093

Table 2. The significantly different frequencies of tenses in the two registers

Considering the frequency of the **present simple tense**, there is a significant difference (p=.016) for the Biology Corpus (BC) (*M*=1.15) and the General English Reference Corpus (GERC) (*M*=.83). The results show that there are significantly more instances of using the present simple tense in biology texts than in general English texts. On average, the present simple appears in every single sentence in the biology texts, more precisely, there are 8 present simple verbs in 7 sentences; the general English texts contain fewer present simple items than sentences, as there are 5 present simple items in every 6 sentence.

Secondly, it is the **past simple tense** that shows a significant difference (p=.03) between the two genres. The BC (*M*=.091) tends to use the past simple three times more often than the GERC (*M*=.364) as an item of the past simple appears in every sentence in a biology text, while it is only about one in every three sentences in a general English text that uses past simple.

The third tense whose frequency shows a significant difference in the two genres (p=.03) is the **present perfect simple**. The BC applies nearly six times fewer present perfect items (*M*=.016) than the GERC (*M*=.093).

The difference in the occurrence of the **present continuous tense**, however, is not significant (p=.5) in the two genres. As the probability coefficient of the present continuous tense is not lower than 5 per cent, the mean values of their frequency for the BC (*M*=.091) and for the GERC (*M*=.364) cannot be used for describing the sample in a generalisable way. That is, the fact that there are five times as many present continuous items in the general English texts than in the biology texts should not be generalised and claimed to be a descriptive fact for the biology text genre; it is true for the sample under investigation only.

About the frequency of the occurrence of the **past continuous tense**, it can be stated that its difference is not significant in the two genres (p=.19). Although there are no items in the past continuous in the BC, it cannot be generally claimed that biology texts do not apply this tense, as the lack of significant difference prevents generalisations about the genre.

The frequency of the occurrence of the **past perfect simple** in the BC (*M*=.006) and the GERC (*M*=.016) is not significantly different either (p=.29). Consequently, the fact that the biology texts use three times fewer verbs in the past perfect simple than the general English texts is a statement true for this sample only; it cannot be given as a description of the genre of biology texts in general.

Giving an account of the frequency of the **'used to'** structure, the difference between the two genres is not significant (p=.34). Similarly to the lack of appearance of the past continuous in the BC, it can be claimed that no 'used to' items are present in the biology texts described. However, as the probability coefficient is too high, no generalisation can be made about the register of biology texts not containing the structure.

Additionally, the frequency of the occurrence of the **present perfect continuous** bears no significance (p=.37) in the two genres. The probability coefficient is not low enough to generalize the results of the sample, thus the fact that the BC (*M*=.001) uses seven times fewer present perfect continuous items than the GERC (*M*=.007) is a sample-specific characteristic feature. The reason behind this might be that both genres use the present perfect continuous sparingly (in the biology texts there is one in every 1000 sentences, in the general English texts there is one in every 143 sentences), which makes the genres indistinguishably similar.

The **future simple** is another tense that does not differentiate between the two genres since its high significance value (p=.3) allows no generalisations. Therefore, the fact that the GERC (*M*=.027) uses twice as many future simple verbs as the BC (*M*=.012) is a sample-specific description, which does not necessarily hold true for a larger corpus, that is, for the register of biology texts in general.

As for the frequency of the **'going to'** structure, it can be noted that the two genres do not differ significantly (p=.34). The BC contains no such future structure at all, and similarly, the GERC (*M*=.004) contains hardly any, as the structure is used in every 250th sentence only. That is, the two genres are similar in avoiding the use of 'going to' future.

There are four tenses which make no appearance in either of the two registers, where obviously no significance and mean values could be counted. Such tenses are the **past perfect continuous** and the more complicated 'will' futures, the **future continuous**, the **future perfect simple** and the **future perfect continuous**. The complete lack of their use in the registers makes statistical operations impossible, which prevents drawing generalisable conclusions about the genre of biology texts from this respect**.**

5.2 Conditional structures

The second aspect of comparison examined was conditional structures, such as zero, first, second, third, and mixed conditional. The range of these aspects shows no significant difference between the two genres at all, the probability coefficient, indicating the percent of coincidence in the sample, being way above 5% in all the cases.

Due to the lack of significant differences, the frequency results of conditional structures are true for the sample only; they cannot be regarded as descriptive ratios of the register of biology text in general. **Zero conditional** structures appear in the sample approximately two times more often in the BC (*M*=.03) than in the GERC (*M*=.012). In contrast, **first conditional** structures are less typical of the BC (*M*=.005) than of the GERC (*M*=.016), as they appear more than three times less frequently in the first genre. In a similar manner, the frequency of **second conditional** structures is twice lower in the BC (*M*=.005) than in the GERC (*M*=.013). **Third conditional** is a hypothetical structure that is not represented in the BC at all. The absolute lack of **mixed conditional** structures in both genres indicates no salient importance of connecting hypothetical past and present in the two corpora.

5.3 Passive voice and causative structures

Subsequently, the passive voice, both with a direct and an indirect object, along with causative structures such as ‘have it done’, ‘get it done’, ‘needs doing’, and ‘make somebody do something’ were examined in the sample. Apart from the passive voice with a direct object, none of the above structures were present in any of the texts; hence their frequency could not be identified in either of the genres in a statistical sense. It can be deduced, however, that the passive voice with an indirect object as well as the aforementioned causative structures are unlikely to be the most prominent characteristics of the two genres. Discussing the frequency of the **passive voice with a direct object** in the texts, it can be claimed that their probability coefficient (p=.053) is high, however slightly, for the texts to be considered significantly different, consequently no generalisations can be made about the genre of biology texts based on this aspect of comparison. The sample, however, indicates that the BC (*M*=.253) uses 1.5 times more passive items with a direct object than the GERC (*M*=.177).

5.4 Relative clauses

The aspect of relative clauses gave space to the comparison of the two genres in terms of defining and non-defining relative clauses, as well as various reduced relative clauses, such as simple and progressive participles both in the present and the past, in active and passive voices. Out of the numerous relative clauses examined, two of them proved to be significantly different in the two registers, see Table 3.

Relative clauses	**Probability coefficient**	**Mean Value**	
		Biology Corpus (BC)	**General English Reference Corpus (GERC)**
Defining relative clause without a relative pronoun	p=.048	*M*=.006	*M*=.034
Non-defining relative clause	p=.03	*M*=.009	*M*=.06

Table 3. The significantly different frequencies of relative clauses in the two registers

Considering the probability coefficient of the frequency of **defining relative clauses with a relative pronoun** (p=.065), it should be noted that the two genres are not significantly different, even if the BC (*M*=.117) contains 1.5 times more such clauses than the general English texts (*M*=.079). In contrast, the frequency of **defining relative clauses without a relative pronoun** shows a significant difference (p=.048) between the two genres. Therefore it can be claimed that the above grammatical item appears in the genre of biology texts (*M*=.006) nearly six times less often than in general English texts (*M*=.034). Likewise, the probability coefficient of **non-defining relative clauses** indicates a significant difference (p=.03) between the two genres. It follows that the appearance of the grammatical item being nearly seven times fewer in the biology texts (*M*=.009) than in the general English texts (*M*=.06) can be claimed as one of the characteristic features of the genre of biology texts. In contrast, the significance of **progressive participle clauses** (p=.765) is far too high to be generalized, thus no general characteristics of the BC can be accounted in this respect. The fact that progressive participle clauses' frequency is apparently the same for the BC (*M*=.0412) and for the GERC (*M*=.046) is a sample-specific statement, not generalisable for larger corpora. In contrast, **progressive participles in the past** are not present in either of the genres, which makes statistical analysis impossible in this respect. Based on the sample, what can be stated with certainty is that progressive participles in the past are of no primary importance for the genres of biology texts and general English alike. Similarly to progressive participle clauses, the frequency of **simple participle clauses** is not significantly different in the two genres (p=.956), thus the fact that the texts contain nearly an identical number of simple participle clauses, (*M*=.044 for the BC and *M*=.043 for the GERC), is a sample-specific observation. In contrast, **passive progressive participles**, either in the present or in the past, cannot be traced in any of the texts, thus these reduced relative clauses cannot be analysed by statistical means. As a consequence, the frequency of their use in either of the genres cannot be demonstrated; however, the lack of their vital importance can be expected with high certainty.

5.5 Nominal relative clauses

Describing the sample from the point of view of the frequency of nominal relative clauses, it can be affirmed that there are no significant differences between the genres of biology texts and that of general English, the reason for this being that all the probability coefficients are too high, above five per cent, to be generalisable. From this it follows that the description of the texts under investigation is sample specific in terms of nominal relative clauses, that is, the differences found are not genre specific. **Nominal relative clauses without a reporting verb without time shift** (p=.95) occur rather often in both types of text (*M*=4.5 for the BC and *M*=4.412 for the GERC). In contrast, **nominal relative clauses**

without a reporting verb with time shift (p=.12) does not appear the BC at all. However, **nominal relative clauses without a reporting verb with an infinite verb** are used with a modest frequency (p=.77), such clauses appear twice in each sentence in all the texts of the corpora on average (*M*=2 for the BC and *M*=2.375 for the GERC). Yet the frequency of **nominal relative clauses without a reporting verb with a preparatory 'it'** is considerably lower (p=.79), as they appear once in every eighth sentence in a biology text (*M*=.125) and approximately once in every sentence in a general English text (*M*=.083) in the sample. Reported speech without time shift, that is, **nominal relative clauses with a reporting verb without time shift,** appear in both types of texts (p=.43), once in every third sentence in the GERC (*M*=.333) and nearly three times less frequently in the BC (*M*=.125). Whereas the appearances of reported speech with time shift, or **nominal relative clauses with a reporting verb with time shift,** are not present in any of the texts in the sample. Slightly similarly, examples of reported speech followed by an infinite verb, that is, **nominal relative clauses with a reporting verb with an infinite verb,** cannot be identified in the BC, however, they are used in the GERC (p=.34) with a considerable frequency, there being present six times in five sentences (*M*=.083). The frequency of reported open questions, or **nominal relative clauses with a reporting verb with an open question** (p=.58) shows that they are applied two times as often in the GERC (*M*=.25) as in the BC (*M*=.125). In contrast, reported yes or no questions, **nominal relative clauses with a reporting verb with a yes or no question,** are not exemplified in any of the texts in the sample. It should be noted, however, that all the frequency ratios of nominal relative clauses are characteristic of the sample itself, not those of the genre of biology texts, as no significant differences could be traced in these respects.

5.6 Infinitives

The comparative aspect of infinitives comprises the analysis of the frequency of simple, progressive, active and passive forms of infinite verbs in the corpora. In this respect, one single infinitive form indicates significant difference between the two registers, see Table 4.

Infinitives	**Probability coefficient**	**Mean Value**	
		Biology Corpus (BC)	**General English Reference Corpus (GERC)**
Passive infinitive	p=.04	*M*=3	*M*=1.08

Table 4. The significantly different frequencies of infinitives in the two registers

Analysing the samples, the frequency of **passive infinitives** (p=.04) shows a significant difference between the two genres. The genre of biology texts (*M*=3) applies three times more passive infinitives than that of general English texts (*M*=1.08). On the other hand, the probability coefficient of **simple infinitive** (p=.46) shows no significant difference between the two genres, its frequency description does not give space for making generalisations. The BC (*M*=18.75) contains a large number of 19 simple infinitives in each sentence, while the GERC (*M*=15.667) uses 17 of them in every sentence on average. The fact that both registers apply an immense number of simple infinitives, however, cannot be claimed about the genre of biology texts in general, it should be treated as a sample specific trait. Similarly, the use of **progressive infinitives** cannot be generalised for the genres, since the probability coefficient (p=.34) is too high to give other than sample specific results. The BC contains no progressive infinitives at all; however, it appears in every fourth sentence in the GERC (*M*=.25). In contrast, **progressive passive infinitives** cannot be identified in the

GERC, while they are used in the BC in every fourth sentence (*M*=.25). However, the difference between the two genres is not significant (p=.23), that is, the fact that the use of progressive passive infinitives is considerably more frequent in the BC is a sample specific statement. Similarly to the frequency of progressive infinitives, **perfect infinitives** are not present in the BC at all, while they can be identified in every one and a half sentences in the GERC (*M*=.667), which difference, however, is not significantly different in the two genres (p=.074), thus it cannot be generalised. It can be claimed that more complex infinitives do not typically appear in either of the genres, neither **perfect passive infinitives,** nor **perfect progressive infinitives** or **perfect progressive passive infinitives** can be identified in any of the texts.

5.7 Prepositions at the end of sentences

Prepositions tend to appear at the end of sentences in questions, in clauses with infinitives and in relative clauses. All three cases were examined in the corpora; however, none of them show significant differences in the two genres. Moreover, none of them are present in the BC; thus it can be claimed that the genre of biology texts is highly unlikely to contain prepositions at the end of sentences.

5.8 Modals

Among the comparative aspect of modals, forty-one different modals were examined in the samples. Analysing the results, it can be observed that most of the modifying auxiliaries show no significant difference between the two registers, that is, the difference in their frequencies is mainly sample specific. In the case of three modals, however, the probability coefficient is low enough, smaller than five per cent, to indicate a significant difference between the two registers, see Table 5.

Modals	**Probability coefficient**	**Mean Value**	
		Biology Corpus (BC)	**General English Reference Corpus (GERC)**
'can' expressing ability in the present	p=.013	*M*=5.5	*M*=2.25
'may' expressing the level of certainty in the present	p=.038	*M*=1.25	*M*=.5
'must' expressing obligation in the present	p=.028	*M*=1.25	*M*=0

Table 5. The significantly different frequencies of modals in the two registers

First, the frequency of the use of **'can'** expressing ability in the present is register specific for biology texts (p=.013). It tends to appear extensively in the genre of biology texts, more than five times in each sentence on average (*M*=5.5), while its appearance in GERC is half as massive as that, being used approximately two times in each sentence (*M*=2.25). Secondly, the frequency of **'may'** expressing the level of certainty in the present shows an account typical of the genre of biology texts (p=.038). This modal appears three times in two sentences on average in biology texts (*M*=1.25), while far more scarcely in general English texts, appearing only once in every second sentence there (*M*=.5). Finally, it is the frequency of obligation in the present expressed by **'must'** that differs significantly in the two genres (p=.028). The significant dissimilarity lies in the fact that the genre of biology texts uses this modal auxiliary three times in two sentences (*M*=1.25), while it makes no appearance in the genre of general English texts at all.

Besides the above three modals, no other modal auxiliary can be described as genre-specific due to their far too high probability coefficients. Hence, the frequency of the modal verb **'able to'** expressing ability in the present and the future (p=.48) being twice as high in the BC (*M*=.375) than in the GERC (*M*=.167), describes the sample under investigation, and not the genre of biology texts. In a similar manner, the frequency of the modal **'could'** expressing ability in the past (p=.66) describes the sample, being present approximately twice in three sentences in the BC (*M*=.625), while appearing twice in five sentence in the GERC (*M*=.417). In contrast, the modal auxiliary **'able to'** expressing ability in the past does not appear in the BC at all, while the GERC contains it in every twelfth sentence (p=.34, and *M*=.083). In an absolutely identical manner, the statistically not significant auxiliaries **'must', 'bound to', 'ought to'** expressing the level of certainty in the present and future, as well as **'may have'** and **'would have'** expressing the level of certainty in the past are not represented in the BC, while they appear in the GERC once every twelfth sentence (p=.34 and *M*=.083). The modal verbs **'would'** and **'would have'** with the function of distancing from reality are used with nearly the same frequency (p=.93), appearing in both registers five times in six sentences (*M*=.833 for the BC and *M*=.875 for the GERC). In contrast, the auxiliary **'will'** expressing the level of certainty in the present cannot be found in the BC, while it is present once in every fourth sentence in the GERC (p=.34 and *M*=.25). Similarly, the modal verb **'should'** expressing the level of certainty in the present cannot be identified in the BC, however, it appears once in every second sentence in the GERC (p=.26 and *M*=.5). Showing a four times more frequent use in the BC (p=.12), the modal auxiliaries **'might'** expressing the level of certainty in the present and **'should'** expressing an obligation in the present appear once in every third sentence in the BC (*M*=.375), yet only once in every twelfth sentence in the GERC (*M*=.083). Obligations in the present and in the past expressed by **'have to'** and **'had to'** are both less frequent in the BC. The first one is used twice as scarcely in the BC as in the GERC (p=.5, *M*=.125 and *M*=.25), the second one is applied nearly four times less often (p=.34, *M*=.125 and *M*=.417). Obligations expressed by **'to be to'** in the present and in the past are present in the GERC with the same frequency (*M*=.167), however they appear slightly more frequently in the BC in the present (p=.68, *M*=.25), while they are not applied in the past in the BC at all (p=.25). In contrast, the obligation expressed by **'need'** in the present is three times more frequent in the BC (p=.38, *M*=.25 and *M*=.083).

A considerable number of auxiliaries, twenty in particular, could not be identified in any of the texts in either genre, these being criticism expressed by **'will'**, wishes expressed by **'may'**, present and past willingness and refusal expressed by **'will'** and **'would'** respectively, polite requests expressed by **'would'**, the levels of certainty in the present expressed by **'could'** and **'can't'**, the levels of certainty in the past expressed by **'must have'**, **'bound to'**, **'will have'**, **'might have'**, **'could have'**, **'can't have'**, obligations in the present expressed by **'mustn't'** and **'had better'** as well as obligations in the past expressed by **'should have'**, **'ought to have'**, **'needn't have'**, and **'didn't need to'**.

6 Summary

This paper gives a thorough description of the register of English language biology texts written for secondary school students with regard to its characteristic use of grammar from the point of view of EFL teaching, a yet untapped area within the field of discourse analysis and ESP. An analytical instrument with seven aspects of comparison embracing 96 grammar items was developed to carry out a grammatical register analysis. With respect to the first research question of the study, the instrument proved to be reliable as it produces "consistent results in a given population in different circumstances" (Dörnyei, 2007, p. 41). Furthermore, external validity, the extent to which the findings are generalisable, is also ensured as the results of the independent sample t-tests were checked to determine whether the scores are generalisable for the register of biology texts or are merely characteristics of the very collection of biology texts. Respectively, the results were reported as either register specific or sample specific. Internal validity of the analysis, the fact that the instrument measures what it is intended to measure, was ensured by seeking expert judgement through interviewing four of my colleagues.

Collecting the findings to the second research question, the BC can be described by the lack of versatile use of tenses, the preference of simple to continuous tenses and the underuse of complex forms. The BC uses significantly more instances of the present simple and the past simple tenses than the GERC. The appearance of the present continuous tense is five times fewer, and there are no occurrences of the past continuous or the past perfect continuous tenses or the '*used to*' structure in the BC at all. The present perfect simple and continuous tenses appear sparingly in the register of biology texts, the first one significantly fewer times in the register under analysis than in general English texts. The appearance of the future simple tense in the BC is half as many as in the GERC. The occurrences of the future continuous, future perfect simple and continuous tenses, along with the '*going to*' structure are nought in the BC.

Although conditional structures do not significantly characterise the register of biology texts, it can be stated that the BC uses twice as many instances of zero conditional than the GERC. In contrast, first conditional structures appear three times fewer in the BC than in the GERC, and the second conditional structure is two times underrepresented in the BC. While third and mixed conditional structures are not present in the BC at all.

The BC uses 1.5 times more instances of passive voice with a direct object than the GERC; however, the difference is statistically not significant. Other types of passive forms (passive voice with an indirect object, causative structures such as '*have it done*', '*get it done*', '*needs doing*', and '*make somebody do something*') are completely absent in the BC.

An informative, academic register that abounds in clarifying concepts through definitions, the BC contains 1.5 times more occurrences of defining relative clauses with a relative pronoun, while defining relative clauses without a relative pronoun appear significantly fewer times, six times less often than in the GERC. Non-defining relative clauses, providing extra information, appear seven times fewer in the BC than in the GERC. The frequency of simple and progressive participle clauses in the present is the same in the two corpora, while progressive participles in the present and in the past do not appear in the BC at all.

Nominal relative clauses without a reporting verb without time shift or with an infinite verb occur rather frequently in the BC, similarly to that of the GERC, however, the above grammar item with time shift is not present in the BC at all. The appearance of nominal relative clauses without a reporting verb with a preparatory '*it*' is eight times fewer in the BC than in the GERC. Occurrences of nominal relative clauses with a reporting verb without time shift are three times fewer in the BC; however, such a grammar item with time shift or with an infinite verb is totally absent in the BC. Reported open questions appear half as many times in the BC as in the reference corpus, and reported yes or no questions are not exemplified at all.

The BC uses simple infinitives as often as the GERC, and significantly more passive infinitives than the GERC. While progressive infinitives and perfect infinitives are not present in the BC, progressive passive infinitives, a grammar item completely absent in the GERC, appear frequently. More complex infinitives (perfect passive, perfect progressive, perfect progressive passive ones) are absent in the register of biology texts.

Prepositions at the end of sentences (either in questions, or in clauses with infinitives or in relative clauses) are not characteristic of the register of biology texts to any extent; no such grammar item appears in any of the biology texts.

Three of the modal verbs are register specific for biology texts; significantly different from general English corpus. The modal '*can*' expressing ability in the present appears two times more often in the BC than in the GERC, the modal auxiliary '*may*' expressing the level of certainty in the present is three times more abundant in the BC, while '*must*' expressing obligation in the present appears three times in every two biology sentence and completely absent in the GERC. The modals used more massively in the BC text than in the reference corpus are '*able to*' expressing ability in the present and the future, '*could*' expressing ability in the past, '*might*' expressing the level of certainty in the present, '*should*' expressing an obligation in the present, obligations expressed by '*to be to*' in the present, obligation expressed by '*need*' in the present; however they show no significant differences between the two registers. Nearly thirty modal verbs are not exemplified in the register of biology texts.

Apparently, no register can be fully described by providing merely grammatical accounts. As a result, further research is needed into other dimensions of the BC. According to the literature, one of the aspects worthwhile investigating is "the associations between words and grammatical structures", called lexico-grammatical features (Biber, 1998, p. 105), that is, features revealing the relationship between words and their environment from a grammatical point of view. Besides grammar and lexico-grammar, biology texts obviously have their own special lexis, including academic English and ESP related vocabulary, which should also be discussed when giving a complete account of the register. Along with the above elements, the register is to be carefully examined from a macro-structural point of view as well, involving sentence and paragraph structure, and text organisation.

Still, the study adds to the field of discourse analysis, more specifically to the genre analysis of textbooks, by providing a thick description of the register of English language biology texts written for secondary school students with regard to its characteristic use of grammar from the point of view of EFL teaching. The results of the analysis presented in this paper can be applied when working out the grammar foci of biology ESP syllabi for secondary school students. Furthermore, the findings can provide insights for general EFL teachers in bilingual secondary schools preparing their students to study academic subjects in English.

Proofread for the use of English by: Courtney Kersten, freelance English teacher

References:

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

Biber, D. (1991). Oral and literate characteristics of selected primary school reading materials. *Text, 11,* 73-96.

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

Coates, J. (1983). *The semantics of the modal auxiliaries*. London: Croom Helm.

Cunningham, S., & Moor, P. (2005). *New cutting edge. Intermediate.* London: Longman.

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford: Oxford University Press.

Falla, T., & Davies, P. A. (2008). *Solutions. Elementary.* Oxford: Oxford University Press.

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

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.

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

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

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

Thompson, S. A. (1983). Grammar and discourse: The English detached participle clause. In F. Klein-Andreu (Ed.), *Discourse perspectives on syntax* (pp. 43-65). New York: Academic Press.

Tottie, G. (1985). The negation of epistemic necessity in present-day British and American English. *English World-Wide, 6,* 87-116.

Vince, M., & Emmerson, P. (2003). *First Certificate language practice.* Oxford: Macmillan Education.

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

APPENDIX A

Linguistic features Biber (1998, p. 148) analysed when describing ESP registers

adverbial subordinators, adverbs, agentless passive, amplifiers, analytic negation, attributive adjectives, *be* as main verb, *by* passives, causative subordination, conditional subordination, conjunctions, contractions, demonstrative pronoun, discourse particles, *do* as pro-verb, final prepositions, first-person pronouns, general emphatics, general hedges, indefinite pronouns, infinitives, necessity modals, nominalization, non-phrasal coordination, nouns, past participial adverbial clauses, past participial postnominal clauses, past tense verbs, perfect aspect verbs, phrasal coordination, pied-piping constructions, place adverbials, possibility modals, prediction modals, prepositions, present participial clauses, present tense verbs, present private verbs, pronoun *it*, public verbs, second-person pronouns, sentence relatives, split auxiliaries, suasive verbs, synthetic negation, tense verbs, *that* deletion, third-person possibility modals, pronouns, time-adverbials, type-token ratio, *wh*-clauses, *wh*-questions, *wh*-relative clauses on object position, *wh*-relative clauses on subject position, word length

APPENDIX B

The finalized analytical tool

Aspect of comparison	**Linguistic feature**
Tense	Present simple
	Present continuous
	Past simple
	Past continuous
	Past perfect simple
	Past perfect continuous
	Used to
	Present perfect simple
	Present perfect continuous
	Future simple
	Future continuous
	Future perfect simple
	Future perfect continuous
	Going to
Conditional	Zero conditional
	1st conditional
	2nd conditional
	3rd conditional
	Mixed conditional
Passive	Passive with a direct object
	Passive with an indirect object
	Causative: have it done
	Causative: get it done
	Needs doing
	Make sy do sg
Relative clauses (RC)	Defining RC with a relative pronoun
	Defining RC without a relative pronoun
	Non-defining RC
	Reduced RC: participle clause: -ing
	Reduced RC: participle clause: having past participle
	Reduced RC: participle clause: -ed
	Reduced RC: passive participle clause: being done
	Reduced RC: passive participle clause: having been done
	Nominal RC (NRC): without a reporting verb without time shift
	Nominal RC (NRC): without a reporting verb with time shift
	Nominal RC (NRC): without a reporting verb with an infinitive verb
	Nominal RC (NRC): without a reporting verb with a preparatory '*it*'
	Nominal RC (NRC): with a reporting verb without time shift
	Nominal RC (NRC): with a reporting verb with time shift
	Nominal RC (NRC): with a reporting verb with an infinitive verb
	Nominal RC (NRC): with a reporting verb with an open question
	Nominal RC (NRC): with a reporting verb with a yes or no question

Infinitive	Simple infinitive
	Passive infinitive
	Progressive infinitive
	Progressive passive infinitive
	Perfect infinitive
	Perfect passive infinitive
	Perfect progressive infinitive
	Perfect progressive passive infinitive
Preposition	Preposition at the end of the clause: in questions
	Preposition at the end of the clause: with an infinitive
	Preposition at the end of the clause: in relative clauses
Modal verbs	Ability in the present: can
	Ability in the present, future: able to
	Ability in the past: could
	Ability in the past: able to
	Present habits, typical behaviour, criticism: will
	Wish: may
	Present willingness and refusal: will
	Past willingness and refusal: would
	Past habit, typical action: would
	Polite request: would
	Distancing from reality: would
	Level of certainty in the present: must
	Level of certainty in the present: bound to
	Level of certainty in the present: will
	Level of certainty in the present: should
	Level of certainty in the present: ought to
	Level of certainty in the present: may
	Level of certainty in the present: might
	Level of certainty in the present: could
	Level of certainty in the present: can’t
	Level of certainty in the past: must have
	Level of certainty in the past: bound to
	Level of certainty in the past: will have
	Level of certainty in the past: may have
	Level of certainty in the past: might have
	Level of certainty in the past: could have
	Level of certainty in the past: can’t have
	Level of certainty in the past: would have
	Obligation in the present: must
	Obligation in the present: have to
	Obligation in the present: ought to
	Obligation in the present: need
	Obligation in the present: mustn’t
	Obligation in the present: don’t have to
	Obligation in the present: should
	Obligation in the present: had better
	Obligation in the present: to be to

	Obligation in the past: had to
	Obligation in the past: should have
	Obligation in the past: ought to have
	Obligation in the past: needn’t have
	Obligation in the past: didn’t need to have
	Obligation in the past: to be to

APPENDIX C

Texts of the biology book taught in the first academic term in the 10th grade

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