Register analysis of English for specific purposes discourse

An in-depth exploratory and descriptive
theory- and corpus-based study
of the case of biology texts
in secondary education in Hungary.

PhD Dissertation

Candidate: **Natália Borza**

Supervisor: Krisztina Károly, PhD, habil.

PhD Programme in Language Pedagogy

Doctoral School of Education

Eötvös Loránd University

Budapest, 2015

Eötvös Loránd University
Faculty of Pedagogy and Psychology
Doctoral School of Education
Head of Doctoral School: Éva Szabolcs, PhD

PhD Programme in Language Pedagogy
Faculty of Humanities
School of English and American Studies
Founder and Honorary Programme Director: Péter Medgyes, DSc,
Programme Director: Krisztina Károly, PhD, habil.
Director of Studies: Dorottya Holló, PhD, habil.

Defence Committee:

Head:	Péter Medgyes, DSc
Internal referee:	Brigitta Dóczi, PhD
External referee:	Zsuzsanna Zsubrinszky, PhD
Secretary:	Dorottya Holló, PhD, habil.
Members:	Zsolt Király, PhD
	Pál Heltai, PhD
	Zsuzsanna Kurtán, PhD, habil.

Budapest, 2015

Acknowledgements

It is a genuine pleasure to express my deep sense of gratitude to my supervisor, Krisztina Károly, whose dedication and keen interest in the research of language pedagogy and overwhelming attitude to provide academic help for her students empowered me to conduct my research in a smooth manner. Her professional guidance, timely suggestions and kindness contributed greatly to the accomplishment of this project.

I thank profusely all the staff of the PhD programme in language pedagogy who were instructing me between 2010 and 2013. In particular, I would like to acknowledge with deep appreciation the invaluable help of Dorottya Holló, whose scholarly advice and academic approach helped me to a great extent to find the focus my research. It is my privilege to thank Péter Medgyes, whose cheerful inspiration, enthusiasm and dynamism encouraged my persistence in times of difficulties. I am also thankful to my friend and peer at the PhD programme, Mária Adorján, who was kindly available for discussions, was willing to share her insights both about the guiding principles and the practicalities of this research.

I wish to express my enormous gratitude to several colleagues of mine at the bilingual secondary school. I am especially grateful to Anikó Bognár, who tirelessly showed great effort in offering professional assistance in the field of bilingual education as well as providing personal support. I feel fortunate to have had the opportunity for years for sharing knowledge and experience considering the beauties and challenges of teaching in the bilingual programme with Bernadett Szabadkai, Ágnes Szili, Móni Fekete, and Lilla Jéri.

My thanks are also extended to Ingrid Hamow, librarian at the Institute of Philosophy at Eötvös Loránd University, who passionately managed to create an inspiring and peaceful environment in the library, which served as an ideal place for the activity of writing the present dissertation.

This research also owes deep thanks to one of my students, Georgiosz Kukumzisz, to whom I wish to offer a special acknowledgement for sharing his natural, insatiable spirit of inquiry.

Finally, I take great pleasure in recognizing the generous help of my friend, Péter Tóth, whose constant encouragement and lively support throughout the research enhanced my enthusiasm in continuing the present investigation. I am grateful for his exceptional patience, with which he accepted the lengthy period of time I devoted to this research project. His reassuring attitude of joyfully welcoming my scholarly interest greatly reinforced the successful execution of this analytical study.

Abstract

While the prevalent linguistic features of academic writing at a tertiary level are widely researched in the field of register analysis, those at a secondary level have not yet been thoroughly investigated. Even less attention has been dedicated to the exploration of the linguistic characteristics of biology textbooks in English for secondary students. The present study seeks to address this lacuna from a pedagogical perspective. Accordingly, the aim of the current research is to design a pedagogically oriented text-analytical instrument (POTAI) which is capable of yielding linguistic data relevant for ESL and ESP teachers. A further aim of the research is to apply the POTAI to the corpus of biology texts (BIOCOR) which the 10th grade students in a bilingual secondary school in Hungary are assigned to process in order to gain insights into the possible linguistic reasons why the target group finds the texts challenging to comprehend. Data was collected through quantitative and qualitative register analytical methods and through interview studies with ESL and biology teachers instructing in the bilingual programme of the secondary school. The findings of the research project reveal that the newly designed POTAI is a reliable tool, which is appropriate for producing valid linguistic data applicable by ESL and ESP teachers. The central finding of applying the POTAI to the BIOCOR exposes that the biology textbook register for secondary students is below the CEFR B2 level, which is the linguistic level students at the bilingual secondary school are expected to pass at the end of the 9th grade. Elucidating the linguistic level of difficulty of the BIOCOR through a fine-grained analytical description is assumed to be of assistance to ESL and biology ESP teachers alike.

Table of contents

List of tables

List of figures

List of acronyms

ARI	automated readability index
AWL	academic word list
BICS	basic interpersonal communicative skills
BIOCOR	the corpus of the biology texts for secondary students
CALP	cognitive and academic language proficiency
CEFR	Common European Framework of Reference for Languages
CLAWS7	constituent likelihood automatic word-tagging system version 7
CLIL	content and language integrated learning
ESL	English as a second language
ESOL	English for speakers of other languages
ESP	English for specific purposes
FCE	Cambridge First Certificate in English
M	mean value
MD	metadiscourse
MDA	multidimensional analysis
NRC	nominal relative clause
p	probability coefficient
POTAI	pedagogically oriented text-analytical instrument
REFCOR	the reference corpus
RQ	research question
TMD	textual metadiscourse
SFL	systemic functional linguistics
UCREL	University Centre for Computer Corpus Research on Language

Register analysis of ESP discourse

An in-depth exploratory and descriptive theory- and corpus-based study of the case of biology texts in secondary bilingual education in Hungary.

1 Introduction

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 and theoretically-driven research to investigate the possible nature of the problem and to offer feasible solutions is motivated by my own experience of going through similar difficulties at the same school in the same bilingual program as a student and later observing the regular reappearance of the same hardships among the 10th graders as a practicing English language teacher.

At the end of their first year at the secondary school, 9th graders are expected to take an upper-intermediate level Cambridge examination, level B2 in the Common European Framework of Reference for Languages (CEFR), the First Certificate in English (FCE). Students who pass this examination are thought to be able to study academic core subjects in English (such as mathematics, history, geography, physics and biology) from the following year on. However, when it comes to studying various subjects in English as a foreign language in the 10th grade, students have considerable difficulties meeting the academic requirements. Although at this point they 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 here in particular as its status differs deeply from that of the other subjects in the school: there is no biology ESP instruction provided for the

students in the 9th grade since the special terminology of the discipline is thought by the biology teachers working at the school to be far too diverse and difficult for 9th graders to grasp without studying the subject itself. 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 the *other* four specialized English classes (history, mathematics, physics, and geography). Accordingly, as an educator teaching general English in the 9th grade, I have become interested in what my students need to know in terms of English language in order for them to handle biology texts successfully in the 10th grade.

Developing a framework conceived in a language-pedagogical perspective is unique of its kind as no model has been devised hitherto that analyses the written register of biology textbooks for secondary students from the point of view of English as a second language (ESL) teaching (for further details see Section 2.5 on pp. 33-37). The aim of the current research is to develop a pedagogically oriented text-analytical instrument (in the study referred to briefly as POTAI) that is capable of producing reliable and valid data concerning the dominant register features of biology texts used in the instruction of mostly monolingually raised Hungarian students in a bilingual secondary school. This theoretical aim serves a practical one too, namely, to apply the POTAI to the biology texts used by 10thgrade students at the bilingual secondary school in order to describe the register of the biology corpus students need to process during their studies from the point of view of ESL teaching. This second aim 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 school as to what extent the language foci of the preparatory year enable students to handle the language use of the biology texts 10th graders are assigned to process. Besides gaining a deeper understanding of the 10th grade bilingual students' needs in terms of English

language and thus supporting my own and my colleagues' professional development as general English teachers, this exploratory and descriptive corpus-based study can provide insights for future biology ESP teachers, once biology ESP has been included in the 'zero year' language programme of the secondary bilingual school. Although the present research launches a close investigation into describing the language use of two types of texts at a particular secondary school in Hungary, the results of the enquiry are not restricted to the secondary school at hand, they can be meaningfully transferred and applied by educators working in any English-language international school where the alumni includes non-native students.

First the study reviews the relevant literature (see Chapter 2 on pp. 5-52) to arrive at the clarification and particular interpretation of the concepts used in the research within the field of text analysis (such as text, discourse, genre, and register) as well as to find the theoretical basis for the most reasonable ways of text analysis in the present research environment. Then the corpora under investigation are contextualized by providing a thick description of the setting where they are used: the bilingual immersion programme of the secondary school (see Section 3.1 on pp. 56-63). Next the two corpora (their sources, size and the process of compilation) are introduced (see Section 3.2 on pp. 64-71). This is followed by the presentation of the methods of data collection and data analysis (see Section 3.3 on pp. 72-111), where the development of each component (including all the linguistic variables) of the POTAI is elucidated. After the presentation of the design of the research project, the data resulting from the application of the POTAI to the two corpora are demonstrated (see Chapter 4 on pp. 118-202). The meaning of the figures in the research environment produced by the components of the POTAI is interpreted in a comparative manner across the two registers. Subsequently, pedagogical implications are formulated for ESL and ESP teachers based on the results of the text analysis (see Chapter 5 on

pp. 203-209). Finally, the answers to the issues addressed by the research questions are summarized and possible future avenues of the current research are drawn (see Chapter 6 on pp. 209-213).

Keeping the 10th graders' difficulty of tackling academic subjects in English in the foreground, the present theoretically and pedagogically motivated study attempts to answer the following umbrella questions:

A) By what means, relevant to English as a second language teaching, is it possible to describe the dominant register features of the biology texts used at an English-Hungarian bilingual secondary school?

B) From a linguistic point of view, to what extent do the general English reading texts 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?

These broad questions, which designate the centre of attention of the research, are explored by systematically searching answers to a number of focally more pointed subquestions (in the study referred to as the Research Questions), which are detailed in the first part of the methods section of the dissertation (on pp. 54-56).

2 Review of the literature

2.1 Shifts and development in English for specific purposes (ESP)

2.1.1 The origins of ESP

The teaching of English for specific purposes (ESP) is a field of English language teaching (ELT) that has a relatively short history of five decades hitherto. Earlier than fifty years ago, the field of ESP was not brought to life as the factors shaping ELT were dominantly different from the ones of today, which did not favour the appearance of this specialization. In an attempt to uncover the reasons for the emergence of ESP in the 1960s, Hutchinson and Waters (1987) list three markedly different ways of justification. Firstly, they argue that after the Second World War the market for ELT changed as technology and commerce expanded on an international scale. The worldwide expansion created the need for an *international language*, which was generally accepted to be English. Compared to learners of English before the mid-20th century, the new generation of language learners of the post-war world had more clearly graspable aims, namely, they wanted to become successful in selling their trade, skills and expertise in English. The second reason for the birth of ESP is claimed to be a new trend in linguistics. Relying on Widdowson's argument (1978), Hutchinson and Waters (1987) maintain that by the late 1960s and early 1970s not only did linguists endeavour to describe the formal features of English grammar, but the importance of characterizing the different ways in which English was applied in *real communication* also came to the foreground. This change in the approach of English linguistics helped the development of a different view in ELT, which favoured the evolution of ESP. The significance of recognizing the fact that there are several varieties of English used in different situations which differ from one another and whose specific traits can be identified and taught to language learners grew evidently. The results of research into distinguishing different varieties of scientific and technical English were promptly incorporated in tailoring English

language courses (Candlin, Bruton & Leather, 1976; Ewer & Latorre, 1969; Ewer & Huges-Davies, 1971; Selinker & Trimble, 1976; Swales, 1971). Finally, as a third force fostering the appearance of ESP teaching, Hutchinson and Waters (1987) mention developments in educational psychology in the late 1960s. Following Rogers' lines of argument (1969), *learners' individual needs and interests* were treated as crucial factors in motivation. This student-centred approach seeped into ELT, which can be traced by the fact that language courses satisfying learners' professional needs through teaching the very variety of English that was relevant to the learners' needs were commenced to be designed.

2.1.2 Major shifts in the course of ESP

Throughout its half a century long development, ESP has focused on describing the langue used in particular professional settings or academic disciplines with the "ultimate goal of developing instructional materials that will help students learn the particular language patterns" (Biber & Conrad, 2009, p. 3). Despite the presence of such a clear aim, the course of ESP has not been homogeneous, four different stages can be distinguished in its history. At its beginnings in the 1960s, it was *grammatical* and *lexical features* at a sentence level that linguists chose to apply when differentiating particular varieties of English (Ewer & Hughes-Davies, 1971; Ewer & Latorre, 1969; Halliday, McIntosh & Strevens, 1964; Swales, 1971). The academic interest in sentence grammar resulted in teaching materials abundantly applying and thus getting learners practice language forms students primarily meet when reading subject-specific texts, while carefully avoiding language forms that have low priority in the given variety of language. Diverging from the sentence level approach of investigation, the second phase of ESP history saw the development of rhetorical analysis (Allen & Widdowson, 1974; Lackstrom 1973; Trimble 1985; Widdowson, 1978). The school of *rhetorical analysis* opened its scope above the sentence level, and turned its attention to the

linguistic ways how sentences combine. Instead of tracking down characteristic grammatical features of a given variety of English, the focus in the 1970s and 1980s shifted to organisational patterns in texts. Linguists aimed at uncovering discourse markers or linguistic features by which organisational patterns are signalled. The pedagogical result of such enquiries was the production of teaching materials bountiful of tasks aiming at revealing textual patterns and text-diagramming exercises. The next stage of ESP development was marked by a synthetizing approach, the method of *target situation analysis* (Cohen & Mannion, 1980; Drobnic, 1978; Hutchinson, Waters & Breen, 1979; Mackay, 1978; Richterich, 1984; Richterich & Chancerel, 1980). The aim of ESP course designs within this approach was the detection of situations in which the learners use English. In order to prepare the learner to function effectively in the target situation, first the learners' needs were mapped and collected in learner profiles, a model developed by Munby (1978). The nature of the approach is synthetizing as it joins diverse points of research, such as the purpose of communication, the setting, the means of communication, language skills, functions, structures, etc. At the same time, it primarily keeps the needs of the language learner in the focal point. The fourth stage of ESP development saw a dramatically different approach from the previous three phases. Rather than paying attention to the different forms of language use, either at a sentence level or above, researchers endeavoured to discover *cognitive processes* that underlie language use (Alderson & Urquhart 1984; Gellet, 1981; Nutall, 1982). Research projects were launched to learn more about the working processes that language learners apply when extracting meaning from discourse (Brazilian National ESP Project, Holmes, J. 2012; UMESPP, University of Malaysia ESP Project, Khairi, 2001), where mainly reading and listening strategies were investigated. The resulting teaching materials considered the learners as essentially thinking beings, and attempted to motivate them to become conscious of and reflective on the interpretive processes that allow them to handle surface forms of the

language. Teaching materials of this view characteristically involve tasks that make the learner aware of the importance of the context when guessing the meaning of an unfamiliar word or ones that put emphasis on how message is conveyed through the visual layout of a text.

The current theoretically and pedagogically oriented investigation makes extensive use of the elements of all four stages of ESP development. This research heavily draws on identifying grammatical and lexical features of a particular variety of English, meanwhile it also analyses overt text organizing patters that create a logical order of the flow of sentences within the ESP variety. Simultaneously, the learning environment where the given variety of English is used is discovered and described in great detail. The manner how the particular texts are processed is also mapped out and the nature of further activities the texts give rise to in the learning environment is examined likewise. Finally, reflections on cognitive processes of the students who read the specific texts are traced through group and individual interviews. Among these ways of investigation, text analysis approaches are the most emphatic ones, whose results are backed and made more comprehensible by information gained from interview studies. As the research primarily relies on analysing texts, let us first clarify the opaque concept of text.

2.2 Discourse and text

In order to communicate with one another it is language that we use. Communication takes place through discourse, a general term used for both spoken and written language (Sandes & Sanders, 2006). The term discourse has several different ways of interpretation in the literature of the linguistic study of discourse. Halliday (1990, p. 41) argues that discourse is "a unit of language larger than a sentence and which is firmly rooted in a specific context."

However, the term is not unanimously defined in the literature. The concept appears in an overlapping way with the notion text, and the two terms are even used synonymously. Yet at other places they convey contrastive shades of meaning (Károly, 2007). Widdowson (1996) applies the term text to the product of the process of discourse, de Beugrande and Dressler (1981) argue that a text is the product itself and the process as well. There is no consensus whether the terms mean written or spoken products either. Some linguists differentiate the two terms as a text being written while discourse being spoken (Coulthard, 1985; Sanders & Sanders, 2006). Nevertheless, others use both terms for written and spoken products (Ford et al, 2001; Trask, 1999). Some even argue that the terms discourse and text are such elemental and underlying ones that their nature cannot be defined categorically with certainty, which in turn might help new theories emerge relying on slightly different connotations of the terms (Kocsány, 2002). The present research follows the de Beugrandian tradition (1981) as far as treating the technical term text as a communicative event that is both a product and a process, as well as using the term discourse synonymously with text (de Beugrande, 1997). Considering the nature of communication, the current investigation is in the wake of Trask (1999) by applying the two compatible terms for written and spoken processes and products alike.

2.3 Genre and register

2.3.1 Genre and register as overlapping concepts

Certain instances of communicative events and a number of discourse samples or texts display several kinds of similarities, on the basis of which they might be labelled as belonging to one common class: a genre, traditionally a literary construct (Hyon, 1996), or a register. According to Swales (1981, 1986, 1990), whose research has been seminal in shaping genre theory, the crucial similarity that groups a pool of discourse items in a shared category does

not lie in the mere resemblance of the surface form of the language used in the items, but more importantly, ”the principal critical feature that turns a collection of communicative events into a genre is some shared set of communicative purposes” (Swales, 1990, p. 46). In this view, the formation of a genre is a response to communicative purposes in common, where the members of a discourse community typify the conventions of the genre while achieving their shared communicative goals. Applying the Swalesian definition to the present research, texts in general English course books definitely share a set of communicative purpose (they aim to provide written samples of the target language for EFL learners), and so do texts in a biology textbook (they intend to inform students of educationally selected topics of biology). Consequently, both groups of texts in the present investigation might be treated as unmistakeably different genres. Following the Swalesian idea, biology textbooks and general English course books can be distinguished as two distinctively different genres since they are written for different audiences with different purposes. As Lee (2001) points out, the term genre is “assigned on the basis of external criteria such as intended audience, purpose, and activity type, that is, it refers to a conventional, culturally recognised grouping of texts based on properties other than lexical or grammatical (co-)occurrence features” (p. 38). A given variety of language, or discourse, is used by a specific community, which Swales (1990) calls a discourse community. Among his criteria of a discourse community, Swales (1990) maintains that “a discourse community has acquired some specific lexis” (p. 28). This point of view is further explained by Ramanathan and Kaplan (2000) by claiming that “members of a discourse community, who become insiders of the community, partially out of long-standing participation in that community, evolve a selective lexis – modes of communication, acronyms, jargons, textual forms – that facilitates easy communication among peers” (p. 177). Since the two sets of texts under investigation belong to two

distinctive genres read by two different discourse communities, it possible and worth considering to what extent their language use overlaps and differs.

The term genre is frequently used interchangeably with that of register, their definitions are compatible (Lee, 2001; Rittman, 2007). As Biber and Conrad (2009) warn, "there is no general consensus concerning the use of register and related terms such as genre" (p. 21), which makes "genre literature a complicated body of scholarship to understand" (Hyon, 1996, p. 693). Several scholars endorse one of the two overlapping concepts and disregard the other, for instance Bhatia (2002), Biber (1988), Bunton (2002), Love (2002), Samraj (2002), and Swales (1990, 2004) apply the term genre solely, while Biber (1995), Biber et al. (1999), Bruthiaux (1994, 1996), Conrad (2001), Ferguson (1983), Hymes (1984), Heath and Langman (1994), and Ure (1982) prefer register over genre. Similarly to the Swalesian concept of genre, the notion of register defined by Biber et al. (1998) relies on non-linguistic or situational characteristics. The Biberian register, 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), also emphasizes the notion of a specific need of communication. In accordance with the Swalesian term genre, the Biberian concept of register groups discourse items on the basis of situational characteristics rather than focusing on the immediate surface similarities of their language use. Although the Biberian definition of register uses different distinguishing elements (such as purpose, topic, setting, interactiveness, and mode) than the Swalesian one of genre (where the idea of a shared set of communicative purposes appears), underlying scheme of the two is the same: it is the situation in common that connects and classifies discourse items. Both approaches treat the situational characters and not the linguistic phenomena to be of primary importance since "linguistic differences can be derived from situational differences" (Biber & Conrad, 2009,

p.9.) but not the other way round. Considering the obvious differences in the purpose and topic of the two types of texts under investigation in the present study, the term register can also be applied to them when making a differentiation between them. That is to say, the two sets of texts, EFL reading materials and biology chapters, belong to different registers in the Biberian sense and as such, their "identifying markers of language structure and language use differ from the language of other communicative situations" (Biber & Finegan, 1994a, p. 20). According to Halliday, this is exactly the reason why registers can be studied analytically, claiming that clusters of "associated features have a greater than random tendency to co-occur" in a register (1988, p. 162). In more general terms, Biber notes that all discourse analysts working in the field of ESP uncover "specialized registers in English" (1998, p. 157), which implies that each and every ESP field forms a different register.

2.3.2 Genre and register as different approaches

Although the terms genre and register are typically used synonymously, covering similar notions in a parallel manner, a clear distinction has been made between them lately. It was Biber and Conrad (2009) who recently separated the two overlapping concepts distinctively by treating them as two different *approaches* of text analysis. In their terminology, the genre approach examines rhetorical organisations and linguistic characteristics that structure whole texts. Such generic features might occur in the text only once or in strictly limited number, for instance the abstract of a research article, the title or the subheadings of a chapter in a textbook. For this reason, studies in the genre approach investigate complete texts instead of analysing a collection of excerpts. Examining texts from a different point of view, the register approach has a focal point of words and grammatical features that are frequently present in representative excerpts of numerous texts. Within the frame of the register approach, the analysis is regularly based on the collection of excerpts of

texts instead of relying on complete, full texts. The present study investigates the characteristic features of eight complete texts of a biology book against twelve full texts of a general English course book. The comparative analysis of entire, complete texts might suggest that the current study follows the genre approach. However, the nature of the investigation is more in harmony with the register perspective as it relies essentially on statistical methods of determining frequencies when discovering various prevalent characteristic features of the biology register. As the research lies in line with the register approach, the term register is used when referring to different text varieties of English in the dissertation rather than that of genre. This decision does not mean the rejection of the importance of the Swalesian emphasis on shared set of communicative purposes; however, aims at consistency through keeping the opaqueness of the various terminologies at a minimum throughout the paper.

2.4 Brief overview of the different ways of text analysis

With regard to text analysis, the diversity of the existing theories are not restricted to the register and genre approaches, though. To see why the present research adopts the latest version of register approach, let us now have a brief overview of the accomplishments of unique approaches and different theoretical perspectives to text analyses, inasmuch as the various ways of analysing texts in the last six decades, the period during which applying text analysing methods became well-established in ESP research. The present overview includes methods which primarily focus on texts; however, ones that are more ethnographic than text analytic in their approaches are not considered here. In this fashion, the New Rhetoric School (or North American School as it is also called) is not discussed, since its orientation principally concerns investigating the context in which the given text is used with the objective of revealing attitudes, values, and beliefs about the text user communities.

2.4.1 Register analysis

The register approach holds that communicative situations predetermine the choice of language use to a great extent. This is the reason why one can find the right words in the right place to convey the intended message (Pickett, 1986). The register perspective postulates that core linguistic features are "commonly used in association with the communicative purposes and situational context of the texts" (Biber & Conrad, 2009, p. 2). Presuming the fact that some linguistic features are more typical in certain communicative situations than in others, the register perspective aims to identify the pervasive linguistic characteristics, typical lexical and grammatical features in a variety. Pervasive linguistic features are not exclusively unique of a given register, they might occur in any other variety; however, they are "much more common in the target register" (Biber & Conrad, 2009, p. 6). Since it is the extent of pervasiveness of linguistic features that is analysed, the register perspective applies mathematical calculations and statistical methods of determining the frequency of certain linguistic items in a set of texts. Besides computing frequencies of lexical and grammatical items, the register approach combines numerical analysis with the examination of the situation of language use. In this way, the fingerprinting of a register consists of the exploration of three major components: the situational context where the texts stem from, the linguistic features whose pervasiveness is determined through statistical accounts, and the functional relationship between these two elements. The functional analysis of the characteristic linguistic features in a register description is possible due to the fact that linguistic features tend to occur in a register when they are "particularly well-suited to the purposes and situational context of the register" (Biber & Conrad, 2009, p. 6). Thus the third component of a register analysis attempts to interpret why certain linguistic features are more abundant in a register than in other contexts. Disclosing functional relationships between linguistic choices and situational contexts is "at the heart of studying register variation" (Biber & Conrad, 2009,

p. 10). In the frame of the register approach it is indispensable to try to explain why pervasive items, for example in the case of near synonyms or roughly equivalent grammatical structures, are applied in the given register.

As a rule, single lexical or grammatical features fail to characterize registers. Rather, it is a set of linguistic features whose level of pervasiveness in the given variety illuminates the typical language use of the texts, as early researchers (Ervin-Tripp, 1972; Hymes, 1974) in sociolinguistics have shown. Accordingly, register analysts discover the functional use of batches of prevailing linguistic items instead of examining specific, isolated linguistic markers. Biber et al. (1998) emphasize the necessity of investigating a group of wide-ranging linguistic features since it is not common for a register to be identified and well-described by the presence of a solitary linguistic feature. On the contrary, sets of several linguistic features tend to describe different registers, it is the frequency of various linguistic patterns that depicts the distinctiveness of a register. The exception to the rule of exploring multi-features is the attempt to identify register markers. These unique linguistic features are fixed expressions or "distinctive linguistic constructions that do not occur in other registers" (Biber & Conrad, 2009, p. 53). A register marker is so genuinely typical of a variety that it immediately reveals the communicative situation where it is naturally applied. Hearing for instance the fixed expression, 'Mind the gap,' one instantaneously identifies that the auditory warning was played at one of the tube stations in London, and the speaker is directly identified as the recorded announcer of the public transport company. Clearly distinctive register markers are infrequent, therefore groups of register features are investigated, instead.

Register analysis is a comparative approach by nature. To claim that the prevalence of any recurring linguistic item is a distinguishing feature of a given register, its frequency needs

to be compared to that appearing in another variety. Average frequencies without comparison across registers mean little, practically it is impossible to give a meaningful description of the distinctiveness of a register using figures without comparing these values to those of other registers. For register analyses to be effective, the data of pervasive linguistic items need to be compared to an adequate basis.

Shortly after the birth of the register approach in the 1960s, its popularity declined among ESP language analysts, dramatically fewer register studies appeared in the 1970s. There might be different reasons why the approach was not widely used. The register perspective has been criticized for being too simplistic since it fails to deal with any characteristics of the text beyond the sentence level (DeMarco, 1986). A relatively homogenous register that shows little variety among its users, for instance the language use of air traffic controllers, can be mapped effectively through describing its typical lexis and grammar. However, more complex ones with greater freedom of lexical and grammatical choices on the part of the language user are more difficult to be depicted through frequency accounts, moreover, the predictive value of these accounts is less reliable. This suggests that in the case of analysing more complex registers additional variables should be introduced. Another problematic point about register analysis voiced by DeMarco (1986) lies in the nature of the method of investigating texts on a linear, word-by-word or sentence-by-sentence basis. It is implied that such linearity results in losing global meaning when overemphasising the parts. Additionally, register analyses based on calculating pervasiveness made authentic representations of what language learners wishing to acquire a specific register use need to know, still there was some discrepancy when applying this knowledge in the compilation of teaching materials. Exposing students directly to the most typical discrete elements of a register did not enable them to handle communicative situations effectively, where pragmatic

knowledge is also required. This complaint was voiced by Selinker et al. (1976) when they claimed that students tended not to understand "the total meaning of the EST discourse even when they understand all the words in each sentence" (p. 82). However, the use of a research method, corpus-based register study in particular, does not strictly entail that language teaching and learning should rely on decontextualized methods (Coxhead, 2000). Despite the above mentioned weaknesses, the register perspective did not come to its end in the course of ESP history. Its revival is the benefit of the rapid advancement of computer technology in the 1980s. Computerized register analysis, which is less demanding to carry out than manual text examinations, is prone to be more reliable, besides, its scope of investigation can be wider-ranging and thus it can encompass greater complexity.

Register analysis has been applied in various academic and professional fields. Among the numerous foci of examining the typical language patterns of different communicative situations, sports announcer talk (Ferguson, 1983; Reaser, 2003), engineering English (Verantola, 1984), note-taking (Janda, 1985), academic prose (Biber, 1988), newspaper, radio and other media registers (Bell, 1991; Biber et al., 1999), personal ads (Bruthiaux, 1994), coaching (Heath & Langman, 1994), classified ads (Bruthiaux, 1996), abstracts of research articles (Connor, 1996; Flowerdew, 2002; Hyland & Tse, 2005), research articles (Conrad, 1996; Hyland, 1998a), textbooks (Conrad, 1996; Hyland, 1998b), scientific prose (Atkinson, 1999; Conrad & Biber, 2001), medical guidebooks (Vilha, 1999), internet registers (Crystal, 2001; Gains, 1999; Herring & Paolillo, 2006), student essays (Hyland, 2002a), PhD dissertations (Hyland & Tse, 2004; Paltridge, 2002), computer-based instant messaging (Fox et al., 2007; Thurlow, 2003), middle English medical texts (Taavitsainen & Pahta, 2004), university lectures (Biber 2006a; Biber et al., 2007; Csomay, 2005), news in

tabloids (Bednarek, 2006), dating chats (del-Teso-Craviotto, 2006), office conversation (Koester, 2006) were discovered through giving register analytical attention to them.

2.4.2 Systemic Functional Linguistics

Systemic Functional Linguistics (SFL) shows parallels with register analysis to the extent that both types of text analysis direct their attention to working out the probability of functional components in a text. The Australian-based discourse analysis, known in the United States as the Sydney School (Hyland, 2002b), also intends to find connections between language functions (governed by situational and social factors) and language use. Halliday et al. (1964) observed that there are "differences in the type of language selected as appropriate to different types of situation" (p. 87). The framework of SFL, appears to be dissimilar from that of the register approach, though. SF theory, based on Halliday's work, treats language as a social semiotic, a means people use to achieve their purposes by expressing meanings in context. Differently from the previous approach, SFL builds upon the idea that language is primarily a systematic resource, which appears in specific communicative situations. Thus the guiding principle, according to SFL, in describing language use is exploring a system, rather than structure. The theory aims to "uncover the general principles which govern the variation in situation types, so that we can begin to understand what situational factors determine what linguistic features" (Halliday, 1978, p. 32). Besides, SFL holds that the language used to express any meaning is implied by the context, therefore language use cannot be described without exploring its context culture. In the frame of SFL, language is regarded as a semiotic potential, which view results in describing language us as an account of choice. By way of using system networks, SF linguists map language analyses by creating diagrams of the choices language users might make in a situation to convey certain message. The choices available are subject to the context in which the language is used. Linguistic choices can be

described on different levels, SF theory deals typically with the semantic, the phonological, and the lexico-grammatical strata of language use, of which the latter includes the investigation of syntax, lexicon and morphology. Within the strictly contoured threefold foci, however, SF theory provides freedom for the analyst to uncover how language is manoeuvred to make meaning. It is the researcher who determines which aspect of language is relevant to be highlighted in a given register description, based on his argumentation or intuition of which patterns are more likely to co-occur in the register under investigation than in another. Whichever aspect the researchers decides to explore, SF theory maintains that the unit of analysis should be the text since the functional meaning is realized in no smaller unit than the text. The analysis of the threefold smaller units (semantics, phonology and lexico-grammar) are viewed from the standpoint of their extent of supporting the entirety of the text. Halliday (1985a) argues that "for a linguist to describe language without accounting for text is sterile; to describe text without relating to language is vacuous" (p. 10).

Due to the primary emphasis on system in function within the framework of SFL, the approach makes a clear theoretical distinction between the concepts of register and genre in its terminology. Martin (1985) clarifies that the two terms refer to two distinctly different semiotic planes. Genre is not considered to be a product (e.g., sets of texts), but it is thought to be a social process in which participants use language in a highly foreseeable sequential structure within the given culture in order to achieve their communicative purposes. In this sense, genres are assumed to be conventionally organized texts. More precisely, genre is the short form for the more sophisticated term "genre-specific semantic potential" (Halliday & Hasan, 1985b, p. 108), which is "tied closely to considerations of ideology and power" (Lee, 2001, p. 42). Following this line of thought, some researchers (Christie, 1992; Cope & Kalantzis, 1993; Johns, 1995) support the importance of genre instruction. In their view it is a

way by which students become empowered with linguistic resources for social success, a tool through which even nonmainstream groups of marginalized students (i.e., Aboriginal students in Australia, the home country of SFL) could gain access to a greater social power due to becoming more able at handling texts (Feez, 2001; Macken-Horarik, 2002; Martin, 2000). In contrast, registers are believed to be the expression plane of genres, thus the concept of genre encompasses that of register, as Eggins and Martin (1997) claim, a genre goes "above and beyond" (p. 243) a register. Typical linguistic choices across different genres are expressed in different registers, which are "recognizable as a particular selection of words and structures" (Halliday, 1978, p. 110), since "each speaker has a range of varieties and chooses between them at different times" (Halliday et al, 1964, p. 77). At the same time, Halliday (1978) warns that "instead of characterizing a register largely by its lexico-grammatical properties, we shall suggest a more abstract definition in semantic terms" (p. 110), in SFL registers are to be defined in terms of meaning. This is what Halliday (1978) underlines when stating that "register is the set of meanings, the configuration of semantic patterns that are typically drawn upon under specified conditions, along with the words and structures that are used in the realization of these meanings" (p. 23). Besides, the importance of the broader social context is also stressed in SFL, as Halliday (1978) maintains that "a register can be defined as the configuration of semantic resources that the member of a culture typically associates with a situation type" (p. 31).

Linguistic choices made in a social context are viewed in SFL as resulting from three contextual variables of register, called field, tenor and mode. Among these situational parameters of variation, field means the topic of the communicative event, tenor denotes the participants in the communication, their social roles and power relationship, while mode refers to organization and the aspects of the channel of communication. In Halliday's (1978)

wording, “register is determined by what is taking place, who is taking part, and what part the language is playing” (p. 31). Respectively, the following metafunctions can be assigned to these contextual variables: ideational to field, interpersonal to tenor, and textual to mode. Pragmatically speaking, ideational semantics (field) contains the propositional content, interpersonal semantics (tenor) is concerned with exchange structures, speech-functions, ways of expressing attitude, etc., while textual semantics (mode) involves elements of how the text is structured as a message, such as theme-structures, given-new. Throughout the course of SFL history, an extensive theoretical framework has been developed with these concepts (Halliday, 1985c, 1989; Martin, 1985, 1997, 2001a; Matthiessen, 1993).

By the application of SFL, a wide range of professional and academic registers have been explored, among them written sports commentaries (Ghadessy, 1988b), science articles (Ghadessy, 1993b; Tognini-Bonelli & Camiciotti, 2005; Vande Kopple, 1998), news reporting (Ghadessy, 1993b), internet-based registers (Herring, 1996), business letters (Ghadessy, 1993b), classroom discourse (Christie, 2002), and popular science articles (MacDonald, 2005).

2.4.3 Genre analysis

In line with SFL and register analysis, genre analysis also underlines the importance of situational context when analysing texts. The approach maintains that genres primarily develop within social formations (Kamberelis, 1995) thus genre analysis involves providing descriptions of communicative purposes and context in which a text variety arose. Although ESP scholars (Bhatia 1993; Flowerdew, 1993; Hopkins & Dudley-Evans, 1988; Thompson, 1994; Weissberg, 1993) working within the framework of the genre approach agree on the need to specify these purposes and the context, Hyon (1996) warns that many of them pay

disproportionately much attention to "detailing the formal characteristics of genres while focusing less on the specialized functions of texts and their surround social context" (p. 695). Similarly to SFL, the genre perspective does not fail to recognize social relationships. The ways in which social relationships are codified in language use form the basis of generic exploration of text varieties. It holds true to such an extent that Kress and Hodge (1979) point out the fact that one tends to identify the conventional aspect of a communicative event as a distinctive genre. The social structures of discourse communities produce disciplinary communication, which relies on their own built-in system of rules. Genres are kept alive and in circulation through the social practices of a discourse community, as Giddens (1979) points out.

Despite these similarities, the linguistic analysis of the genre approach contrasts with that of the register perspective by aiming at identifying conventional structures used in the entirety of the text instead of finding pervasive linguistic features. The genre approach tends to discover the conventional ways of language use in the genre, for example the beginning or ending of business letters. Focusing on the rhetorical elements that organize a text, the genre approach is characterized by top-down analysis, "where the starting point is the macrostructure of the text with a focus on larger units of text rather than sentence-level, lexico-grammatical patterning" (Flowerdew, 2005, p. 324). The target of genre analyses is to unveil the linguistic repertoire of structuring texts from a particular genre and to clarify for what communicative purposes they are applied. This vantage point is in stark contrast with the view of the register analysis, which relies on bottom-up descriptions starting out from smaller units of lexical and grammatical features limited by the sentence level. Genre markers, or distinctive expressions and devices that give a structural flow to the text are explored in the genre approach. These formulaic and typically once-occurring genre-marking expressions can

be found at a particular location of the text, such as ‘To be continued’ at the end of the episode of a series. Through describing the typical structuring phrases and expressions at various places of the text, the genre approach exposes the otherwise covert macrostructure of the text. When discovering the macrostructure of a text or a specific part of a text, the genre approach makes extensive use of the Swalesian move structure analysis, which “classifies segments of text according to their prototypical communicative purpose for a particular genre” (Flowerdew, 2005, p. 323). The Swalesian moves are divisions of the text, which are further subdivided into steps; for example the genre of introduction to a scientific article typically follows the moves of the CARS model, whose starting point is the text’s communicative purpose, that is, Creating A Research Space (CARS) for the new piece of work. In the model, each move contains specific information, which is systematically divided into steps through which the communicative purpose is reached. Move structure analysis collects syntactic and lexical features that are characteristically used in the steps and moves. In finding conventional structures and explaining their communicative functions, genre analysis does not aim to map out the myriad of different possible ways of expressing a message, in comparison with SFL, but focuses on the comparatively small set of codifications that have become typical and conventionalized in the genre.

Genre-based pedagogy has typically focused on written texts and made use of genre studies at writing classes (Hyon, 1996). The instruction of the results of genre analyses, how and why linguistic conventions are used for particular rhetorical effects, in second language writing courses is not without debates. Form-focused model introducing instruction has its advocates and opponents. Genre researchers (Gosden, 1992; Love, 1991; Miller, 1984; Swales, 1981, 1990) hold that conventionalized forms are typical means by which information is dispersed in a discourse community with shared interests. In their view, teaching genre

markers and discussing textual organization is of great importance since through developing students' awareness of the communicative purposes of generic typifications learners become more able participants of the genre community and can better control the organizational and stylistic features of texts. Not all scholars believe, however that employing generic knowledge in the service of language education is beneficial. Some challengers of the approach (Fahnestock, 1993; Freedman, 1993; Martin et al., 1987; Raimes, 1991; Reid, 1987; Threadgold, 1988; Zamel, 1984) assign more importance to the individual originality of the writer and to the process of writing itself, and put lesser emphasis on the specific elements of genre and organization. This, however, does not mean the complete ignorance of generic elements in second language instruction, genre markers are still advised to be addressed in the phase of rewriting, with a secondary importance compared to the verbalization of the message of the writer. More ardent opponents of genre-based instruction (Berkenkotter & Huckin, 1993; Dias, 1994; Freedman, 1993; Freedman & Medway, 1994) argue that the use of the conventions of generic knowledge in social context cannot be taught explicitly, it is a skill acquired tacitly through enculturation as students become active participants of the disciplinary community. Other scholars (Freedman, 1993; Williams & Colomb, 1993) claim that genre instruction has serious negative impacts on genres themselves as teaching textual rules to future writers acts in favour of rigidifying writing conventions.

Applying genre analytical methods, the language use of text organizing elements have been uncovered in numerous academic and professional fields. Among these are research articles (Biber et al., 2007; Marco, 2000; Swales, 1981), research article introductions (Gledhill, 2000; Samraj 2002a; Stotesbury, 2003; Swales, 1990), grant proposals (Connor, 1996; Connor & Mauranen, 1999; Swales, 1990), business faxes (Louhiala-Salminen, 1999), research abstracts (Salager-Meyer, 1990), popularized medical research reports (Nwogu,

1991), sales letters (Bhatia, 1993), university lectures (Thompson, 1994), fundraising discourse (Bhatia, 1998), promotional genres (Connor & Mauren, 1999), property transaction reports (Kong, 2006), academic e-mails (Gains, 1999), job application letters (Connor et al., 2002; Henry & Roseberry, 2001; Upton & Connor, 2001), editorial letters (Flowerdew & Dudley-Evans, 2002); direct mail letters from organisations (Upton, 2002), PhD dissertations (Swales, 2004), PhD conclusion chapters (Bunton, 2005).

With the aim of comparing and contrasting the above three different approaches of text analysis, a quick overview of their similarities and differences are summarized in Table 1.

Characteristics	Register Analysis	Systemic Functional Linguistics	Genre Analysis
Length of the text(s)	• various samples of text excepts or • complete text(s)	complete text(s)	complete text(s)
Linguistic focus	lexico-grammatical feature(s)	• semantics • phonology • lexico-grammar	• conventional expressions • rhetorical organization • textual organization
The rate of occurrence	frequent items	frequent items	typically once-occurring, in a particular place in the text
The method of analysis	bottom-up	bottom-up	top-down
The scope of explanation	the features are functionally connected to the situational context of the variety	according to field-tenor-mode	how language features conform to the culturally expected way of constructing texts belonging to the variety

Table 1 A contrastive overview of three different theories of text analysis: register analysis, systemic functional linguistics and genre analysis

2.4.4 Corpus linguistics

Relying on the advancements of computer technology in the new millennium, more advanced register studies are carried out within the framework of corpus linguistics (Grabe & Kaplan, 2006) using sophisticated methods of data analysis. Corpus-based register analyses can be classified as the latest, most modern current of the register perspective, at times partly overlapping with that of the genre approach (Biber & Conrad, 2009).

A corpus is a pool of samples of naturally occurring language, either written or spoken texts, which is stored by electronic means (Hunston, 2006) and is computer-readable through linguistic software (Stubbs, 2004). Small corpora tend to be highly specified (Stubbs, 2004), while large ones contain millions or even hundreds of millions of running words. Small or large, corpora embrace either complete texts or longer extracts from texts. The samples in the corpus represent a variety of language specifically designed for linguistic analysis, which makes the corpus homogenous to some extent. The careful selection of texts in a corpus embodies a broad and balanced sample of a register.

According to Biber, Conrad and Reppen (1998) the essential characteristics of corpus-based analysis can be summarized as follows:

"It is empirical, analysing 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 for analysis; it makes extensive use of computers for analysis, using both automatic and interactive techniques; it depends on both quantitative and qualitative analytical techniques" (p. 4).

Observing the collection of characteristic traits of corpus-based analysis, it is evident that computational linguist count on mechanized procedures of data analysis rather than relying on numerical methods carried out manually. As O'Keffee and McCarthy (2010) point

out it was not a linguistic paradigm change that stimulated corpus linguistics but the rapid boost in computer technology. Computerized analyses have several advantages of over manual ones. Dedicated software examinations provide undeniably more consistent analyses than manual investigations, which inherently might contain a number of errors due to miscalculations. Besides, computer programmes allow for the identification of complex patterns of language use in a large collection of texts, which could not have been dealt with by hand. Studying a large corpus is beneficial as it "reduces the burden that is placed on individual text or on intuitions" (Hyland, 2000, p. 138), characteristic in the case of manual investigations. Computerized studies also have the benefit to allow for exploring grammar, lexis and semantics interacting in an intertwined way, as Tognini-Bonelli (2001) suggests that grammar is not an abstract system underlying language but the same layer of language as lexis.

Corpus linguistics has been widely and prolifically applied in teaching English as a second language, the use of corpora became indispensable for lexicographers, grammarian, and teaching materials compilers. Numberless dictionaries (CIDE, 1995; COBUILD, 1995; LDOCE, 1995; OALD, 1995), grammars of English (Biber et al., 1999; COBUILD, 1990; Francis, Hunstor, & Manning, 1996) and teaching materials (Bernadini, 2000; Johns, 1991; Lewis, 1998) rely on patterns of language use detected in electronic corpora. Additionally, the databases of corpus linguistics have seeped into language courses directly when advanced learners are exposed to searching online corpora themselves. Corpus-based data-driven discoveries are favoured as they heighten students' awareness of language use through carrying out their own guided observations (Bernardini, 2004; Hunston, 2002; Willis, 2003).

The nature of computer-assisted corpus studies is inherently quantitative. Computational linguists search for and record repeated events in texts, with the aim to reveal which linguistic features recur frequently. The presupposition behind corpus linguistics is that high frequencies of language use cannot possible be due to blind chance. Thus recurrent words, phrases, collocations, phrasal schemas, multi-word units, grammar choices and semantic preferences are tapped with the aid of computer software programmes. Owing to its quantitative characteristics, criticism of the deficiency of deeper explanation has been levelled against corpus linguistics. Hunston (2006) voiced disapproval of the narrow scope of corpus linguistics by claiming that it does not allow more than the observation of quantity and fails to "expand the explanatory power of linguistic theory" (p. 234). Such worries tend not to be valid about corpus linguistics, as raw data are not believed to be the end of computer-assisted research. Frequencies are applied as "springboards to more qualitative study" (Hyland, 2000. p. 141), simple counts and other sophisticated quantitative patterns form the basis for describing wide-ranging similarities and differences in the language use of particular communities. The same expectation is stressed by Biber and Conrad (2009), who emphasise that "the quantitative and computational aspects of corpus analysis do not lessen the need for functional interpretations in register studies" (p. 74). Numerical findings are presented to increase our knowledge about the register under investigation and the quantitative data are primarily used to propose a "viable candidate explanation of underlying communicative purposes and interactional practices" (Hyland, 2000. p. 138). Corpus linguistics has also attracted criticism for being atomized in the sense that it analyses corpus data in a bottom-up approach, typical trait of the register perspective. However, the bottom-up manner of analysis does not inevitably result in atomized finding as long as the computational linguist works with whole texts. The particular place of an item in a text can suitably reveal the overall rhetorical structure of the text (Flowerdew, 2005). Moreover, corpus-based studies (Connor et al., 2002;

Thompson 2000; Upton, 2002) might follow the genre approach by examining the interaction of lexico-grammatical features in move structures rather than observing truncated linguistic items or sentences. Another focus of the criticisms fired against corpus linguistics is its decontextualized nature from the texts' original communicative setting. Allegedly, corpus linguistics fails to take into account the contextual features of the text, which can be particularly problematic when exploring pragmatic features obtainable from the socio-cultural context. According to Hunston (2002), the lack of visual and social context makes the interpretation of corpus data unmanageable. Widdowson (1998, 2002) goes as far as refusing to treat corpus data as samples of authentic language due to the absence of familiarity with the communicative context where the data were produced. Decontextualisation may be a valid argument against analyses using large corpora where contextual features are not easily retraceable from the text alone. However, where the analyser is also the compiler, there is plenty of knowledge about the broader socio-cultural setting in which the texts were born. Studying the institutional setting in which the registers is used gives a better understanding to the implicit conventions followed by participants in the communicative situation. Situated research, locating texts in contexts, allows for examining registers with the considerations of the views of those who use the texts. In Flowerdew's (2005) words, the "compiler-cum-analyst can act as a kind of mediating ethnographic specialist informant" when interpreting corpus data.

Within the frame of corpus linguistics, Biber (1988) developed a comprehensive approach to describing patterns in register variations, the computerized method of multidimensional analysis (MDA). This method aims at finding underlying linguistic parameters, or dimensions, as well as specifying linguistic parallels and dissimilarities among registers along the dimensions identified. MDA relies on multivariate statistical techniques,

especially factor analysis, to investigate the co-occurrences of linguistic features when discovering systematic patterns of variation among registers. Characteristically of the register approach, the complexity of linguistic features to be explored is emphasised in the process of obtaining adequate descriptions of registers. In line with the early recognition of the importance of linguistic co-occurrences (Brown & Fraser, 1979), MDA follows Biber's (1988) observation that statistically significant features tend to cluster in texts. Consequently, the method finds it misleading to focus on specific, isolated linguistic features and does not investigate single parameters individually. The MDA perspective aims at finding groups of linguistic features that co-occur in registers. To map registers onto the groups of linguistic markers or dimensions, texts in the corpora are automatically analysed, or tagged, for linguistic features representing numerous major grammatical and functional characteristics. After carrying out the quantitative, numerical analyses, the frequent (positive) and rare (negative) features in the dimensions detected through factor analysis are interpreted in terms of communicative functions. The qualitative analysis specifies how the language features with statistically significant values are well-suited to the communicative purposes of the text.

Using the numerical and functionally interpretive method of MDA, numerous registers have been explored, among them are letters (Biber & Finegan, 1989), medical academic prose (Atkinson, 1992), 18th century authors across different registers (Biber & Finegan, 1994b), spoken and written registers in variety of languages (Biber & Finegan, 1994a; Biber, 1995), research articles and textbooks (Conrad, 1996), internet-based and computer-mediated communication (Herring, 1996), newspapers (Biber & Finegan, 1997), scientific prose (Atkinson, 1999), newspapers, magazine articles and medical writing (Vilha, 1999), disciplinary texts (Conrad, 2001), historical and contemporary registers (Conrad & Biber,

2001), speech and writing in the university (Biber et al., 2002; Biber, 2006), radio and TV sports commentary (Reaser, 2003), and university classroom talk (Csomay, 2005).

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

and needs great expertise. The reason for its laborious nature is the fact that it requires the sophisticated statistical analysis of a large number of linguistic features to identify the groups of features that co-occur in the text with high frequency. In their research to show that MDA fails to be irreplaceable with a less arduous tool for register analysis, Xiao and McEnery (2005) undertook a keyword analysis to compare three registers (conversation, speech, and academic prose) by producing wordlists of corpus files extracted from large American corpora (the Santa Barbara Corpus of Spoken American English, the Corpus of Professional Spoken American English, and the Freiburg-Brown corpus of American English), which were compared to a reference corpus, the British National Corpus to detect and compile those words whose frequency differed from the reference corpus either by being unusually high (positive keywords) or extremely low (negative keywords). The results of their study confirmed that applying the keyword approach is capable of producing comparable results to the MDA approach and can identify important register patterns, despite creating a less nuanced comparative contrast of registers.

Considering the benefits and limitations of the reasonably different ways of text analyses reviewed above, the present research follows the register perspective to gain insights for ESL and biology ESP teachers into the characteristic linguistic features of the biology texts 10th grade bilingual students process in their studies. It is the register approach which gives space for identifying a great number of linguistic features simultaneously that characterise a pool of texts, thus can serve ESL and biology ESP teachers with substantial information on the possible foci their teaching materials should be directed at. SFL explores various potential choices of different language use, which uncovers relevant information for language users who actively produce the register under examination and wish to make a well-informed choice of the most appropriate language use in the given situation. In contrast, the

register approach discovers the very text in hand, producing knowledge that is more applicable for secondary students, who are in need of processing biology texts but are not expected to create textbook chapters during their studies. Likewise, the genre approach unveils structural rhetorical information about texts which is indispensable for writers of similar texts to become accepted members of their discourse communities, while less informative for those who aim at understanding and processing but not at creating generically similar texts. Furthermore, the register perspective is comparative by its nature, which is advantageous in the present case since it allows a direct comparison of the two registers students meet in the course of their studies. Following the register approach, the set of 10th grade biology texts can be compared to that of the previous year's, 9th grade general English texts, which have been specially chosen in the bilingual program to prepare students for their academic studies in English the following year. In this manner, the comparison generates relevant linguistic information for general English teachers instructing in the 9th grade and biology ESP teachers alike.

2.5 Research on secondary-level biology textbooks

In the field of text analysis, textbook analysis is a prolific subfield. In the last century, infinite types of textbooks, the primary teaching aid in the classroom (Khine, 2013), have been examined from a plethora of different viewpoints depending on the intent of the analysis.

2.5.1 Analyses of biology textbooks in secondary education

In order to support my research with relevant methods and results in the field of textbook analysis, biology textbook analyses were sought which investigate textbooks for secondary school students written in English where the point of analysis is the language use of the textbook. Analyses with different approaches were not considered since the results of their

views are of little consequence to ESP instructors, let alone to ESL teachers, whose task is to prepare bilingual students for studying academic subjects in English and not that of making a well-founded decision on choosing which biology textbook to use in the 10th grade. Thus the specific area of concern for my investigation did not include studies exploring the following otherwise typical textbook analysing foci. First, no literature discovering the content of the subject matter was compiled, information was not collected about the ways of evaluating whether the biology topics are developed in a sequentially logical order starting from the simple and moving to the more complex; if the level of the teaching material corresponds to the target readers' age; if the technical terminology is appropriate to the subject; whether the examples are suitable; what kind of ideology the textbook follows etc. Secondly, analyses assessing the quality of the tasks following the texts were not considered, no data was compiled about how different biology textbooks phrase questions to grasp the main message of the text; if the tasks are written in an understandable manner; if the instructions are preceded by clear instructions, etc. Thirdly, knowledge about textbook illustrations was not gathered either, studies investigating whether the diagrams in the textbook are appropriately labelled; if they are easily drawable by students; whether the diagrams reflect the local environment; if the illustrations are attractive, etc. were left uncollected. Finally, in a similar manner, investigations into the layout of textbooks were not amassed, data was not collected about appropriate printing styles; quality of paper; binding and durability of textbooks etc. Whether the different approaches of textbook analyses yield applicable information for the current research is summarized in Table 2.

Point of analysis	Applicability in the present research
Content of the subject matter	not relevant
Quality of related tasks	not relevant
Textbook illustrations	not relevant
Layout of the textbook	not relevant
Language use of the textbook	relevant

Table 2 An overview of the different foci of secondary school biology textbook analyses and their relevance to the present research

2.5.2 Analyses of biology texts and secondary textbooks

With the aim of finding already existing descriptions of the language use of secondary school biology texts in English that can advance my understanding of the register, a systematic search of the literature was conducted. All the existing volumes of the specialized journals *English for Specific Purposes* and *Journal of English for Academic Purposes* were thoroughly checked in order to collect relevant articles. In this manner, the search embraced the publications of 29 years of the previous and 14 years of the latter journal, including 105 and 44 issues respectively. Besides, several academic search engines were applied to find appropriate studies. Academic Search Complete, Education Research Complete, ERIC, JSTOR, Science Direct, Taylor & Francis, and Web of Science were run with the combination of the following keywords: register analysis, discourse analysis, genre analysis, biology text, biology textbook, science textbook, secondary school textbook. Although a huge body of literature was consulted, to the best of my knowledge no study was carried out in the secondary context that would take the linguistic perspective when analysing a biology textbook to the benefit of the ESP or the general English teacher.

The reason for not finding studies that analyse English language biology textbooks for secondary school students from a language teaching point of view might be the fact that most of the research into academic English focuses on teaching and learning in a post-secondary environment, as Hamp-Lyons and Thompson (2006) pointed out, the milieu of secondary education is given much lower priority. The growing interest in exploring academic English use in secondary school setting has been recognized, the embodiment of this increasing attention is the edition of a special issue on the topic launched by the Journal of English for Academic Purposes in 2006. It is important to note that not even this special issue led to a heightened awareness of the register of biology textbooks, as none of the articles in the issue dealt with the academic language use of the discipline of biology. The explanation of the editors (Hamp-Lyons & Thompson, 2006) for the dearth of investigation in the secondary school context is the fact that secondary school teachers do not appear to find the time or lack the incentive to carry out research and write articles. Simply phrased, secondary educators prefer to be teachers than researchers.

As the description of the register of biology textbooks for secondary students from the point of view of English teaching is not represented in the field, the scope of my search of existing empirical studies was widened in order to discover analyses that are relevant in describing the corpus of scientific texts for secondary students.

Type of scientific reading	Type of text	Medium	Level
Subject	educational chapters	textbooks	elementary and secondary (precollege)
Discipline	research articles	specialized journals	tertiary (college, postgraduate)
General interest	popularizing articles	magazines	common knowledge of public science, general understanding

Table 3 Three types of scientific reading according to Widdowson (1981)

Widdowson (1981) distinguished three types of scientific reading, such as science as a subject, sciences as a discipline and science as a topic of general interest, summarized in Table 3. The first category, science as a subject, can be read through textbooks; in the present case the subject is biology, and the textbook is the corpus under scrutiny. The second, more academic category refers to research articles published in learned journals of a discipline taught at tertiary level, a medium which provides specialized knowledge in the field. Finally, the third type includes the less academic register of journalistic texts and popularizing magazine articles introducing science to the general public. With regard to the extent how much education each type of scientific reading requires, the second one is the most specialized and academic, while the last one is the least academic, which communicates information to people with no specific training in sciences.

In the broadening of the horizon of the present research, a principled decision was made to include the analyses of texts of the second category but none of the third. Based on the similarities between popular literature and secondary-level textbooks, science textbook analyst Shapiro (2012) believes that precollege science education can be best understood if textbooks are considered to be a form of popularization. His arguments point out that secondary students form an audience of non-scientists to whom science is presented in a less or even non-technical language by authors who try to relate their academic profession to the experience of society. Despite this view, the opening of the angle of my investigation shifted in favour of the more academic type of scientific reading rather than the third type of more popular texts. The reason for my choice of extending the investigation to include studies on the more academic type of scientific reading lies in the fact that the English teaching program of the bilingual secondary school run in the 9th grade was developed based on the educational idea that from the 10th grade on students need to be able to process academically challenging

texts. Considering register characteristics of less academic and more journalistic science texts might have resulted in arriving at features which are not comparable with the supposedly highly scientific corpus 10th graders are expected to read according to the developers of the bilingual program of the school. From the various science disciplines in the tertiary setting, biology and its closely related sister-discipline, medicine were screened for relevant register descriptions. Besides, studies analysing textbooks of other subjects than biology were also consulted to gain information about different possible ways of analysing their language use. With the help of opening up the scope of my search for relevant methods in text analysis, the notions of readability indices, lexical density and thematisation devices proved to be applicable in yielding useful linguistic data for ESP and EFL teachers. Their usefulness in providing linguistic data on textbooks for EFL and ESP teachers is discussed in the following chapters.

2.6 Readability indices

Readability is an attribute of the degree of clarity of a text, it is the quality of how effortlessly one can understand a reading material due to its style of writing (Klare, 1963). The extent to which a given group of people find a text gripping and comprehensible shows the text's readability (McLaughlin, 1969). The ease of reading words and sentences can be expressed through readability measures (Hargis et al., 1998), mathematical formulae have been developed and improved to predict the level of difficulty of texts by objectively calculable means since the 1920s. The primary goal of classical readability research was to develop a technique that matches reading materials with the abilities of the target reading public, students and adults alike. The alarmingly poor results of adult literacy surveys collected in the USA in the 1930s urged the creation of graded texts for adults. Aiming at the solution of such practical problems, readability indices were elaborated to assess the level of

difficulty of a text objectively and with consistency on the basis of counting the length of sentences and words as well as the number of syllables and characters by the use of mathematical formulas that require no great sophistication to apply (DuBay, 2004; Harrison & Baker, 1998). The principle behind using these variables is the generally believed view that longer sentences are more difficult to read than short ones, and short words are easier to process than longer ones (Jacobson, 1998; Ulusoy, 2006). Some of the more than hundred readability indices (Fry, 2002) assign a grade level to the figure they count, which grade refers to the number of formal education needed to understand the text without difficulty. The use of readability indices has been mandatory in some documentation contracts, particularly for the US government (Harrison & Baker, 1998). Besides that, there are numerous other fields where readability formulae are regularly applied, such as textbook publishing, journalism (e.g., United Press, Associated Press, National Geographic tend to use them), giving information about bank loans or insurance policies, issuing rental agreements and property-purchase contracts, in the business of health insurance, sharing technical-training materials in the army or research on literacy. Despite the fact that readability indices are widely used in different areas of life, their value has been questioned.

Criticism levelled at readability formulae emphasize the notion that the numerical analysis is restricted to the surface features of texts (Bruce et al., 1981; Duffy & Kabance, 1981; Harrison & Baker, 1998; Kern, 1979; Manzo, 1970; Redish & Selzer, 1985; Schriver 2000; Selzer, 1981; Valdes et al., 1984). Undoubtedly, readability indices ignore many features in their assessing of a text's level of difficulty, for instance content, conceptual load, coherence and organization of the text or the reader itself is not taken into account in any ways. When calculating the difficulty of a text, readability indices fail to differentiate according to the reader's reading style, the reader's reasons for reading, the reader's prior

subject knowledge, the reader's beliefs about the word, the reader's motivation, etc. Furthermore, the excessive use of short words results in the low ranking of the text, implying a quick and easy understanding of it; however, short but unfamiliar words can make a text unduly problematic to process. Deeper syntactic structures of a text are also disregarded by these indices (Valdes et al., 1984). The perplexing discrepancy that different formulae give different results when applied to the very same text has also been pointed out (Duffy, 1985). One of the reasons behind the difference is the fact that there is no common baseline of understanding defined for the various formulae. There is no common zero point for reading success (Klare, 1982), and different formulae draw the minimum of understanding at rather different levels of accurate understanding indicated by the score of correct answers on a reading test, for example. Secondly, formulae do not use the same variables and applying different variables results in different figures, whose suggested grade levels cannot be truly compared. For these reasons Duffy (1985) argues that even if readability formulae rank texts successfully, they cannot predict the actual number of years of education required from the reader to process the text with ease. Duffy's cautious view strengthens Klare's (1968) vigilant stance claiming that scores of readability indices are better treated as rough guides of probability statements than highly precise values.

One might wonder how readability indices have continued to be applied in such wide-ranging fields for nearly a century long if they fail to grab what is beyond the surface of a text. The sceptical view (Harrison & Bakker, 1998) formulates that the absence of any better alternative keeps traditional readability formulae in use. Research, however, clearly shows that despite all the limitations of readability indices, "these formulas are correlated with conceptual properties of text" (Kintsch & Miller, 1981, p. 222). Data gained from the formulae of readability variables produce precise predictions of text difficulty as measured by

comprehension tests, such as multiple-choice reading tests, ranking or cloze procedures (Bormuth, 1966; Carver, 1990; Chall, 1984; Chall & Conrad, 1991; Chall & Dale, 1995; Coupland 1978; Davison, 1984; Klare, 1984; Maxwell, 1978). In addition, the figures resulting from readability formulae correlate with another measure of text difficulty, the density of propositions in a text (Kemper, 1983; Kintsch & Miller, 1981). The high predictive power of readability indices, according to DuBay (2004), can be explained by the measurement's ability to reveal the "skeleton of a text" (p. 57), whereas tone, content, organization, coherence and design do nothing else than "flesh out" (p. 57) the skeleton. Although readability indices have proved to have a high predictive power with regard to rating the level of difficulty of a text, and their validity has been verified for long (Chall & Dale, 1995; Fry, 1963; Klare, 1963, 1974, 1976, 1984; Smith & Kincaid, 1970), readability researchers recommend that formulae are most trustworthily used in conjunction with other methods of grading rather than applying numeric analysis mechanically. (Chall & Conrad, 1991; Dale, 1967; Dolch, 1939; Fry, 1988; Gilliland, 1972; Gunning, 1952; Klare & Buck, 1954).

2.7 Lexical Density

When measuring the level of difficulty of a text, one of the possible ways to complement the results gained from applying readability indices is to gauge the lexical density of the text, which displays no direct connection with the grade levels suggested by readability formulae (Vinh et al., 2013). The numerical measure of lexical density as a readability predicator was introduced and validated by Ure (1971), who recognised its importance as a measurable factor of complexity of discourse. In order to measure lexical density, or the proportion of lexical items in the texts, it is indispensable to divide words into two main categories. Words can be classified as either content words (e.g., *read, academic,*

text, easily) or grammatical words (such as *it, and, in, the*). Content words are also referred to as lexical ones, while grammatical words as structural or function words. The main difference between content words and function words is the quality of the previous having autonomous meaning even in isolation. Besides, to the category of content words new members can be added (Le, Yue & Le, 2011), they are part of an open system rather than a closed set (Halliday, 1985a), which typically embraces nouns, verbs, adjectives and often also adverbs. In a different fashion, words that do not have clear lexical properties but possess a more grammatical-syntactic function are grouped as grammatical words. Grammatical or function words belong to a closed system containing a fixed set of items, where new members are impossible to be added. This set comprises articles, pronouns, most prepositions, conjunctions, finite verbs, interjections, and count verbs (Cindy & James, 2007; Halliday, 1985a; Johansson, 2008). Adverbs are on the borderline between lexical and grammatical items: modal adverbs form closed sets, while adverbs derived from adjectives can be freely added to an open set. The proportion of the number of lexical items per the number of orthographic words of a text expressed as a percentage defines lexical density in Ure's (1971) terminology. Later Ure (1977) added that "all words have grammatical values" (p. 207), emphasizing the fact that lexical words contain some grammatical information as well, thus they are not in sharp contrast with function words. In this way Ure (1971) draws attention to the fact that lexical density is a "part : whole relation" (p. 207). Ure's (1971) formula was refined by Halliday (1985a), who determined lexical density in relation to the clausal richness of the text. Halliday's (1985a) definition counts lexical density to be "the number of lexical items as a ratio of the number of clauses" (p. 67). It is noteworthy to remark that Halliday (1985a) uses the term 'item' rather than 'word' when distinguishing lexical items from grammatical ones, since he reasons that lexical items might contain more than one word. For instance, phrasal verbs such as *sit down, take up* or *call for* contain two words, a lexical word

and a preposition, however, Halliday (1985a) considers each to be one single lexical item. This is in contrast with Ure's (1971) classification, who treats phrasal verbs as consisting of two separate words, a lexical word, e.g., *sit, take* and *call* in the examples and a function word, such as *down, up* and *for*.

Applying the measure of lexical density to spoken and written texts, Ure (1971) and Halliday (1985a) have shown that written English contains significantly more lexical items in relation to grammatical items than spoken English (written English higher than 40% while spoken English under 40%). This implies that written discourse tends to have a higher level of information content for a given number of words than its spoken counterpart, which is expressed through the abundant use of content words. This means that investigating the lexical density of a text uncovers its level of information packaging (Johansson, 2008), a text containing a high proportion of lexical words conveys more information than a text with a high proportion of function words. To examine whether the rate of lexical density indicates the level of difficulty of a text, Harrison and Bakker (1998) designed a research project in which passages expressing the same information with different level of lexical density were to be ranked according to how demanding they were perceived to be understood by native participants. The research has demonstrated that lexical density clearly predicts the level of difficulty of a text as less lexically dense texts were consistently perceived as more readily understandable by the participants, whose preference for the less dense passages was markedly significant.

Relying on the measure of lexical density, various fields have been described and compared, among them are the language use of students (Cheryl, 1995; Narelle et al., 1994); textbooks (Vinh et al., 2013), spoken discourse (Reads & Nation, 2006; Nesi, 2005);

comparison of spoken and written language use (Camiciottoli, 2007; Halliday, 1985b; Johansson, 2008; Yu, 2007); various languages other than English (Henrichs, 2010; Laurén, 2002; Linnarud & Thoursie, 2008; Stegen, 2005, 2007).

In order to gain insight into the reading difficulty of the present corpus through determining the level of information packaging it contains, this research follows Ure's (1971) definition of lexical density by counting the proportion of lexical words in the texts. The proportion of lexical items in relation to the clausal abundance of the texts, as Halliday (1985a) developed the term, was not considered since the present study uncovers sentence complexity of the corpus separately in much greater detail. Besides, lexical density per clause has proved to be less informative than lexical density per text since the latter is independent of clause length (Stegen, 2007). In the case of borderline categories, adverbs and prepositions, however, the current empirical research relies on Halliday's (1985a) views. Non-grammaticalised adverbs, all adverbs derived from adjectives, were counted as lexical items since they form an open set; while grammatical, non-productive adverbs were classed as grammatical items. Phrasal verbs, containing one or two prepositions, were treated as one single lexical item since they tend to be non-separable in the process of reading and correct understanding of a text.

2.8 Sentence complexity: sentence length, packet length and syntactic structure

The information packed in a text, whose density can be examined through computing lexical density, is expressed through a series of sentences, some of the longer, others shorter. 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 his longitudinal study, Sherman (1893) noticed that the average sentence length of English prose shortened

dramatically over time. The length of sentences shrank from 50 words per sentence in the 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 extremely 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 suggesting 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 claiming that long sentences effectively broken up do not increase the level of reading difficulty of texts. Their research considered the length of packets, a mechanically modelled unit of a group of words between any syntactical punctuation marks (full-stop, comma, colon, semi-colon, exclamation mark, question mark, long dashes and parenthesis). The 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. 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.

The present research does not fail to notice the limited nature of the information gained through computing sentence length, thus it also examines to what extent the sentences of the corpus are organized into easily recognizable packets. Furthermore, the current research also explores the level of complexity of syntactic structure of the register, as processing a string of simple sentences poses less serious challenges to the reader than comprehending a stretch of complex and compound sentences.

2.9. Textual metadiscourse

Besides gaining information about sentence complexity (sentence length, packet length, readability and syntactic structure), the level of difficulty of a text can also be predicted by investigating how explicitly it structures its claims to convey the author's message. The indicator that reveals to what extent a text expresses overtly the directions of the intended logical flow of its ideas is called metadiscourse. The term *meta*discourse might suggest discourse *about* discourse, however, it covers a slightly different concept in the literature of text analysis. Hyland (1998b) defines metadiscourse with three focal points as the "aspects of the text which explicitly organise the discourse, engage the audience and signal the writer's attitude" (p. 437). The implicit idea behind Hyland's (1998b) definition is that authors never write without intentions, their aim is to convince the reader or to guide their audiences' understandings of texts. In order to reach this aim, authors need to manipulate rhetorical features effectively, which requires a sense of audience. Merely informing the reader is not persuasive enough, writers need to "weave discourse into fabrics that others perceive as true" (Harris, 1991, p. 289) through applying appropriate metadiscourse. Metadiscourse is the author's response to the readers' potential difficulties of interpreting the text or to the anticipated objections to the content of the text. In addition to directing their readers' comprehension of the text, authors also intend to clarify their own stances towards the content of their texts (Hyland, 1998b, 2005). All these intentions are expressed through metadiscourse, whose function is the creation of contact at different levels. According to Hyland (2000, 2005), metadiscourse creates a threefold contact: one between the parts of the text, a contact between the writer and the reader, and also a contact between the text and the writer.

In the sense of creating contact between parts of the text, or making the text coherent, metadiscourse includes the non-propositional parts of the text (Crismore at al., 1993). This is the focal point to which many text analysts constrict the meaning of metadiscourse in their research (Ädel, 2006; Beauvais, 1989; Mauranen, 1993a; Valero-Garces, 1996). Hyland (1998b), in contrast, distinguishes coherence as a subtype of metadiscourse, called *textual* metadiscourse, and differentiates it from *interpersonal* metadiscourse, which involves the author's intentions to show his relation to the text and to affect the audience. Textual metadiscourse is the collection of linguistic devices that supports the formation of a convincingly cohesive and coherent text (Vande Kopple, 1985) through linking the otherwise discrete propositions to each other by explicating e.g., "topic shifts, signalling sequences, cross-referencing, connecting ideas, previewing material" (Hyland & Tse, 2004, p. 158). Among textual metadiscourse items, Hyland (1998b) differentiates five broad functional categories as follows: logical connectives (mainly conjunctions, which express semantic relations between main clauses), frame markers (which explicitly refer to the stages or boundaries of the schematic structure of the text), endophoric markers (which make reference to other parts of the same text), evidentials (which refer to information in other texts, that is, exophoric markers that creates intertextuality), and code glosses (which provide additional information to help the reader grasp the meanings of ideational material through explanation, examples or paraphrasing).

Interpersonal metadiscourse, on the other hand, establishes a relationship between the author and the reader. It contains the linguistic strategies which self-reflectively refer to the writer and to an imagined reader of the text (Hyland & Tse, 2004). Language is never purely used to express bare pieces of information about the world (Hyland, 2010), the writer has a tendency to support the text with interpersonal metadiscourse cues to ensure that the reader

receives the propositional content of the text the intended way. Interpersonal metadiscourse is primarily interactional, it is essentially applied to facilitate communication between the writer and the reader. The level of interaction the author wishes to establish with the readers is expressed through interpersonal metadiscourse (Vande Kopple, 1985). It also reveals the writer's persona, a "created personality put forth in the act of communicating" (Campbell, 1975, p. 394). How personal the personality of the persona becomes is determined not only by the personal intentions of the writer but it is also strictly affected by the conventions of the discourse community the writer belongs to; it cannot be ignored that writing is a social engagement. Thoughtfully added to the text, metadiscourse does more than change a juiceless text into a reader-friendly, coherent piece of writing that conveys the author's personality (Vande Kopple, 1985), it also expresses the writer's "credibility, audience-sensitivity, and relationship to the message" (Hyland, 2000, p. 157) by ways of evaluation and appraisal (Hunston & Thompson, 2001; Martin, 2001b). Using devices of metadiscourse, writers endeavour to create a credible representation of both themselves and of their writing (Hyland, 2010).

Ironically, it was one of the founding-fathers, the most influential representative of metadiscoursal analysis, Hyland, who criticized the approach for its misleading dichotomy, which separates two single, distinct functions of metadiscourse. To overcome the confusion implied by the unhelpful juxtaposition of textual versus interpersonal metadiscourse, Hyland and Tse (2004) proposed a new model for the analysis of metadiscourse arguing against Hyland's earlier developed straightforward differentiation (1998b, 2000). The new model (2004) advocates the overriding significance of interpersonal metadiscourse, and disagrees with distinguishing textual metadiscourse as purely and categorically different from interpersonal metadiscourse. It treats the hitherto discrete textual aspect of metadiscourse as

the expression of interpersonal relations. Textual metadiscourse is claimed to be primarily interpersonal on grounds that cohesion is a textual function that is "intrinsic to language and exists to construe both propositional and interpersonal aspects into a linear whole" (Hyland & Tse, 2004, p. 162). The role of linguistic devices of textual metadiscourse is regarded to be central in articulating propositional and interpersonal functions rather than expressing mere links between clauses. Textual metadiscourse, even in the case of straightforward cohesive markers, is reasoned to be problematic when not treated as an expression plane of interpersonal discourse. The explanation for this is that besides establishing cohesion, conjunctions also have the function of indicating the writer's comprehension of the relationship between individual ideas. Understood this way, conjunctions "not only glue the text together, but extend, elaborate, or enhance propositional meanings" (Hyland & Tse, 2004, p. 162). In this new framework, conjunctions are seen as interactionally triggered linguistic features, which essentially support the elucidation of the flow of interpersonal orientations. For instance, the use of mitigating counterclaims in an argument or introducing alternative statements, while raising doubts helps the author direct the reader's understanding of the text towards accepting his claims and reasoning against other possible counterarguments at the same time. Indicating connections overtly between propositional elements in a text reflects the author's awareness of the audience and his or her self-awareness of the writer as a persona. The level of the presence or the absence of the threefold referencing in a text (connecting to the text, to the reader, and to the writer) shows the writer's sensitivity to the context of the piece of writing, i.e., to its discourse community. The extent and the quality of textual metadiscourse is thus the result of the writer's decisions to underline certain relationships, and to organize the text in such a way that appropriately guides the readers' understandings and increases the audience's acceptance of the argumentation through making them cognizant of the writer's preferred interpretation. By negating the possibility of splitting metadiscourse into

two distinct types, textual and interpersonal, Hyland and Tse (2004) also voices a more general doubt claiming that imposing disconnected categories on the "fluidity of actual language use inevitably conceals its multifunctionality" (p. 175).

Although Hyland and Tse (2004) complain about the fact that metadiscourse, understood in either a mono- or a dichotomous way, has never become a key analytical focus in the research of discourse, the social distinctiveness of several disciplinary communities have been detected and described from this angle. The most noticeable studies have identified the characteristic metadiscoursal features of casual conversations (Schiffrin, 1980), school textbooks (Crismore, 1989; Hyland, 1999), science popularizations (Crismore & Farnsworth, 1990), undergraduate textbooks (Hyland, 2000), postgraduate dissertations (Bunton, 1999; Swales, 1990), company annual reports (Hyland, 1998c), cultural differences in texts (Crismore et al., 1993; Mauranen, 1993b; Valero-Garces, 1996), medieval medical writing (Taavitsainen, 1999), scientific discourse in the 17th century (Atkinson, 1999), student writing (Cheng & Steffensen, 1996; Intraprawat & Steffensen, 1995; Jalilifar & Alipour, 2007), academic writing (Thompson, 2001) and EAP classroom discourse (Lee & Subtirelu, 2015).

The present research welcomes Hyland's (2004) recognition of the importance of interpersonal functions of textual metadiscourse, the attempt to conceptualize textual metadiscourse more broadly as an interpersonal feature of communication and to avoid a simplistic duality. However, the current study intends to adapt Hyland's earlier dichotomous model of metadiscourse (1998b, 2000) for different theoretical and practical reasons. Although textual metadiscourse undoubtedly backs interpersonal metadiscourse, in secondary education it is beneficial to separate the two overlapping ideas clearly and distinctively. On the one hand, secondary-level textbooks differ from research articles and specialized tertiary

textbooks. Contrary to tertiary publications, textbooks written for secondary students address the dissemination of knowledge that has already been accepted in the discipline, consequently their writers have little pressure of making their own persona appear credible to the discourse community or of making their findings and speculations seem convincing. In the case of secondary textbooks, no crucial emphasis is placed on the overt expression of the author's stance, rather pre-college level science texts intend to communicate impartial, objective, canonized pieces of knowledge (Shapiro, 2012). On the other hand, the audience in the secondary setting expects to read science textbooks for their propositional content primarily; the credibility of the author and the writer's attitude to the information conveyed in the texts are of little significance to them, if any. The principal concern for secondary students processing science texts is to form a clear understanding of the topics discussed. Thus the essential function of textual metadiscourse in this environment is to guide the readers' comprehension effectively, which is the main characteristic feature of the formerly (1998b, 2000) distinguished textual metadiscourse. Since sentence initial frames (Gosden, 1992) or differently termed by Halliday (1985a) as marked non-subject themes help the reader manipulate the thematic components of a text, the intimate familiarity with the devices of textual metadiscourse is vitally important when handling texts. In order to enhance students' understanding of texts, raising their awareness of the text-organizing function of textual metadiscourse is indispensable. With the aim to be able to draw pedagogical implications that finally support the reader of secondary science texts, the present research focuses on the connectedness of texts through examining the overt elements of textual metadiscourse when considering textual metadiscourse in the threefold interconnectedness of text, writer, and reader.

Keeping the merits of the above introduced various possible ways of texts analyses in mind, let us now turn our attention to the biology corpus under examination and the environment where the secondary-level texts are applied.

3 Methods

The following chapter lays out the design of the current research. First the setting where the two corpora are used is depicted through the detailed introduction of the bilingual immersion programme of the secondary school. This is followed by the characterization of the participants: it is explicated what bilingualism means in the present research environment. Then the two corpora are described through presenting the textbooks of their origin, their size and the principles and practice of their compilation. Subsequently, the methods of data collection and data analysis are shown. The reliability of the text-analytical instrument and the validity of the data it produces are examined against each linguistic variable of the POTAI, that is, each component of the POTAI is addressed separately in this chapter from these points of view. However, the validity of the instrument and the research in general was also guaranteed by taking three steps in the research.

There is little doubt that any instrument can yield data which describe a register exhaustively. However comprehensive an instrument is, there is necessarily a touch of restriction and limitedness to it since perspectives of gathering information are boundless, while all instruments are finite. In order to ensure that the present instrument effectively measures what it is intended to measure, that is, it has a high rate of construct validity, a panel of experts (several professors teaching in the language pedagogy PhD programme at ELTE) were sought to provide their expert judgment on the components of the POTAI to what extent the instrument's range of the linguistic elements contains features which are indispensable to be investigated in order to obtain register-specific data relevant for ESL teaching. Thus the sampling validity of the linguistic phenomena of the POTAI, i.e., the fact that the content of the instrument is adequately sampled (Carmines & Zeller, 1991), was confirmed.

Secondly, a high rate of external validity (or the extent to which the results of the study are discernible whether they are transferable to other research environments or not) was assured by providing contextual information about the corpora through thick, detailed descriptions of the setting, the learning environment, the source of the compiled corpora, and the participants.

Thirdly, the overall content validity of the research was also increased by setting clear goals and well-defined objectives. As a result, the theoretically and pedagogically motivated umbrella questions, which outlined the main focus of the research (see Chapter 1 on p.4), were specified into more tightly focused Research Questions (abbreviated as RQ in the study).

The channelling, focus-providing umbrella questions of the research are as follows:

A) By what means, relevant to English as a second language teaching, is it possible to describe the dominant register features of the biology texts used at an English-Hungarian bilingual secondary school?
B) From a linguistic point of view, to what extent do the general English reading texts 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?

RQ1: Is the pedagogically oriented text-analytical instrument (POTAI) capable of providing reliable and valid data concerning the dominant register features of the biology texts used in the instruction of 10th grade students in a bilingual secondary school?

RQ2: What linguistic features characterise the biology texts used by 10th grade students at an English-Hungarian bilingual secondary school in the first academic term (BIOCOR) in comparison with the B2-level general English texts read by 9th grade students at the school (REFCOR) with regard to the texts'

1) lexis

    1.1 frequently used words

    1.2 keyness

    1.3 lexical density;

2) grammatical phenomena;

3) sentence complexity

    3.1 sentence length

    3.2 packet length

    3.3 readability indices

    3.4 syntactic structure; and

4) textual metadiscourse?

RQ3: Based on the findings of the research, what pedagogical recommendations may be formulated for educators (e.g., ESL teachers instructing in the 9th grade language preparatory year of the bilingual programme and biology ESP teachers) selecting finely-tuned texts which are at the appropriate level for preparing secondary students for their biology studies in English?

The methods of investigation which were applied in the current research are summarized and displayed in Table 4.

Research Question		Methods of data collection	Methods of data analysis	Chapter in the dissertation
Umbrella question	Sub-question			
A	RQ1	Qualitative Qualitative Quantitative	• Developing the instruments: seeking expert judgement • Developing the instrument: interview studies with teachers instructing at the bilingual school • Statistical methods	Section 3.3 Methods of data collection and data analysis
B	RQ2	Quantitative and Qualitative[1]	Corpus-based register analysis	Chapter 4 Results and discussion
B	RQ3	Qualitative	Interpreting the data (constant comparative method)	Chapter 5 Pedagogical implications

Table 4 The methods of investigation used in the study

The methods of data collection (whether its nature is qualitative or quantitative), along with the different methods of data analysis used when addressing the Research Questions are clarified here. With this methodological overview in mind, let us now turn to the pedagogical context of two corpora: the bilingual immersion programme of the secondary school.

3.1 The setting: 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

[1] Describing a register based on characteristic frequency accounts might seem to imply the application of quantitative methods only. However, register analysis, as Biber and Conrad (2009) emphasize, applies qualitative methods at the same time, since there is a definite attempt to provide an interpretation as to why particular linguistic features are more abundant in one register than in another.

that newly introduced a five-year-long bilingual programme in various languages, which meant an absolutely new genre 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 was 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 her indispensable support to promote the progress of the English-Hungarian bilingual programme at the 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.) who are capable to study, be engaged in research and work in English. Besides, it was also important to educate students in a framework that refuses to accept the singularity of a cultural vision 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 while the others in Hungarian, and minimum 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 language is used as the medium of instruction in the case of Hungarian language and literature, IT, chemistry, music, and physical education. At the time of the foundation of the programme, bilingual education meant the instruction of at least five academic subjects in the target language. However, 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 the time of data collection (2011-2013), the same academic subjects were taught in English as the ones 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 5. The name immersion was given to this educational model by Lambert and Tucker (1972), whose metaphor 'language bath' emphasised the intensive presence of the second language in the educational environment into which the students are immersed.

	Features characterizing immersion programmes (Swain & 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 5 Characteristic features of immersion programmes (Swain & Johnson, 1997)

As it can be seen from Table 5, nearly all the features of an immersion programme are present in the bilingual education programme of the school. English as a second language is used as a medium of instruction in the majority of the academic subjects. The curriculum runs parallel to that used in the first language in non-bilingual classes in the case of all subjects.

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. 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. Students in the bilingual programme use English as a second language mostly in the classroom, their exposure to English is confined to studying curricular content and conversing with the native teachers at times. Despite English being 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. 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. 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). Finally, classroom culture is also similar to that of typical immersion programmes, namely, 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 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 strict sense (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, characteristic of bilingualism (Bloomfield, 1933). The macro-context in which the students were brought up is not different either. English language is not significantly present in the 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 that 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 was 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. The notion of bilingualism can be understood in a less strict sense, however. On the other end of the spectrum of interpretations, 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, p. 7). In this particular educational context, I apply the term bilingual 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 constantly being exposed to a linguistic environment different from their first language. When this process is induced through schooling, I apply 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 first one 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 megacorpus containing one million running words. Being trained 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 its aspect of sociolinguistic competence, somewhat lags 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 introduces to check similar skills. Students have never been not 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 her new principles of bilingual education (Regulation No 26/1997). The regulations favoured numbering the academic years consecutively, thus the 'zero-year' became 9th grade and the following first year of the national academic secondary school programme became 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' was not in official use between 2011 and 2013, 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 study five core subjects, history, mathematics, physics, geography, and biology, in English the following year on for four years of the Hungarian academic secondary school programme. Respectively, there is one history ESP, one mathematics ESP, one physics ESP, and one geography ESP provided a week for 9th graders. 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 English language, the aim of the preparatory year is to enhance students' knowledge of English to reach a firm B2 level. In harmony 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," and they are also predicted to "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, p. 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 a upper intermediate level mock Cambridge Exam, the First Certificate in English (FCE), administered by the school[2].

[2] Despite the fact that the data collection of the present research occurred between 2011 and 2013, the mock FCE exams were still structured according to the composition of the examination in practice before the 2008 modifications. The immersion programme's principle behind not updating the mock exams was a practical reason: the majority of the resources (practice books and test samples) available at the school were published before 2008.

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, where complete beginners are instructed separately from students with higher levels of English. The decisive factor whether to arrange students in 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 numerous complete beginners tend to be grouped homogenously, as long as they attend the same class. Between 2011 and 2013, the period when data were collected at the school, 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 minimum 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 The corpus

3.2.1 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 from the past that describe registers. Table 6 shows the seven major situational characteristics the framework consists of and the 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 6 The situational parameters of the biology textbook (Roberts, 1981) according to the framework of situational characteristics of a text (Biber &Conrad, 2009)

The biology textbook was produced by a readily identifiable single author, M. B. V. 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 GCSE exam in biology, and apparently who study biology in English. The addressees are definitely multiple individuals, however, their exact number or further identification cannot be more closely specified, and thus the large 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 intended to be contacted 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, which is less due to the inequality of social power, but it is more the result of the complete lack of bidirectional encounters among the participants in any forms of communication. The shared background knowledge of the participants covers a specialist field, the addressor communicates information purely within the field of biology. The addresses are not expected to have expert background in the field, their novice status is connected to their lower social power in the field.

The physical channel of the register is writing, its specific medium of communication is the printed form. Although the textbook exists in an electronic form too, 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, the level of unintendedness is extremely low, if any. The editor of the text is the single addressor himself. However, instances of revision to 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 texts, the circumstances are completely controlled, too. 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, thus 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, p. 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. That is, the textbook is not designed to impart newly tested hypotheses but it 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 for the addressees. As a result, the epistemic stance of the biology textbook expresses a high rate of certainty, the information it imparts leaves little space for doubts. 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.

3.2.2 The size of the corpus

The number of words in the collection of the biology texts in the present research project does not come close to a million, the approximate benchmark of a large corpus; thus, considering its size, it can be treated as a mini-corpus (Biber & 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, in the case of examining a specific area of the language a small corpus is advised to be compiled by various scholars for several different reasons. 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 & McCarthy, 2010, p.6). Besides the obvious convenience of a mini-corpus of being more manageable to handle than a large one (O'Keffee & McCarthy, 2010), there are several serious considerations for its compilation. 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 Beugrande, 2001). Additionally, all occurrences, including low-frequency items, can be examined, which is not possible in the case of a large corpus

(O'Keffee & McCarthy, 2010). The examination of items entails the possibility of establishing a close link between the corpus and the context (Biber & Conrad, 2009) since the language use is kept intact in the sense that the texts are not de-contextualised in the mini-corpus.

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

The above benefits of a mini-corpus in general conform accurately in a particular case as long as the corpus under investigation is representative of its register. Besides, representativeness is crucial from another perspective, namely, that of validity. The outcome of 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.

In order to make the biology corpus (hereafter referred to as the BIOCOR for short) representative of what the 10th grade bilingual students are expected to read and process in their first academic term, it was checked which biology texts exactly are assigned for them to read. 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 not to remember precisely what has been covered in class. Each of the five interviewees named the same eight chapters, which are listed in Table 7. To affirm 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. Through observing the electronic register, it was confirmed that the biology chapters compiled by the students was an exhaustive list. Next, the

eight chapters were typed in order to make them computer analysable, and first a word count was run. It can be stated that 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	1613
2	Classifying, naming and identifying	875
3	Amoeba and other protists	767
4	Bacteria	689
5	Viruses	777
6	The earthworm	517
7	Harmful protists	1085
8	Parasitic worms	698

Table 7 The BIOCOR: the eight chapters of the biology textbook (Roberts, 1981) and their lengths given in words

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. Where the application of the text analytical instrument produced data which are statistically generalizable, i.e., the description is not constrained to the very texts of the corpus, the broader sense of the term 'register' is explicitly indicated.

3.2.4 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 (referred to in the study in its acronym form as the REFCOR) against which the results of the corpus of the biology texts are compared was that the pool of general English texts should also contain approximately 7,000 words in total. 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 FCE reading exam the 9th graders take. Although the data for the present research were started to be

gathered in 2011, the four parts of the reading paper represent a former FCE version, the one before the 2008 modifications. The reason behind not choosing the most up-to-date version of the exam is that 9th grade students tackle to solve the previous version as their end-term exam (see Footnote 2). A complete FCE reading exam consists of about 2,000 – 2,500 words, thus 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 number of texts and, as much as possible, an equal number of words. Finally, twelve texts were selected from the general English course book the 9th graders use when preparing for their end-term FCE exam (Prodromou, 1998). The total length of the twelve general English texts measures 7,098 words. Table 8 displays the reference corpus of the general English texts, their lengths given in number of words, and the total length of each part of the exam is summed up 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 8 The REFCOR: the general English texts chosen from the 9th graders' FCE course book (Prodromou, 1998) and the lengths of the texts given in words

3.3 Methods of data collection and data analysis:

Linguistic variables of the POTAI

Since no single measure is capable of describing a register completely and of assessing its level of difficulty, a combination of various different measures is recommended (Biber, 1998). Based on the insights gained from the literature review (see Sections 2.4 – 2.9 on pp. 13-52), and the expert judgement of the specialists instructing in the language pedagogy PhD programme at ELTE, the POTAI finally contains the following main components: lexis, grammar, sentence complexity and textual metadiscourse.

3.3.1 Lexis

The most outstanding linguistic features along which registers differ from one another is considered to be vocabulary (Atkins, Clear, & Ostler, 1992; Biber, 1989; 1993; Sinclair, 1991). The lexical component of the POTAI offers register analysis of texts from three different points of view: frequently occurring words, keyness and lexical density.

Frequently occurring words of the BIOCOR were explored for several reasons. From a pedagogical point of view, words which are frequently applied are regarded as more useful for language learners to acquire than words which appear infrequently in a register (Nation, 2001; Biber, Conrad, & Reppen, 1994; West, 1953). Considering a different aspect, frequent words tend to be processed more quickly and understood better than ones which are used infrequently (Haberlandt & Graesser, 1985; Just & Carpenter, 1980). For this reason, frequent words might be claimed to be the ones that make a text easier to process, since their rapid or in some cases even automatic decoding increases the effectiveness of reading performance (Koda, 2005). Texts which contain a great proportion of high frequency words can be regarded as easier to process. Furthermore, frequent words assist higher level meaning

building (Crossley, Greenfield and McNamara, 2008). The more frequent a word is, the more likely it is to be process with a high degree of automaticity, which (besides increasing reading speed), frees working memory for higher level cognitive functions.

3.3.1.1 Frequently occurring words

After having compiled the biology and reference corpora, the hard copies of the texts were digitalised by use of keyboarding to carry out the following analyses with the expectation of describing the special language use of the register of biology texts as far as its frequent lexis is concerned. To find the typical lexis prevalent in English language biology texts written for secondary school students, the frequency of lexical words in the corpus of the biology texts was computer counted by using text analysing software program WordSmith version 5 (Scott, 2008). The frequency of grammar words was ignored in this part of the analysis, as it was examined through the measure of lexical density (see Section 3.3.1.3 on pp-72-84).

In order to find the most common lexical elements of the register, words of the same root were lemmatized by the software program so that it was the frequency of word families determined, not that of individual word forms. Lemmatization was considered to be crucial as it is more valuable for ESP teachers to possess knowledge about the frequency of word families than that about conjugated verb forms or different word formations when it comes to working out the lexis part of ESP syllabi. This is the practical reason why word lists for learners of English also tend to group words into families (West, 1953; Xue & Nation, 1984). Besides, compiling words in word families instead of listing isolated elements of different word forms was chosen for theoretical reasons too, namely, word families form a unit in the

mental lexicon (Bauer & Nation, 1993; Nagy et al., 1989). Lemmatization of the words allowed the following different word forms to be considered as one batch:

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

Compound words, however, were not joined in one batch. Respectively, 'flat' and 'flatworm,' 'stream' and 'streamlined' for example were computer counted separately. The reason for not lemmatizing compound words lies in the strong possibility that the parts of the compounds cover relatively distant meanings, for instance 'cow' and 'cowslip' or 'Mary' and 'marigold.' After lemmatization, the most common words in the biology corpus were listed in rank order, arranged and displayed in bands of frequency. Band 1 contains the most ubiquitous, most typical words in the text, the ones that appear minimum 30 times in the investigated chapters of the biology book, while Band 10 involves more unusual items, word families that occur only four times in the corpus. Table 9 shows how often items of particular bands appear in the register expressed both in the number of their raw occurrences and in percentages.

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

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

Table 9 The frequency bands in the BIOCOR

In each band the individual words and lemmas were manually sorted out into one of the following three categories: biology term, academic English lexis and general English item. First, the category of *biology terms* contains lexical items that carry a specific meaning within the context of biology, a meaning or a shade of meaning which is different from the everyday use of the word. A current dictionary of biology (Thain & Hickman, 2004) was applied as the baseline when determining whether a word should be labelled as biology term or if it is nothing else but a general English word that happens to be related to a certain biology topic. The entries of Thain & Hickman's biology dictionary (2004) were chosen to be the reference line since the dictionary claims to clarify the most essential concepts in biology for teachers and students alike. In this research, lexical items that appear as entries in the biology dictionary were labelled as biology terms. Within a word family, all the members of the lemmatized batch were checked in the dictionary, thus it was ensured that a lexical item was labelled as biology term irrespective of its word class. For example the noun 'reproduction'

appears as an entry in the biology dictionary; however, the verb ‘reproduce’ does not. In this case the lemmatized word family including the items ‘reproduce’, ‘reproduction’, ‘reproductive’ was marked as biology term. On the other hand, dictionary entries where a lexical item appears in conjunction with other words, that is, biology terms that contain more than one word, were not labelled as biology terms unless they appeared in the biology corpus with the exact same word combinations. For instance, the lexical item ‘body’ is not a separate entry in the biology dictionary, while ‘carotid body’ is. Consequently, the word ‘body’ was not labelled as biology term in the present analysis unless it was used in the biology texts in conjunction with the word ‘carotid.’ Biology terms that occur frequently in the corpus were gather in order to provide pedagogical implications for biology ESP and general English teachers.

The second category, *academic vocabulary* was assigned to those lexical items that appear on Coxhead’s (2000) list of academic vocabulary, a collection of 570 word families. Coxhead’s academic word list (AWL) was selected to be applied in the research since it is a systematic collection of academic English, a set of wide-ranging lexis typically used in the register of academic English, which was particularly compiled for pedagogical purposes. The AWL was gathered in order to provide insights for English teachers preparing students for their tertiary studies in English, that is, the aim of Coxhead’s collection of words is to show clearly what specific lexis is prevalent in academic texts. The AWL has proven to pinpoint the lexis that makes academic registers markedly different from other registers (Coxhead, 2000), thus it is a reliable instrument to find academic vocabulary in texts in English. The corpus in which the frequency of words was run by Coxhead (2000) embraces four subcorpora of the following faculty sections: arts, commerce, law, and science. Each of these faculty sections are further divided into seven subject areas. Biology is one of the subject areas of the science

sub-corpus, which allows using it as a baseline with a high-rate of construct validity in my research environment. In the AWL only those word families were involved that appeared in over half of the twenty-eight subject areas. Words that occurred in fewer than fifteen of the subject areas were labelled as narrow range words and were excluded. This principle ensured that the list could be used for any academic subject area, its coverage is not restricted to specific subjects. In the development of the list, frequency played a key role, word families that were used more than 100 times in the 3,500,000-word-long corpus were shortlisted. Basic vocabulary, words that are among the first 2,000 most frequently occurring words of English as compiled by West in his General Service List (1953), were not involved in the short list, since academic reading presupposes the learner's familiarity with basic vocabulary at tertiary level. From this aspect, AWL is advantageous to be used in my research environment since 10th grade students are also expected to be familiar with the most widely used words in general English. The similarity ensures a high-rate of criterion related validity for my instrument. Besides basic lexis, proper nouns, for example names of places and people, as well as Latin forms, such as *etc, i.e.*, were also removed from the short list of AWL. Finally, the list was organized into ten sublists based on the frequency of the particular word family. The sublists were numbered consecutively, where sublist one contains the most common academic words in the corpus, while sublist ten comprises less frequent academic lexis. The present research uses Coxhead's (2000) findings in order to see whether the biology texts assigned to 10th grade students in the bilingual secondary school are difficult to read for the fact that they contain a large number of academic lexical items.

Finally, words that belong neither to the category of biology terms nor to that of academic vocabulary were labelled as *general English* in my research. The prevalent lexical items within the general English category were collected and listed in order to help general

English teachers and biology ESP teachers gain insights into the nature of the general English lexis used in the biology textbook for secondary school students.

From a pedagogical point of view, the description of the lexical environments of the most frequent biology terms was treated as essential as our "knowledge of a word includes the fact that it co-occurs with certain other words" (Hoey, 2005, p. 8). The lexical environments of the biology terms appearing in the first three bands in the corpus were described by compiling the words that they go together with. The biology terms in the following bands (bands 4-10) can be found in Appendix A, however, the restrictions on the length of the dissertation curtailed the detailed description of their lexical environments. In order to look more deeply behind the quantitative results collected through frequency analysis, collocations were searched using the KWIC (key word in context) application of the same software (WordSmith version 5, Scott, 2008) within the range of the boundary of the sentence. Compiling all the word combinations with which the frequent biology terms are used in the corpus gives the possibility to gain pedagogical implications for biology ESP teachers working out biology ESP syllabi. The words that collocate with the frequent biology terms were sorted out according to their part of speech. To produce an easy-to-follow list, collocations were recorded in an alphabetic order, in their dictionary forms. That is, tenses in which the given verbs that go together with the biology terms were not kept, one can find for instance 'parasites make for John's liver' instead of 'parasites made for John's liver'. In a similar manner, modals that appear in the biology texts were not accounted here, thus 'viruses are released' appears in the description of the biology term's environment and not 'viruses may be released'. Finally, to keep the descriptive list as easy-to-grasp as possible, relative clauses used in the biology texts were also omitted, even if it resulted in a slight change of meaning. Minor changes of content information of the biology texts were not considered

crucial in the present analysis since the description of the environment of the biology terms is of lexical nature. In other words, the main focus of the lexical accounts is to tap the possible collocations used with the frequently applied biology terms, while the descriptions do not attempt to collect information in the field of biology. That is the reason why for example the phrase 'animals transmit parasites' is listed in the research instead of recording the defining relative clause 'animals which transmit parasites.'

3.3.1.2 Keyness

Although Biber's (1988) multidimensional analysis (MDA) has a long record of reliably uncovering linguistic patterns of registers, the present research follows a more novel analytical method which is considered to be a replacement of MDA (Tribble, 1999). The reason for choosing the keyword application of WordSmith program (Scott, 2008) instead of carrying out an MDA analysis on the BIOCOR is not simply due to the recentness of the previous software. The decision was also based on considering Xiao and McEnery's (2005) empirical research results. Their study proves that revealing keyness with WordSmith program is a method that provides comparable results to MDA since the new application can identify similar register patterns. Secondly, MDA uses seven dimensions (Biber, 2001) to explore the characteristic linguistic features of texts (see Appendix B on p. 240). Yet, most of the seven dimensions fail to appear to be utterly relevant considering the focus of the present research as neither the ESL teachers instructing 9th grade students in the bilingual programme nor ESP teachers would benefit directly from the linguistic data of these dimensions in their teaching practice. Thirdly, the multivariate statistical technique on which MDA is based is factor analysis, which is a sophisticated method that can be applied effectively on large corpora only. Sadly, factor analysis does not work on a small corpus (Csizér, personal

communication, 16th February, 2012), thus the current corpus of 14,000 running words cannot be investigated along the Biberian lines of factor analysis.

Keyness describes the distinguishing lexical characteristics of a register by comparing its language use to that of another register (Xiao & McEnery, 2005). The keyword application of WordSmith version 5 (Scott, 2008) extracts lexical items that are present in the register under examination, however are not typically used in the reference corpus. That is, keyness shows the lexical uniqueness of a corpus by compiling lexical items that make the register markedly different from another one. Inversely, the application also collects items that are underrepresented in the register compared to a baseline corpus This trait is labelled as negative keyness. Either positive or negative, keyness does not reveal lexical items that frequently or infrequently occur in one single register under investigation but ones which are characteristically different with respect to their frequencies when the register is compared to another register. This method ensures that lexical items which are not register specific, ones which occur with similar frequencies in both corpora, are not compiled.

Keyness is determined by statistical comparison carried out by keyword programs. A word is considered to be key if its frequency in the corpus when compared with its frequency in a reference corpus is such that the statistical probability as computed by the appropriate procedures described below is smaller than or equal to a p value of 1E-6.[3] To compute the keyness of an item, the WordSmith version 5 (Scott, 2008) calculates four values, which are consequently cross-tabulated. The four values include the raw frequency of the item in the corpus, the number of running words in the corpus, the raw frequency of the item in the

[3] 1E-6 is a standard scientific notation for the value of one times 10 to the power of -6, which equals one over 1million, or 0.000001.

reference corpus, and the number of running words in the reference corpus. The statistical procedure of finding key words includes the chi-square test of significance with Yates' correction for continuity to reduce the error in approximation. The test of keyness in the case of WordSmith program (Scott, 2008) relies on a log-likelihood test, Dunning's procedure (1993). The fact that Dunning's procedure is not based on the presupposition that data have a normal distribution in the text (McEnery et al., 2006) increases the instrument's reliability. The application of a log-likelihood test, disfavouring normal distribution, was especially important in my research environment, where the REFCOR does not contain considerably more running words than the target corpus but was compiled to be approximately of the same size as the BIOCOR. WordSmith version 5 (Scott, 2008) treats words which are not represented in the reference corpus as if they occurred 5.0E-324 times (that is $5.0x10^{-324}$) in the baseline corpus. To apply a keyword program that assigns such a small value to non-represented lexical items in the corpora was a decisive factor in the choice of the software. Without this slight modification, uncovering stark contrasts between the two registers would have been impossible since cross-tabulation with values of zero does not produce any meaningful result. An infinitesimally small number, however, allows for the handling of lexical items that do not occur in either of the two corpora, and due to the number's close-to-zero value, it does not affect the calculation materially. To ensure reliability, WordSmith version 5 (Scott, 2008) defines those items as key whose p value is smaller than or equal to 1E-6, that is 0.000001. The p-value shows the danger of being ungrounded when claiming relationships. Consequently, an extremely low p-value threshold increases reliability. In the present case the chance of erroneously listing words with similar frequency in the two corpora as key words is 0.00001%.

In order to arrive at data which are practically useable for ESL teachers instructing in the bilingual programme of the school and for biology ESP teachers alike, words of the same root were lemmatized by the keyword program before running the keyword application. The principles of lemmatizing isolated words into word families were the same as in the case of lemmatization when finding frequently occurring lexis, which is described in detail in Section 3.3.1.1 on pp. 73-79. After running the appropriate statistical procedures, the key words of the BIOCOR were listed by the software in an order of outstandingness. The computer-counted keyness values of the lemmatized items on the list reveal to what extent the frequency of the particular item is different when compared to that in the REFCOR. Subsequently, the key words were manually correlated to the most frequently occurring lexical items in the BIOCOR. Such a correlation was considered to be important in order to find out more in depth about the nature of the biology register. Next, with view of gaining a better understanding into the degree of the use of specific lexis in the register (biology terms and academic English) might make the biology texts difficult for 10^{th} graders to process, the key words were classified according to the three categories used and defined before (biology terms, academic English and general English; for the principles of differentiation see Section 3.3.1.1). Keeping the ESL and ESP teachers' focus in the foreground, the need to avoid teaching lexical items in an isolated manner was treated to be crucial. Thus the lexical environments of the biology key words were also described by running the KWIC concordancing application of the software, where the range of investigation was the sentence boundary. All the words that co-occur with the biology key words were compiled and organized according to their part of speech. To make the list straightforward, collocations were entered in the list in an alphabetical order. The lexical items which describe the environment of the key biology terms were recorded in their dictionary forms, which resulted in several changes of form and some of meaning. The previous section on the data analysis of

frequently used lexis (3.3.1.1 on pp. 73-79) has already explained what was altered and how. The same guiding principles were followed in this section, too. Finally, items with negative keyness in the BIOCOR were also collected, and their role in shaping the register of the biology textbook was investigated.

3.3.1.3 Lexical density

The level of difficulty of a text can be measured by its lexical density. The lexical density measure uncovers the text's level of information packaging (Johansson, 2008) through calculating the lexical complexity of the discourse. The present research follows Ure's (1971) formula of computing lexical density by counting the proportion of lexical words in the entirety of the text. The validity of the computation, i.e., the extent to which the formula measures what it is purported to measure, was first shown at the beginning of the 1970s (Halliday, 1985a; O'Loughlin, 1995; Ure, 1971). Ure demonstrated that counting lexical proportion with the formula when attempting to find register-specific traits, that is, typical features which distinguish one register from another, results in a measure which accurately assesses the degree of information-content typical in the register. The construct validity of the measure was also shown by Harrison and Bakker (1998), who pinpointed that lexical density predicts the level of difficulty of a text with distinct precision since the measure has a high correlation with the level of difficulty of processing a text.

In order to compute the lexical density of the BIOCOR and the REFCOR with Ure's (1971) formula, all the lexical items of the two corpora were organized into two categories: content / lexical words and grammar / function words (for the guiding principles of categorizing words see Section 2.7 on pp. 41-44). To find the number of lexical words in the BIOCOR, the part-of-speech tagging software designed and developed by UCREL

(University Centre for Computer Corpus Research on Language) at Lancaster University (http://ucrel.lancs.ac.uk/cgi-bin/claws7.pl) was applied. The software called constituent likelihood automatic word-tagging system version 7, abbreviated as CLAWS7, was chosen to be applied in the current study as its reliability is outstandingly high. The CLAWS7, which was also used to tag the 100 million words of the British National Corpus in 2012, can be characterized by a 96-97% rate of consistency; that is, its measuring procedure yields nearly the same result on any repeated trials. The computer program works by assigning a part-of-speech tag to each word of the text fed into it by categorizing it with either of the 137 part-of-speech labels (for detailed information about the CLAWS7 tag set see Appendix C). Once all the words of the two corpora were tagged (for an example sentence see Appendix D), the labels of the corpus annotation were organized into categories as either lexical or functional ones. To increase the internal validity of the research, labels whose categorization in terms of lexical-grammatical dichotomy is not straightforward (ones that might cover a lexical or a grammatical item depending on the item's function in the sentence), were marked as dubious and were revised manually in the corpora (see Table 10).

Part-of-speech codes	Meaning of the part-of-speech code (with examples)
IF	*for* (as preposition)
II	general preposition
IO	*of* (as preposition)
IW	*with, without* (as preposition)
RP	prep. adverb, particle (e.g., *about, in*)
RPK	prep. adverb, catenative (e.g., *be about to*)
RR	general adverb
VD0	*do*, base form (finite)
VDD	*did*
VDI	*do*, infinitive (*I may do, to do*)
VDZ	*does*
VH0	*have*, base form (finite)
VHD	*had* (past tense)
VHG	*having*
VHI	*have*, infinite
VHZ	*has*

Table 10 The dubious CLAWS7 labels that were manually revised in the corpora

The function of every single word with a dubious label was checked in the sentence in which it appears in order to decide with great accuracy whether it is a lexical or a grammatical token in the particular corpus. Accordingly, the words labelled with dubious categories were manually re-labelled by extending their two- or three-letter-long CLAWS7 codes with either an extra letter L, standing for lexical, or a letter F, short form for functional. To avoid corrupting the high-reliability rate of the computer program, the manual revision of the items in question was carried out by one of my colleagues, who was previously informed about the guiding principles (see Section 2.7 on pp. 41-44). Inter-rater reliability, the degree to which the two raters' decisions appeared to be consistent, proved to be outstandingly high, as none of the dubiously tagged items were found to fall into different categories by the two raters. Next, the number of items in each coded category was computer-counted and added up to find the sum total of the lexical words in the BIOCOR and that in the REFCOR. Finally, the ratio of lexical words in the entire corpus, expressed in percentages, was counted for both corpora separately, which were then compared and contrasted.

3.3.2 Grammatical components

The units and modules of all the three books which are used in the language preparatory year of the bilingual immersion programme (Cunningham & Moor, 2005; Falla & Davies, 2008; Prodromou, 1998) are built around various grammar points, thus 9th graders are given a thorough training in grammar and are also expected to master grammar profoundly at the B2 level. For this reason it was considered to be inevitable to involve the aspects of grammar in the POTAI.

3.3.2.1 Procedures of designing the grammatical component of the POTAI

3.3.2.1.1 Investigating grammatical features

The grammatical components of the POTAI were developed in five steps. First, the kinds of grammatical features that have already been studied in ESP register analysis were investigated. It can clearly be seen that various grammar items were examined individually in a systematic manner, such as nominal structures (de Haan, 1989; Geisler, 1995; Johansson, 1995; Verantola, 1984), negation (Tottie, 1991), apposition (Meyer, 1992), clefts (Collins, 1991), the passive voice (Granger, 1983), and infinitival complement clauses (Mair, 1990). The intention to uncover the distinctive grammatical features of registers in a comprehensive description was also attempted (Biber, 1998; for an exhaustive list of the items investigated by Biber see Appendix E). The subfield of describing registers along distinctive grammar features cannot be claimed to be untapped within the field of ESP, however, the existing frameworks of research fail to be compatible with the present research. The above listed high-validity frameworks lose some of their validity in the current research environment as their focal point is not that of the ESL teacher. To keep a high rate of construct validity, only those grammar items from Biber's (1998) comprehensive study were chosen to be included among the phenomena of the grammatical component of the POTAI which pose possible challenges of understanding a text clearly for 9th grade students in the bilingual programme. Grammatical phenomena which are considered as simple and straightforward to process while reading a text in English for 9th grade bilingual students were not included in the POTAI (e.g., negation) since their irrelevance from the point of view of the research questions of the current study decreases the content validity of the instrument. Grammatical phenomena which are examined in other components of the POTAI were not included either, e.g., conjunctions (for the extensive list of all the items selected from Biber's (1998) comprehensive study see Appendix F).

3.3.2.1.2 Compiling the grammatical component of the POTAI

In the next phase of designing the grammatical component of the POTAI, the grammar topics covered in the 34 grammar units in the FCE grammar book (Vince, 2003) used as a supplementary book in the 9th grade were compiled in order to extend the Biberian (1998) framework with grammar phenomena specifically relevant for ESL learners at B2 level. Relying on my own professional experience gained through teaching for then six years at the bilingual secondary school, those grammatical phenomena of the grammar book were chosen that are challenging and to some extent even confusing for 9th graders by the end of the academic year. In the cases when the list of grammatical phenomena was more detailed in the grammar book than in Biber's (1998) framework, the more thorough grammatical scheme developed by Vince (2003) was adopted. For example, Biber analysed necessity modals as such, which might further be specified along their temporal aspect according to Vince (2003). At this stage the grammatical component of the POTAI contained nine groups of grammatical phenomena, comprising 74 grammatical phenomena altogether.

3.3.2.1.3 Piloting the grammatical component of the POTAI

In the third stage, the grammatical component of the POTAI developed so far was piloted by analysing two texts, each of approximately 500 words in length. The texts to be analysed 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 presupposition that the texts they find hard to process might abound in challenging grammar features, which are likely to contain register specific language features as well. Nearly unanimously, the students chose a newspaper article from the FCE course book (Unit 3), and it was the chapter on viruses in the biology textbook that all the low-achieving students in English found hard to understand. The pilot study tested to what extent the various grammar phenomena included so far in the grammatical component of the POTAI are present in either of the two texts. As a result of the pilot analysis, various infinitive forms and several additional grammatical phenomena that appeared in these texts but were not yet involved in the instrument were added to the POTAI, e.g., zero conditional, passive with an indirect object. Thus the list of grammatical phenomena extended up to 100 along the same nine groups of grammatical phenomena.

3.3.2.1.4 Teacher interviews

In the fourth phase of the development of the grammatical component of the POTAI, interviews were conducted with four teachers. 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 instrument. Through the expert judgement of the four teachers the internal validity of the analysis, i.e., the fact that the instrument measures what it is intended to measure, was also ensured. For the sake of anonymity, the participants of the interviews chose pseudo names in Hungarian, their mother tongue, to cover their identities. In what follows, the methods and the outcomes of these interview studies will be demonstrated so that the complete form of the grammatical component of the POTAI may be reached.

As mentioned earlier, altogether four interviewees participated in this phase of the study. Erna, 46, is a female teacher of modern languages, English and Russian in particular. She graduated from Eötvös Loránd University in Budapest in 1989 in both subjects. Since then Erna has been instructing English in the bilingual immersion programme of the secondary school. In the third year of her teaching career, Erna was granted a Fulbright scholarship, which gave her the possibility to spend an academic year in the USA teaching English for non-native speakers in a state high-school. Not counting the six years of maternity leave, which she took a decade ago, Erna has been continuously teaching English language and culture for bilingual secondary students. Her teaching experience in the bilingual programme amounts to sixteen years. What makes her an excellent interviewee for developing the grammar part of the POTAI is the fact that she has been regularly involved in the 'zero-year' programme of the school. That is, Erna has seen the linguistic needs and progress of hundreds of bilingual students in the 9th grade. She is one of the language teachers with the most extensive experience in bilingual education in the school.

Szilvi, 30, is a female teacher of history and English language and culture. She graduated from Pázmány Péter Catholic University in 2008. Szilvi has been teaching English and history in English for bilingual students for three years in the secondary school. Before that she was working with international students in a boarding school in the United Kingdom for two years. Taking both ESL environments into consideration, she has been instructing students in English as a second language for five years altogether. The fact that she has been involved in the 'zero-year' programme for three years without a break and her enthusiasm in language pedagogy made her a promising participant of the interview.

György, 53, is a male biology teacher, who graduated as a teacher of biology and geography from Eötvös Loránd University, Budapest in 1983. His studies at the university were completed in Hungarian language. Besides being a science teacher, György is not a qualified teacher of English language and culture. At the time of the interview, he had 28 years of experience in teaching sciences, including 15 years of experience in teaching biology in English in the English-Hungarian bilingual programme of the secondary school. Besides the fact that he was teaching in the 10th grade at the time of data collection, his extensive teaching experience gained through a nearly three-decade-long instruction was crucial in the project.

Tomi, 27, is a male biology teacher, who also graduated as a teacher of biology and geography from Eötvös Loránd University, Budapest in 2009. His science studies at the university were pursued in Hungarian. Similarly to György, Tomi did not qualify as a teacher of English language and culture. When the interview was conducted, Tomi had two years of experience in teaching biology in English. Tomi teaches sciences both in the English-Hungarian and the German-Hungarian bilingual programmes of the secondary school, which means he gives biology and geography classes in English and German as well. He was chosen as one of my interviewees since he was teaching10th graders when the interviews were conducted and thus he had fresh memories of the target group.

The interviews with the English and biology teachers were carried out in 2012 following semi-structured one-to-one interview schedules. To gain insight into the participants' expertise, the interview format was chosen for data collection as it leaves more space for interpreting and explaining opinions in the frame of a "professional conversation" (Dörnyei, 2007, p. 134) than a questionnaire. As to the degree of the structuring of the

interviews, the semi-structured interview form was selected for several reasons. Firstly, its pre-prepared guiding questions along with the prompts brainstormed before conducting the interview ensure that nothing important is left out when carrying out the interview. Secondly, a semi-structured interview is not too tightly controlled, there is "flexibility in the way questions are asked" (Dörnyei, 2007, p. 135), which means spontaneity and variation can arise in the answers as the interviewee elaborates on the issues. Thirdly, no grand tour questions, ones that allow the interviewee to set the direction of the interview through open-ended questions, were necessary in the four teacher interviews since the development of the analytical instrument was beyond its completely initial, exploratory phase.

Following the guidelines of the semi-structured interview protocol, interviews of nearly 102 minutes in total length were conducted with the four participants. Erna gave an approximately 18-minte-long interview, Szilvi shared her insights in 56 minutes, and György imparted his opinion in about 14 minutes, while Tomi expressed his views in 15 minutes. Since Szilvi elaborated extensively on the topic, a single one-shot interview was not enough to discuss all the interview questions, a sequence of three interviews were administered with her to arrive at a full account. All the interview questions were asked in Hungarian, the mother tongue of the interviewees, while in some cases the English teachers gave their answers in a mixed English-Hungarian language as they found speaking about English grammar easier in English. All the interviews were recorded and transcribed. The multiple sessions conducted with Szilvi were written in one single transcript, the beginning and the end of each interview was marked, however, her lines were numbered consecutively.

With regard to the contents of the interviews, both interview schedules (see Appendices G and H) contained two parts, where the first part focused on the professional

history of the interviewee. The introductory questions fulfilled the double aim of breaking the ice, and at the same time creating an atmosphere grounded in professionalism. In both interview guides, the second part included content questions with the intention of initiating answers which ensure the construct validity of the grammar part of the POTAI. The questions involved relatively numerous grammatical terms, with which I was not convinced the science teachers were familiar. In order to avoid embarrassment on the science teachers' part and to help them understand the ESL jargon, I prepared written prompts for them. The biology teachers were invited to read eleven flash cards at their own pace, where the terminology of grammar was exemplified with a biology related sample sentence.

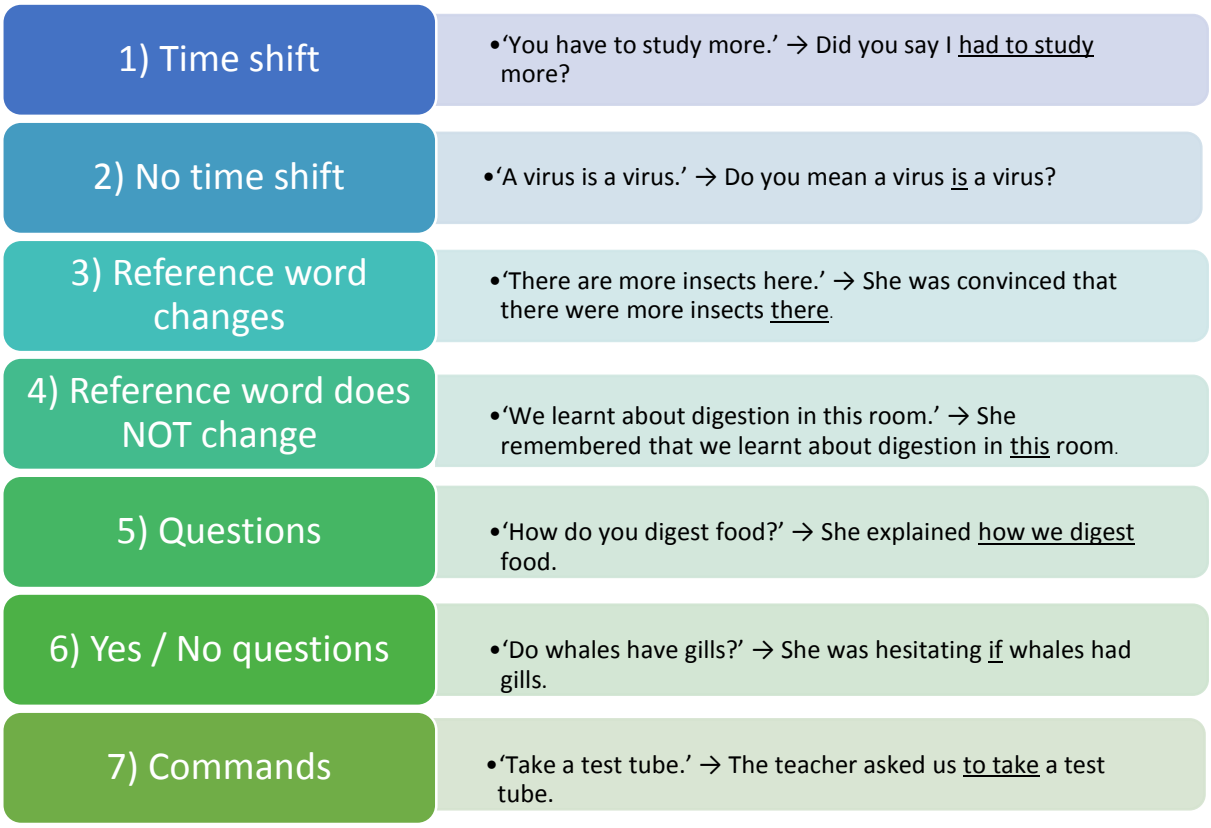

Diagram 1 Indirect speech exemplified on a flash card

The grammatical phenomena on each card were numbered and the term in focus was underlined in the sample sentence in order to make them absolutely clear. As an illustration,

see how the group of the grammatical phenomena of indirect speech was brought closer to the interviewees' worlds on a flash card (Diagram 1).

The biology teachers found the cards giving examples of grammatical terms practical and helpful. The fact that the biology related sample sentences were written on colour cards instead of handing over a long, black and white list of grammatical terms created a playful atmosphere, the interviewees did not feel intimidated by the unfamiliar jargon but rather opened up, which resulted in a positive rapport. Both biology teachers took the flesh cards in hand, were turning them over several times, and used all the references on the cards in their answers (such as the numbers, the names of the particular linguistic phenomena, the words underlined in the sample sentences, and the contents of the examples). György found the cards so instructive and amusing that after the interview he even asked if he could keep them for future use. The final question of the interview schedule was a closing question broadening the topic in order to invite the interviewees' observations and interpretations in related matters.

As the four teachers' opinions were expected to differ, the guiding principle of incorporating their various views in the grammatical component of the POTAI was the following: if any of the four teachers treated a certain grammatical phenomena as important for the students to be familiar with when handling texts, it was accepted part of the analytical instrument. If none of the four teachers considered a grammatical phenomenon to be essential for the students to have mastered in order for them to handle the texts assigned, the grammatical phenomenon was abandoned, and it was not included in the final version of the grammatical component of the POTAI. The reason behind this principle of discarding grammatical phenomena from the POTAI was that the grammatical phenomena that all the

four teachers found dispensable were highly likely not to be present in either the BIOCOR or in the REFCOR. However, if at least one of the four teachers regarded a grammar phenomenon to be important, it was potentially likely to appear in either text of the two corpora – even if its frequency of appearance was expected to be low.

The outcomes of the interviews motivated slight modifications in the instrument. Eight out of nine groups of grammatical phenomena remained exactly the same as the teachers regarded those grammar phenomena as essential to be familiar with when processing the given texts. Although there was a tendency for Szilvi and György to treat grammatical phenomena less vital for the students to be familiar with when handling texts than for Erna and Tomi in general, in the overwhelming majority of the grammatical phenomena of the nine groups of grammatical phenomena at least one of the four teachers thought the grammatical phenoman to be vital. Following the suggestions of my colleagues, the group of question tags (containing four grammar items) were decided to be abandoned completely. Question tags were viewed by my colleagues as not being typical of written discourse in contrast to their excessive presence in spoken registers. Szilvi explained her reasoning by claiming that question tags were more typical of spoken English (Szilvi, line 135), while Tomi expressed the same idea by saying that it was rather him, the teacher, who used question tags in class (Tomi, line 79). As a result, question tags were not included in the grammar part of the POTAI.

3.3.2.1.5 Finalising the grammar component POTAI

Finally, after the outcomes of the interviews were included in the grammatical component of the POTAI, one more alteration was made to the instrument. 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 group of grammatical

phenomena labelled as indirect speech. In this finalizing stage, however, the group practically termed as indirect speech so far in the research was submerged into the aspect of relative clauses, since indirect speech utterances, which function in some respects like noun phrases, are nominal relative clauses from a linguistic point of view (Quirk et al., 1989). Despite the fact that English language course books tend not to emphasise the overlap of the two linguistic phenomena (indirect speech and nominal relative clauses) for the sake of simplicity, a more theoretical linguistic precision was chosen to be followed in the current research. Besides theoretical reasons, a practical point was also considered. Indirect speech in English is a phenomenon whose grammatical rules (e.g., forming the appropriate verb form, word order, leaving out the relative pronoun, and punctuation) are completely different from those in Hungarian. For this reason, ESL learners with Hungarian mother-tongue are inclined to find certain indirect speech sentences just as perplexingly complex to process as nominal relative clauses. Thus treating the two features together through finding their accumulated frequency in the BIOCOR can provide data about one of the possible reasons why target readers in the current research environment might find processing the texts difficult. At this stage, the grammatical component of the POTAI comprising seven groups of grammatical phenomena with 96 grammar phenomena was finalized (see Table 11).

Group of grammatical phenomena	**Grammatical phenomena**
Tenses	Present simple
	Present continuous
	Past simple
	Past continuous
	Past perfect simple
	Past perfect continuous
	Used to structure
	Present perfect simple
	Present perfect continuous
	Future simple
	Future continuous
	Future perfect simple
	Future perfect continuous

	Going to structure
Conditionals	Zero conditional
	1st conditional
	2nd conditional
	3rd conditional
	Mixed conditional
Passive voice	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
Infinitives	Simple infinitive
	Passive infinitive
	Progressive infinitive
	Progressive passive infinitive
	Perfect infinitive
	Perfect passive infinitive
	Perfect progressive infinitive
	Perfect progressive passive infinitive
Prepositions	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*

Table 11 The grammatical component of the POTAI

3.3.2.2 Procedures of data collection and analysis

The grammatical component of the POTAI was applied to both corpora as follows. First, the basic unit of analysis was decided to be the sentence, as it is the most straightforward conscious unit in a written text. Theoretically it is also possible to compute the frequency of various grammatical phenomena in a register compared to the total length of the particular text (expressed in the number of words it contains), however, comparing this

figure to the number of sentences the texts consists of was considered to be more reasonable in the pedagogically motivated present research. If the unit of analysis is the sentence rather than the number of words in the text, the research results in data which is more readily revealing and applicable to practicing ESL teachers who are not experts in text-analysis. For instance, information stating that a particular grammatical phenomenon appears in every second sentence in a register is more easily worded than claiming the same idea by saying that the grammatical phenomenon appears 254 times in a text of approximately 7000 words. Sentence boundaries were marked by a word processing software (Microsoft Word, 2013) and counted in all the texts. Next, the grammatical phenomena of the POTAI were manually tagged in the corpora, each grammatical phenomenon was labelled by a code number in the texts. Then the appearance of the code numbers was totalled in each individual text and added up in both registers. The frequency of each grammatical phenomenon was counted against the basic unit of analysis, thus the ratio of grammatical phenomena per number of sentences in the registers was computed.

In order to ensure the researcher's reliability of the analysis, inter-rater reliability was computed. A colleague was asked to test the coding scheme on one text from the BIOCOR and on one from the REFCOR. The two longest texts in the two corpora were chosen (the chapter on 'The characteristics of living things' from the BIOCOR, and the reading entitled 'The good, the bad, and the unbearable' from the REFCOR) so that the most extensive lengths provided a greater chance of testing the extent of the reliability at which the researcher applied the coding scheme to the texts. The co-rater was introduced into the coding system, in which training special attention was paid to the grammar phenomenon of indirect speech, which is treated as a nominal relative clause in the present research. The two sets of outcomes (the coding of the colleague and that of the researcher) were compared. In the case of the text

from the BIOCOR, the co-rater applied exactly the same grammatical phenomena for all the 378 tags, which shows an outstandingly high rate of reliability (100%) on the researcher's part. While in the case of the text from the REFCOR, the co-rater categorized one single grammatical phenomenon differently out of the 144 tags, still proving an extremely high rate of reliability (99%). From these two high percentages of inter-rater reliability, it can be concluded that the analysis was reliable.

To arrive at a meaningful description of the register of biology texts for secondary students, the frequency of grammatical phenomena in the BIOCOR was compared with that in the REFCOR by means of computing t-tests. Since the frequency ratios of grammatical phenomena in the two registers do not depend on each other, independent-sample t-tests were counted. In many cases, it was only one of the corpora which contained a given grammatical phenomenon. If either of the two registers contained a grammatical phenomenon, 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 it was that frequency ratios with too high probability coefficients ($p>.05$) in the sample do not show generalizable characteristics but sample-specific traits. In order to choose the appropriate probability coefficient of a given frequency ratio of a grammatical phenomenon, Levene's tests were conducted. In the cases where the Levene's tests showed 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. Checking the results of the Levene's tests was a step in the procedure which ensured the interpretation of the results to be reliable in distinguishing register specific traits from sample specific features. The statistical method of Levene's test allowed the grammar part of the POTAI to yield both register specific (generalizable) and sample specific results, which increased the precision of

transferability of the findings. Finally, the mean value of each grammatical phenomenon was compared and contrasted in the two registers. To make the meaning of the mean values more easily graspable for ESL teachers who are not well-practiced in applied statistics, the mean values were also expressed in terms of sentences-based frequency. In this fashion, the mean value M=.5, for instance, was explained as the given grammar phenomenon appearing in every second sentence.

3.3.3 Sentence complexity

The level of the complexity of sentences in a text affects reading performance. To find the extent to which sentence complexity might pose difficulty of comprehension for 10th grade bilingual students, four readability predicators are investigated in the current research: sentence length, packet length (i.e., words between given punctuation marks, forming a logical unit, to be defined in Section 3.3.3.2 on p. 102), the relationship between the length of words and that of sentences, and syntactic structure.

3.3.3.1 Sentence length

Sentence length is an important factor affecting the level of sentence readability. Sentences which are longer than conventionally accepted tend to be more difficult to comprehend than shorter ones (Harrison & Bakker, 1998; Jacobson, 1998; Ulusoy, 2006). The number of conventionally accepted words in a sentence, which is still perceived as of appropriate length, has changed over time, as Sherman (1893), one of the forerunners of text analysis observed. Recent writing manuals (Blakesley & Hoogeveen 2011; Hart, 2007; Williams, 1995) suggest 17 to 20 words or fewer in a smoothly-comprehensible sentence. This figure harmoniously corresponds to the length of sentences published in the newspaper *Economist* for its intended target group of educated readers (Harrison & Bakker, 1998). When

computing the length of sentences in the corpora, the current research relied on a word processing software (Microsoft Word, 2013) to ensure high reliability. Titles and headings are strictly speaking not full sentences, consequently the word processing software did not count these elliptical structures as sentences. This is due to the lack of sentence final punctuation mark at the end of titles and headings. However, these text organizing devices were still counted as sentences in the present investigation for two mainly pedagogical reasons. Students in the bilingual immersion programme are expected to read and comprehend titles and headings, too. Their reading assignments are not confined to the running texts, in contrast, they are advised against skipping these parts of the text already in the 9th grade (as some of the FCE tasks focus on the messages conveyed in the title). Thus the presence of titles and headings in the BIOCOR cannot be ignored. Even if there is no punctuation mark to close these units, the layout of the texts clearly separates them from other neighbouring sentences. Accordingly, there would be no sense in supposing that the target readers of the corpora fail to recognize that these lines are conscious units of the text, separated from other sentences in their textual environment. Besides counting titles and headings as sentences, there was one more adjustment of the results computed by the word processing software. In the case of sentences which contained sequences of numbers (e.g., 1 2 3 4 5 6), the series of numbers were counted as one unit, as if the string of numbers was one single word. The reason for this grouping was that two or three of these sequences in one sentence would have increased the number of words in the sentence dramatically, in certain cases with even 10 to 20 words. However correct mechanically the increase is, keeping it in the research unchanged would have distorted the validity of the computations. The reason for this is that a lengthy sentence packed with a few strings of numbers is not as challenging to process in a linguistic sense as an extremely long one containing words only and no sequences of numbers.

The results in the current research were gained through counting the average sentence length of the BIOCOR and comparing it with that of the REFCOR. To increase the validity of the description of the BIOCOR from the point of view of its sentence length, it was not only the average sentence length of the corpus that was computed. Calculating one single average gives limited information; besides, averages can be quite different from typical, frequently occurring values. For these reasons, the distribution of various sentence lengths throughout the corpus was investigated. To make the interpretation of the distribution of different sentence lengths in the BIOCOR meaningful, the results were compared to that of the REFCOR. The comparison of the two corpora was made possible by converting the frequency values of the sentences of various length into percentages first.

3.3.3.2 Packet length

It has already been noticed that even sentences which are longer than conventionally expected might not be extremely difficult to comprehend as long as they are broken up effectively (Harrison & Bakker, 1998). The effective splitting up of a sentence is carried out by the means of eight punctuation marks: full-stop, comma, colon, semi-colon, exclamation mark, question mark, long dashes and parenthesis. The words between these punctuation marks, indicating logical units for the reader, were termed a packet by Harris and Bakker (1998). A packet might be the same cluster of words as a clause; however, the identity of a packet and a clause is not necessary. Dividing sentences into packets is based on a rather mechanical model, which avoids making complex distinctions according to different grammatical functions of punctuation (Quirk, 1989). Despite this apparent simplicity, packet length was proved to be a valid measure of predicting reading performance (Harris & Bakker, 1998). To investigate the level of readability of the BIOCOR, the present research computes the measure of packet length the Harrison-Bakkerian way. That is, adjacent punctuation

marks (e.g., a parenthesis followed by a comma), which result in zero packet length, were excluded in the computation. Punctuation marks which denote contractions (e.g., '*didn't*') were not treated as packet indicator marks, since their function is not that of revealing a logical unit in the sentence. To find the packets in the corpus, a word processing software (Microsoft Word, 2013) was applied, which warranted a high rate of reliability. First the average packet length of the BIOCOR was counted and contrasted with that of the REFCOR. The comparison was feasible since the raw frequencies of various packet lengths were converted into percentages initially. Characterizing a register by one single average is rather simplistic and of low content validity. With the intention of avoiding these flaws, the distribution of different-sized packets in the BIOCOR was therefore also analysed to increase the sophistication and content validity of the account of the register. Relative frequencies (expressed in percentages) of packets with various lengths were compared within the research environment, that is, with those of the BIOCOR.

3.3.3.3 Readability indices

The extent how easy a text is to read can be expressed through readability measures (Hargis et al., 1998), i.e., mathematical formulae which have been developed to predict the level of readability of texts by objectively calculable means. Readability indices compute the number of characters or syllables in words and sentences. Although these variables might be argued to be nothing else but surface features of a text (Schriver, 2000) and thus not appropriately applicable for defining its level of difficult, readability indices correlate with conceptual properties of texts (Kintsch & Miller, 1981) and compare well with other measures of text difficulty (for more details see Section 2.6 on pp. 38-41).

Different readability indices apply different variables in their calculations, which variety leads to differences in their results. Furthermore, as there is no commonly agreed reference point of understanding a text, different readability formulae can set the minimum rate of understanding at reasonably different levels of accurate processing. For these reasons, it cannot be justly claimed that the results of a particular readability index are more precise or more correct than those of another. As it would be far too simplistic to attempt to find the readability index that predicts the level of difficulty of a text the most accurately, the present research takes the line of a comparative investigation. The current analysis computes and compares several indices which determine readability based on different variables rather than calculating one single readability index. Finding the comparative results of various readability indices instead of relying on one single one also has the advantage of gaining more balanced, less extreme data, which increases the content validity of the research.

In the course of choosing which few of the hundreds of readability measures to incorporate in the POTAI, it was taken into consideration that there are different types of readability indices. Based on the core measure in which the length of words is expressed, all readability indices can be grouped into two main categories. One of the main categories works out the length of words by counting characters per word, while the other is slightly more complex, as it defines the length of words in syllables per word (Coleman & Liau, 1975). Character-based readability indices were developed to reduce the difficulty of mechanically calculating the index in the case of hard-copy texts (Coleman & Liau, 1975; Smith & Senter, 1967). The merit of this type of readability index lies in the fact that computers are more accurate at counting characters than syllables, that is, the reliability of such readability indices is high. While the second main type, syllable-based indices, might be closer to how longer words are processed when reading, that is, by cutting lengthy words into chunks, either into

meaningful units or into conveniently manageable syllables. It is widely believed that the more syllables per word and the more words per sentence a text contains, the more difficult it is to process, thus the higher its readability index is (White, 2011). This belief is based on the notion that multisyllabic words require more time to read and process than monosyllabic words (Rayner & Pollatsek, 1989). The current research applies conventional, widely-validated and extensively-used readability measures that calculate the ease of reading a text in both ways: ones that use the advantage of character-based calculations and ones that rely on the merit of syllables-based computations. Secondly, readability indices can also be grouped based on a different aspect: the type of data their computations yield. The formulae of some of the readability indices result in abstract, raw numbers, while others are designed in such a way that their raw scores are directly converted into grade levels. That is, the value of such readability indices immediately reveal how many years of formal education is expected from the reader of the given text in order to be able to process it easily. Keeping the pedagogical aim in the foreground, the present research applies only the latter type of readability indices (grade level indices), and avoids arriving at theoretically abstract figures. The choice of applying grade level readability indices ensured maintaining the pedagogical implications of the data clearly visible. To find the grade level of a text, no further complex calculation is needed in the case of the latter type of readability indices but a simple rounding to the nearest unit. For instance, a readability index with the value of 10.5 should be rounded up, which indicates that the text is expected to be read without considerable difficulties by eleventh graders, or 16-17 year-old old teenagers. On the other hand, a readability index with the value of 10.4 should be rounded down, which shows that the text is easily read by 15-16 year-old tenth graders. Table 12 clarifies which age group is meant by the rounded output of the grade level readability indices (www.fulbright.org.uk).

Grade level	Age group
Kindergarten	5-6 years old
First grade	6-7 years old
Second grade	7-8 years old
Third grade	8-9 years old
Fourth grade	9-10 years old
Fifth grade	10-11 years old
Sixth grade	11-12 years old
Seventh grade	12-13 years old
Eight grade	13-14 years old
Ninth grade	14-15 years old
Tenth grade	15-16 years old
Eleventh grade	16-17 years old
Twelfth grade	17-18 years old
College	18-20 years old

Table 12 Grade levels and their corresponding age groups

As all grade level indices are based on the principle of rounding, using various readability indices instead of one single index increased the content validity of the research. Namely, index figures which slightly differ in their values (e.g., 0.1) but are rounded to a different unit (one up, the other down) can imply a two-year difference in the number of presumed formal education need to process the given text without considerable difficulties. The comparative analysis of various indices, however, balances the distorting effect of the rounding process.

Grade level indices provide grade levels which are used in the American education system. These levels are identical to the ones used in the Hungarian education system (www.okm.gov.hu/letolt/english/education_in_hungary_080805.pdf); nevertheless, the grade level index of a text is obviously not transferable directly to a Hungarian grade level. Hungarian students learn English, the medium of the texts, as a foreign language, while it is the mother tongue for American students. Thus the required number of years spent in education is not predictable through these measures in the current research environment, where the reading community of the texts is a group of monolingually raised Hungarian

students, who pursued the first eight years of their studies in Hungarian, not in English. Still, using grade level readability indices provides the chance to form an image about the age group of the expected target readers of the texts under investigation, which is more tangible, and in effect, more helpful for educators than a figure which summarizes the readability level of a text in abstract data.

Consequently, the following five grade level indices were applied as part of the POTAI in the current analysis: the automated readability index (ARI), the Coleman-Liau index, the Flesh-Kincaid index, the SMOG index, and the Gunning fog index. These grade level readability indices differ not only in their specific formulae and the core measures they use but also in their approaches. Some of them analyse the whole text, while others take samples of different lengths to generalize the results for the whole text.

1) The automated readability index **(ARI)** was created early in the 1960s by Smith and Senter (1967), and it was validated a few years later (Kincaid & Delionbac, 1973; Smith and Kincaid, 1970). It is a character-based index, which takes the whole text into consideration, and computes the grade level for the entirety of the text. The formula reads as follows.

$$ARI\ index\ =\ 4.71 \times \frac{number\ of\ characters}{number\ of\ words}\ +\ 0.5\ \times \frac{number\ of\ words}{number\ of\ sentences}\ -\ 21.43$$

2) The **Coleman-Liau** index, which is also a character-based index, takes three random samples of the text and counts an average grade level by using the following formula.

$$Coleman - Liau\ index\ =\ 0.0588 \times \frac{number\ of\ characters\ in\ 100\ words}{100}\ - 15.8\ -$$

$$0.296 \times \frac{number\ of\ sentences}{100\ words}$$

3) The **Flesh-Kincaid** index, one of the most tested and reliable (Chall, 1958; Klare, 1963) readability indices, was designed for the US navy. Its syllable-based formula counts the grade level for the whole text.

$$Flesh - Kincaid\ index\ = 0.39 \times \frac{number\ of\ words}{number\ of\ sentences} + 11.8 \times \frac{number\ of\ syllables}{number\ of\ words} - 15.59$$

4) The most popular readability index for teachers (Ruddell, 2005), the syllable-based **SMOG** index, created by G. Harry McLaughlin (1969), analyses three ten-sentence long samples of the text and counts the average of those thirty sentences according to the following formula.

$$SMOG\ index\ = 3\ + \sqrt[2]{polysllable\ count\ in\ 30\ sentences}$$

The term 'polysyllable count' covers the number of words which are at least three-syllable long.

5) The syllable-based **Gunning fog** index, designed by Gunning (1952), predicts the readability of a text based on one single paragraph of the text by applying the following equation.

$$Gunning\ fog\ index\ = \frac{number\ of\ words}{number\ of\ sentences}\ + 0.4\ \times \frac{number\ of\ polysyllabic\ words \times 100}{number\ of\ words}$$

Similarly to the terminology of the SMOG index, the term polysyllabic word also means words which contain more than two syllables.

The above five grade levels of the BIOCOR and those of the REFCOR were determined by using a readability index software (www.online-utility.org). In the case of the BIOCOR, the grade levels were calculated in two ways. They were computed for the entirety

of the texts to determine the collective average grade of the BIOCOR. Then, they were calculated separately, chapter by chapter as well, in order to pinpoint which chapters deviate from the average grade level of the BIOCOR, either by being more difficult or by being easier than the others.

3.3.3.4 Syntactic structure

The readability of a text also depends on the complexity of its sentences, which can be revealed by the complexity of its syntactic structures (Huddleston, 1984). The cognitive demands a text poses varies considerably according to the complexity of the syntactic structure of its constituent sentences (Crossley et al., 2008; Perfetti et al., 2005). 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*. 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*. 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 minimum one dependent clause.

To describe the sentence complexity of the BIOCOR from the point of view of its characteristic syntactic structure, the frequency of ten types of sentence structures were tapped in this research (see Table 13). These ten syntactic categories of the POTAI were set up and finalized as a result of a pilot study on two texts of the corpora (for the guiding principles of choosing the texts for the pilot see Section 3.3.2.1 on pp. 86-88).

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 clause
6	Complex sentence	three dependent clause
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 13: The types of syntactic structures analysed in the corpora

The syntactic structures listed in Table 13 were chosen so that the possible combinations of syntactic categories covered all the variety of different syntactic 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 structures in the BIOCOR and the REFCOR were tagged manually. All the ten types of syntactic structures were labelled by a code number in the texts (for the meaning of the code numbers go back to Table 13). Next, 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, thus the ratio of syntactic structures per number of sentences in the two registers were 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 computed. In two cases, code numbers nine and ten, it was only one of the two corpora which contained the given syntactic structure. 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 the cases where the Levene's tests revealed a significant difference ($p<.05$) equal variances were presumed; while the lack of significant difference when running the Levene's test ($p>.05$) lead not to presume equal variances. Examining the results of the 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 the Levene's tests on the

syntactical 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.

3.3.4 Textual metadiscourse (TMD)

Texts differ in the extent how explicitly the logical flow and organization of their ideas are signposted through overt markers, i.e., textual metadiscourse (TMD). The abundant presence of the various markers of TMD dramatically affects the ease with which the reader follows the content of the text; in contrast, texts that use TMD markers sparingly and thus keep TMD covert tend to be more challenging to process (Gosden, 1992). The frequency and variety of the markers of TMD in the BIOCOR was analysed in order to examine whether the corpus poses challenges for the 10th grade bilinguals due to the corpus's limited use of TMD devices that support the apparent cohesiveness of the ideas in the text.

The current study relies on the analytical scheme of TMD in academic texts developed by Hyland (1998b; 2000), which contains five main categories (see Table 14). The categories of Hyland's (1998b; 2000) academic text analytical model were formed through gathering various TMD functions into five groups.

Category	**Examples** (listed by Hyland, 2000, p. 111)
1) Logical connectives	*in addition; but; therefore; thus; and*
2) Frame markers	*finally; to repeat; our aim here; we try*
3) Endophoric markers	*noted above; see Figure 1; table 2; below*
4) Evidentials	*according to X; Y, 1990; Z states*
5) Code glosses	*namely; e.g.; in other words; such as*

Table 14 Hyland's (2000) scheme of textual metadiscourse in academic texts

The elements of the first category, that of logical connectives, provide signposts for the readers to interpret pragmatic connections between notions of the text by overtly

expressing three types of relations: additive, resultive and contrastive ones. The category of frame markers, on the other hand, signal text boundaries or structural parts of the text, and shifts in the discourse. The functions of this category covers sequencing; showing various stages of a text; explicating discourse goals; and illuminating topic shifts. The linguistic elements of the third main category of endophoric markers make reference to other parts of the text, and at the same time the writer's intentions with regard to argumentation become clearly noticeable through them. Contrastingly, linguistic elements of the fourth main group, evidentials link sources of information which originate from outside the text. Through establishing relations with other texts, evidentials serve various functions. They increase the credibility of the text by demonstrating the writer's awareness of research knowledge; support the reader's interpretation of the text through placing the particular ideas and data within the academic field; and also help creating intertextuality. Finally, elements of the fifth category, code glosses, provide additional information which are intended to help the reader recover the writer's message by way of explaining, comparing or expending what has already been stated.

In the present research, the analytical scheme of TMD in academic texts (Hyland, 1998b, 2000) was modified in order to adapt its categories to the current research environment of pre-college level academic texts read by monolingually raised bilingual students. The extension of the scheme was carried out from a pedagogical perspective. Gaining insights into the salient similarities and differences of the two register practices (those of the BIOCOR and of the REFCOR) with regard to the TMD markers that might be focal points of ESL and biology ESP classes was kept in the foreground. As a result, various TMD functions were not merged into one single category, unlike in Hyland's (1998b, 2000) scheme, but were taken apart into separate categories. The lack of blending different TMD functions into one category (e.g., the three different functions of addition, result and contrast into the fusional

category of logical connectives) increased the content validity of the research since distinct rhetorical functions and various logical relationships became more clearly visible in detail. After the separation of the above registered TMD functions, several academic writing course books (Bailey, 2011; Boyle & Warwick, 2014; Chazal & Moore, 2013; Ruiz-Garrido et., al. 2010; Hogue, 2008; Jordan, 2002; Mann & Taylore, 2007; Swales & Feak, 2012) were consulted in order to incorporate in the present analytical scheme any yet unlisted TMD markers which noticeably create links between ideas in a text. Consequently, four more functional categories were added: reason, purpose, explanation and summarizing.

Functional categories	TMD markers
Addition	*and, besides, in addition, additionally, also, moreover, furthermore, what is more, too, as well*
Contrast	*but, although, though, even though, while, whereas, however, nevertheless, nonetheless, despite, in spite of, on the other hand, in contrast, by contrast, by/in comparison, on the contrary, yet*
Purpose	*so, so that, in order to, so as to, for, infinitive of purpose*
Reason	*because, because of, as, since, for, on account of, due to, owing to, for this reason*
Result	*consequently, as a consequence, as a result, thus, therefore, hence*
Sequencing	*first, first of all, first and foremost, to start with, to begin (with), second(ly), in the second place, next, then, after that, subsequently, finally, lastly, last of all, last but not least, in the end, numbering (1, 2, 3), listing (a, b, c)*
Explanation	*for example, for instance, such as, i.e., e.g., that is, that is to say, namely, in other words, this / which means, so, specifically, known as, defined as, called*
Topic shift	*well, now, so, to move on, to look more closely, to come back to, in regard to, with regard to*
Endophoric markers	*see / noted / discussed below / above / earlier / later / before; section X, chapter X, Figure / Fig. X, Table X, Example X, page X, Investigation X, Table X*
Evidentials	*according to X; Y, 1990; Z states; Z states; Z estimates; Z maintains; Z suggests; Z affirms; Z proposes; Z recommends; Z offers; Z asserts*
Summarizing	*in conclusion, to conclude, to sum up, in sum, overall, on the whole, all in all*

Table 15 The extension of Hyland's (1998b, 2000) scheme: the TMD component of the POTAI

To ensure a high rate of content validity, one of the colleagues of the researcher, who teaches academic writing in the bilingual school (for students participating in the International Baccalaureate Diploma Programme), was asked to extend the list with any further TMD categories on which she puts emphasis in her academic writing classes. Her teaching material did not include any additional functional categories with which the comprehensive analytical scheme could be supplemented; however, she extended the list with a few more exemplary elements of TMD markers. As a result, the following TMD analytical model containing 11 functional categories was developed for the examination of precollege level academic texts (see Table 15).

The newly revised and extended analytical scheme was applied to the BIOCOR and to the REFCOR in order to compare and contrast the TMD practices of the two registers. To keep the reliability of the TMD analytical part of the POTAI high, TMD markers were collected in the corpora by using the word search function of a word processing software (Microsoft Word, 2013). To raise the level of content validity of the research, the sentences in which the particular TMD markers appear were individually checked to make certain that the items were TMD markers indeed. Namely, some of the TMD markers, for example, 'see,' 'now,' and 'investigation,' can function as lexical items and as TMD markers as well, which profoundly different functions cannot be distinguished by the word processing software. The manual double-checking also gave the chance to clarify the functions of various TMD markers, i.e., 'so' can mark triple functions: that of topic shifting and of giving purpose or explanation; or the TMD marker 'for' can serve the functions of providing reason and of purpose as well. In a similar manner, being aware of the possibility that the list of TMD markers might not contain all the conceivable TMD markers made a further re-reading of the two corpora indispensable in order to collect instances of additional TMD markers. These

newly listed TMD markers were also recorded among the 120 TMD markers of the comprehensive scheme displayed in Table15. Next, the raw appearance of the particular TMD markers was computed in each functional category. Then, the ratio of the TMD markers was calculated in the entirety of the BIOCOR and the REFCOR against the number of sentences in the two corpora, which conversion made the data comparable across the two registers. Since the unit of measure is that of the sentence, the value of percentages signifies the frequency of the occurrence of TMD markers against the sentence. For example, a 50% TMD frequency means that every second sentence on average contains a TMD marker in the text. The frequency of TMD markers in each functional category was examined in order to become cognizant of the extent to which various logical relationships and rhetorical functions are overtly marked in the two corpora to help the target reader follow the logical flow of the text.

3.3.5 Summary of the methods

To settle the accomplishments of the data collection and data analysis of the linguistic variables of the POTAI, the results upon which the next chapters are built are reviewed in Table 16.

Components of the POTAI	
Lexis	frequently used words
	keyness
	lexical density
Grammatical phenomena	tenses and tense related structures
	conditional structures
	passive voice and causative structures
	relative clauses
	nominal relative clauses
	infinitives
	prepositions at the end of sentences
	modal auxiliaries
	sentence length
	packet length
	readability indices
	syntactic structure
Metadiscourse	textual metadiscourse

Table 16 The components of the finalized instrument (POTAI)

As a conclusion of the previous four sections, a quick overview, which summarizes the outcomes of Section 3.3 of the Methods Chapter of the current research, is provided in the form of a chart. Table 16 displays the components of the finalized POTAI, listing all the groups of linguistic variables which were eventually integrated as elements of the text analytical instrument.

4 Results and discussion

The ensuing chapter provides a thorough linguistic description of the register of the biology texts for secondary students from the point of view of ESL teaching by presenting and interpreting the data yielded by the application of each component of the POTAI to the BIOCOR and to the REFCOR, which served as a baseline of comparison. The results are examined not only from a sheer theoretical perspective of offering a depiction of the biology textbook register; however, the data are examined from a pedagogical point of view. The discussion aims to clarify the extent to which the typical linguistic patterns of the BIOCOR might pose challenges of comprehension for 10th grade bilingual students, who have successfully processed the REFCOR.

4.1 Lexis

The lexical component of the POTAI discloses information about the biology textbook register from three angles: 1) its frequently occurring lexis, 2) its keyness pattern, and 3) its lexical density.

4.1.1 Frequently occurring words

In the following section, the BIOCOR is described from the point of view of its prevalent lexical items. Frequently occurring words are compiled in three categories: biology terms, academic English lexis and general English lexical items. The lexical environments of the most repeatedly occurring biology terms in the first three bands, that is, the bands containing biology terms that appear at least fifteen times in the BIOCOR, are also recorded here, in order to reveal with which words the highest-frequency biology terms are typically and possibly used.

4.1.1.1 Frequently occurring words in Band 1

After carrying out lemmatization of the frequently occurring words in the BIOCOR that share the same root, and computer counting the frequency of the word families, the results were arranged in frequency bands. The lexical items that appear most recurrently (minimum 30 times) in the BIOCOR are listed in Band 1. Table 17 contains these high-frequency items, showing the number of their raw occurrences and also their relative frequency in the BIOCOR expressed as a percentage.

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

Table 17 Band 1: the most frequent lexical items in the BIOCOR

For instance, the top most frequent biology term *'parasite'* appears 57 times in the BIOCOR (raw occurrence), which constitutes 0.8% of the corpus (relative frequency). There are not more than five biology terms among the most frequently used lexical items, *'parasite,' 'cell,' 'bacteria,' 'virus,'* and *'growth.'* However, the majority of the typically applied items in Band 1 is general English lexis, not biology terms. Although most of these items are related to the topic of biology (e.g., *'animal,' 'plant,' 'body'*), they still do not form specific biology vocabulary. Contrary to the expectations expressed by both biology teachers and students of the bilingual programme claiming that biology texts are full of academic vocabulary (Cserép, 1997), the band of the most frequently used lexical items contains no academic English vocabulary at all.

The lexical environment of the most frequently used biology terms is described in detail so that the data gained here are readily applicable for drawing implications for biology ESP teachers. Table 18 shows all the collocations the lexical item '*parasite'* takes in the biology corpus. It can be seen that the token appears in various noun phrases, such as *'life cycle of the parasites,' 'malarial parasite,'* or *'worm-like parasite.'* The term '*parasite'* is even richer with regard to the verbs it takes, there are 15 different verbs used with it in the BIOCOR. Drastically more sparingly in number, it also appears as an object of verbs, for instance '*kill the parasites,'* or *'transmit parasites,'* and with verbs in the passive voice, e.g., '*parasites are carried to humans*.'

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

Table 18 Lexical environment of the biology term '*parasite'* in the BIOCOR

The second most frequent biology term, '*cell*,' has plentiful word combinations in the BIOCOR (see Table 19). It forms numerous noun phrases, including *'cell membrane,' 'cell wall,'* and *'red blood cell'* among the more than dozen combinations. However, the variety of verbs it takes in the BIOCOR is not that vast, including *'become,' 'burst,'* and *'contain.'* Nevertheless, it has a tendency to function as the object of verbs, for instance *'attack,' 'fill,'* and *'rob.'* It is also typically applied with verbs in the passive voice, such as *'the cell is bounded,' 'the cell is released,'* and *'the cell is surrounded.'*

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

Table 19 Lexical environment of the biology term '*cell*' in the BIOCOR

The third most frequent token, '*bacteria*,' has a modest number of collocations in the BIOCOR (see Table 20). It appears in noun phrases both as an adjective (e.g., *'bacterial cell'* and *'bacterial colonies'*), and it also functions as the head of the noun phrase, for example '*streptococcal bacteria*.' The selection of verbs it takes is wide-ranging, including *'clump,' 'multiply,' 'survive,' and 'vary.'* Neither is its appearance as an object of a verb scarce, it is applied for instance with *'grow,' 'hold back,'* and *'remove'* among others. However, it is not typically used with a verb in the passive voice. There are no more than two examples for such combinations, namely *'bacteria is given moisture'* and *'bacteria are surrounded.'*

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

Table 20 Lexical environment of the biology term '*bacteria'* in the BIOCOR

The fourth most repeatedly applied biology term, '*virus*,' has a humble set of collocations in the BIOCOR (see Table 21). There is hardly any noun phrase where it is the head, such as in *'new virus,'* and *'structure of the virus.'* However, it combines with a fair number of verbs, among them are *'attach,' 'attack,'* and *'reproduce.'* No more than two verbs

take the token '*virus'* as an object, namely *'cultivate'* and *'grow.'* The most numerous collocations of the lexical item are verbs in the passive voice, for instance *'viruses were discovered,' 'a new virus is formed,' 'viruses are released,'* and *'viruses are set free.'*

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

Table 21 Lexical environment of the biology term '*virus'* in the BIOCOR

The fifth most recurrent word family in the biology corpus, '*growth*,' takes a moderate number of collocations (see Table 22). In the form of a past participle modifier, it appears in one single noun phrase, *'full-grown earthworm.'* The verbs it combines with are related to the time span of growth, for example *'speed up,' 'stop,'* and *'go on.'* Signifying an action, it appears both as a transitive verb, for instance increasing the number of *'bacteria'* and *'viruses,'* and an intransitive verb, such as *'living things,' 'moulds,'* and *'worms'* become larger. The token as a verb is also typically applied with prepositional phrases, either showing directions (e.g., *'in a particular direction'* and *'towards light'*), or indicating a place (e.g., *'on the agar'*), or showing dimensions, such as *'to their full size.'*

In a noun phrase	*full-grown earthworm*
Verb it collocates with	*go on growing*
	growth takes place
	growth stops
	speed up their growth
	stop growing
	amoebas grow

Nouns it collocates with	*grow bacteria*
	grow viruses
	living things grow
	moulds grow
	parasites grow
	worms grow
Verb and a prepositional phrase	*grow in a particular direction*
	grow on the agar
	grow to their full size
	grow towards light

Table 22 Lexical environment of the biology term '*grow'* in the BIOCOR

4.1.1.2 Frequently occurring words in Band 2

The second most frequently applied lexical items in the BIOCOR belong to Band 2, which contains word families that appear at least 20 times in the corpus (see Table 23).

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

Table 23 Band 2: the second most frequent lexical items in the BIOCOR

While Band 1 includes five biology terms, Band 2 comprises no more than two: *'amoeba'* and *'reproduce.'* Similarly to the previous band, the word families of Band 2 are characterized by the abundance of general English lexis, which is four times more prevalent among the lemmas of this band than biology terms. While general English lexis in Band 1 is mostly biology related, general English vocabulary in Band 2 is not closely connected to biology topics. Such common items as *'do,' 'make,' 'take,' 'person,' 'thing,'* and *'small'* belong to basic vocabulary, they are not associated with biology areas at all. It is only the item '*worm'* that is related to the field of biology. In the same way as in Band 1, the complete lack

of appearance of academic English vocabulary goes contrary to the biology teachers' and the bilingual students' assumptions alike (Cserép, 1997).

The sixth most frequent biology term, '*amoeba*,' is used with a small number of collocations in the BIOCOR (see Table 24). It appears with no more than three modifiers in a noun phrase, taking the adjectives *'dysentery,' 'live,'* and *'ordinary.'* The variety of verbs the token combines with is not rich either; what is more, most of the collocating actions denote basic verbs, such as *'change,' 'eat,'* and *'live.'* In a similar manner, the biology term is narrowly used as an object of verbs; its appearance is limited to *'examine'* and *'see.'*

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

Table 24 Lexical environment of the biology term '*amoeba'* in the BIOCOR

The seventh most recurrent biology term, '*reproduce*,' appears in a twofold way in the BIOCOR (see Table 25).

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

Table 25 Lexical environment of the biology term '*reproduce'* in the BIOCOR

The token either combines with a noun phrase as its subject, namely, the living thing that reproduces (e.g., *'amoeba,' 'bacteria,' 'euglena,' 'parasite,' 'virus'*), or in more general terms *'offspring'* and *'organism';* or it collocates with an adverb of manner or with a prepositional phrase describing how the reproduction takes place, for instance *'quickly,' 'sexually,' 'by splitting in two,'* or *'on their own.'*

4.1.1.3 Frequently occurring words in Band 3

The third most frequently used word families, which appear at least 15 times in the BIOCOR, constitute Band 3 (see Table 26).

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

Table 26 Band 3: the third most frequent lexical items in the BIOCOR

Similarly to Band 2, this band scarcely contains biology terms, there being only three, such as '*malaria,' 'blood,'* and *'tapeworm.'* The appearance of general English vocabulary, however, is four times as abundant as the use of biology terms in this band. Most of the general English lexis in Band 3 are part of basic vocabulary, such as *'see,' 'use,' 'cause,' 'move,' 'place,' 'shape,'* and *'water.'* It is only a small part of the general English lexical items here that are related to biology topics, for instance *'substance,' 'disease,'* and *'mosquito.'* Not differing from Bands 1 and 2, this band does not contain a single academic English lexical item either. The complete lack of academic English is highly unexpected of

the register, biology textbooks are supposed to use a large number of academic English vocabulary by both the biology teachers and the bilingual students of the school (Cserép, 1997).

The variety of the use of the eighth most recurring biology term, '*malaria*,' is rather limited in the BIOCOR (see Table 27). In an adjective form ('*malarial'*) it combines with nouns, such as *'area'* and *'parasite,'* in addition, it also takes the negative prefix *'anti'* to form the collocation *'anti-malarial tablet.'* The range of verbs it collocates with is extremely narrow; apparently, there is no more than one single combination in the BIOCOR with the verb '*occur.'* The scope of the token to function as an object of a verb is wider, there are four such instances, namely, *'conquer,' 'get,' 'have,'* and *'carry.'* The passive voice is also typical with the lexical item, it is used in combination with *'malaria is controlled,' 'be cured of malaria,'* and *'malaria is spread by mosquitos'* in the BIOCOR.

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

Table 27 Lexical environment of the biology term '*malaria'* in the BIOCOR

The ninth most frequently applied biology term, '*blood*,' appears to be bounded in its use in the BIOCOR (see Table 28). The token is used in a varied manner in noun phrases; it appears in *'blood-sucking tsetse,' 'blood system,' 'dorsal blood vessel,'* and *'red blood cell'* among others. However, the lexical item rarely combines with verbs. There is no instance of it taking a verb in the BIOCOR at all; however, two verbs take it as an object. The verb *'cause'*

collocates with its gerund form, and the phrasal verb *'suck up'* also appears together with the token. Similarly scant is the choice of verbs in the passive voice it collocates with, there is no other such instance but *'blood is pumped by the heart.'*

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

Table 28 Lexical environment of the biology term '*blood'* in the BIOCOR

The 10th most typical biology term in the texts, '*tapeworm*,' shows a similarly restricted selection of collocations (see Table 29). The area where it forms collocations multifariously is the noun phrase. As a noun phrase, it demonstrates a wide range of combinations, for instance *'beef tapeworm,' 'life cycle of the tapeworm,'* and *'tapeworm bladder.'* Its tendency to combine with verbs is not that diverse, however. There are no more than two examples of it taking a verb; it collocates with the phrasal verbs *'pop out'* and *'get round.'* The token's use as an object of a verb is even more restricted, no other verb but '*get'* takes it.

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

Table 29 Lexical environment of the biology term '*tapeworm'* in the BIOCOR

Biology terms and academic English vocabulary that appear fewer than 15 times in the corpus were also collected. However, as the dissertation regulations define a dearth of plethoric space, these items are not recorded here in the running text, and thus their lexical environments are not presented either. For a list of specific lexical items, biology terms and academic English vocabulary, which appear minimum four times in the biology corpus, see Appendix I and J respectively. It is worth noting, however, that only 34 biology terms and no more than 13 academic English lemmas were found in the entire BIOCOR. This shows that the BIOCOR does not abound in specific lexis; it is the general English lexis that is massively present in the biology texts. It can be claimed that academic English is extremely rare in the corpus, moreover, at a more recurrent level totally absent in its language use. The most frequently occurring biology terms (the ones that appear more than thirty times in the corpus), show a wide range of collocations. However, biology terms that are used less frequently than 30 times in the corpus display a much more limited, less diverse scope of lexical combination. The reason for the lack of abundant use of specific lexis (either biology terms or academic English) might be the fact that biology textbooks for secondary school students are written for non-experts of the field, in which sense these textbooks are more popularizing than academic. Shapiro (2012) emphasises that science textbooks tend to be read by non-scientists, moreover, by young audiences whose majority does not even incline to become scientists. For this reason, he argues that pre-college level science textbooks form a subset of literature popularizing sciences. As the authors of secondary school science textbooks attempts to connect sciences with the non-professional needs and experiences of a community much larger than the discourse community of scientists, less technical language is preferred in such textbooks.

Since the BIOCOR can hardly be characterized by the profuse application of specific lexis different from general English, the difficulties the 10th grade bilingual students face when processing the biology texts cannot be attributed to the abundance of unfamiliar biology terms or academic English vocabulary. Moreover, the main reason for the bilingual students' finding the biology texts difficult to handle can barely be recognized in the texts' specific biology terminology as many of the anyway small number of specific biology vocabulary items are similar in the students' mother tongue, in Hungarian. The obvious similarity in the two languages with respect to the Greek and Latin origin biology terms in English in the BIOCOR qualifies even monolingually raised students to understand these terms without hesitation and with great certainty. Consider the similarities in the cases of, for example, *parasite (parazita), bacteria (baktérium), virus (vírus), amoeba (amőba), malaria (malária), microscope (mikroszkóp), cytoplasm (citoplazma), nucleus (nukleusz), photosynthesis (fotoszintézis), membrane (membrán), spore (spóra) or chlorophyll (klorofil).* Consequently, to reveal which characteristic features of the biology corpus might pose difficulties for the tenth grade students when they attempt to process the biology texts further steps of investigation are needed. Accordingly, let us now turn our attention to those lexical items in the BIOCOR which might not be the most frequently occurring word families but are markedly different in their frequency when compared to the REFCOR, the general English reading tasks the 9th graders are assigned to process.

4.1.2 Keyness

The lexical uniqueness of a corpus can be effectively described by keyness values, which compare the frequency of lexical items in the corpus with that in the reference corpus (Xiao & McEnery, 2005). The across-register nature of the method allows for the comparison of two registers, for pinpointing lexical characteristics that distinguish one register from

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

4.1.2.1 Positive keyness

The biology register is described here through listing lemmatized key words in their order of outstandingness. The key words are organized in three categories: biology terms, academic English and general English lexis (see Table 30; for the guiding principles of categorizing lexis into these three categories return to Section 3.3.1.1 on p. 73). The correlation between lexical items with significantly high keyness values and their level of frequency in the BIOCOR was examined (displayed in Table 30), and the particular frequency bands are also indicated (for the methods of developing 10 frequency bands see Section 3.3.1.1 on pp. 73-79). Finally, the lexical environments of the biology key words are also uncovered.

Key word	Keyness (k value)	Type	Raw frequency	Band
bacteria	136,6856537	biology term	41	1
virus	83,84221649	biology term	34	1
tapeworm	72,29248047	biology term	18	3
parasite	68,68048096	biology term	57	1
body	68,47974396	general English	35	1
mosquito	60,57646179	general English	17	3
amoeba	60,50664139	biology term	20	2
plant	54,08612823	general English	40	1
organism	53,18547821	general English	55	1
name	48,05667114	general English	32	1
call	45,21485138	general English	62	1
substance	43,9251976	general English	19	3
cell	41,82590866	biology term	51	1
malaria	40,34447479	biology term	19	3
host	38,4834137	biology term	8	6
live	36,06760406	general English	55	1
group	35,42422485	general English	11	5
figure	33,97449493	general English	39	1
segment	33,66395569	biology term	13	4
worm	33,66395569	general English	20	2
key	33,58996582	general English	11	5
thing	32,68690109	general English	25	2
water	32,53123093	general English	15	3
genus	30,98904419	biology term	8	6
process	30,39152718	academic English	11	5
food	30,23670197	general English	46	1
blood	30,08574486	biology term	19	3

intestine	28,84708214	biology term	7	7
drug	28,22444725	biology term	7	7
energy	27,01064491	general English	8	6
gut	26,9503727	biology term	11	5
earthworm	26,9503727	general English	7	7
cavity	26,9503727	general English	5	9
John	26,2686615	general English	8	6
agar	24,25426292	biology term	6	8
sickness	24,18762207	general English	6	8
sleep	24,18762207	general English	8	6
disease	24,18762207	general English	17	3
bloodstream	24,18762207	general English	6	8
egg	24,03279495	general English	14	4

Table 30 Key words and their frequency in the BIOCOR

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

The lexical environments of the eight biology key words that belong to the first three frequency bands *('bacteria,' 'virus,' 'tapeworm,' 'parasite,' 'amoeba,' 'cell,' 'malaria,'* and *'blood')* have been described in the previous section (4.1.1 on p. 118), thus they are not repeated here. The highest keyness value token among the less frequently appearing lemmas, '*host*' (k=38), shows a narrow range of word combinations (see Table 31). It tends to form noun phrases, such as *'intermediate host,'* or more typically genitive constructions, *'the host's digestive food,'* or *'the host's digestive juices,'* and *'the host's faeces.'* Even greater scarcity is displayed by the verbs it combines with, the single example of the verb with which it appears together is '*carry.'*

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

Table 31 Lexical environment of the biology term '*host'* in the BIOCOR

The second highest keyness value word among the less frequent biology terms, '*segment*' (k=34), combines in a rich manner (see Table 32).

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

Table 32 Lexical environment of the biology term '*segment'* in the BIOCOR

It appears in various noun phrases, such as *'gut segments,'* or *'mature segments'* and shows an even more diverse set of verbs it collocates with (e.g., *'pass a segment,' 'produce segments,' 'segments drop off,'* or *'segments mate'*). The token does not disagree with the passive voice either, even if the BIOCOR displays no more than one single example of it *('the body is divided up into a series of segments')*. The lemma also shows the possibility of being combined in a prepositional phrase (e.g.,*'in each segment').*

The lemma '*genus,'* with the third highest keyness value (k=31), shows a rather scarce variety of collocations. It combines only in noun phrases and verb phrases (see Table 33). There is one single noun with which it goes together in the BIOCOR *('name')*. In a similar fashion, neither is the number of verbs it collocates with more numerous, since it is used in no more than one verb collocation, with the verb *'belong.'*

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

Table 33 Lexical environment of the biology term '*genus'* in the BIOCOR

The token *'intestine,'* with an outstandingly high keyness value (k=29), forms word combinations within a narrow range (see Table 34). It appears in noun phrases which refer either to its type, *'large intestine'* or *'small intestine,'* or to its structure *'wall of the intestine.'* The number of verbs it combines with is even less manifold; the lemma appears only within one verb phrase, *'live in the intestine.'*

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

Table 34 Lexical environment of the biology term '*intestine'* in the BIOCOR

The next significantly high keyness value item (k=28), *'drug,'* is applied in the BIOCOR in a slightly more versatile way (see Table 35). It forms verb combinations both in the active voice *('drugs save lives')* and in the passive voice *('drugs are taken' and 'be treated with certain drugs')*. Also, the lemma is capable of forming an adjective phrase with *'resistant.'*

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

Table 35 Lexical environment of the biology term '*drugs'* in the BIOCOR

The lemma *'gut,'* with a high keyness value (k=27), shows a diverse set of lexical collocations (see Table 36) in the BIOCOR. It appear in various noun phrases (e.g., *'human gut'* and *'gut wall'*) and verb phrases alike *('the gut has a special region,'* or *'gut segments contain')*. Besides, the token is also used as a reference of location in prepositional phrases, such as *'above the gut'* or *'beneath the gut.'*

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

Table 36 Lexical environment of the biology term '*gut'* in the BIOCOR

The last lemma with a significantly high keyness value (k=24), *'agar,'* appears in relatively few combinations (see Table 37) in the BIOCOR.

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

Table 37 Lexical environment of the biology term '*agar'* in the BIOCOR

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

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

4.1.2.2 Negative keyness

Lemmas with high negative keyness value reveal words which are systematically untypical in a particular register compared to a reference corpus. In the present study, lemmas with high negative keyness value show the set of lexical items which occur in the REFCOR but are significantly less often used in the BIOCOR. In other words, tokens with high negative keyness value shed light on a special group of words which 9th grade students process during their general English studies: it is the collection of word families which are underrepresented

(or not present at all) in the biology texts the students read the following term. The findings of running the keyword application of WordSmith version 5 (Scott, 2008) strikingly show that the BIOCOR contains no such item. Notably, there is not one single lemma in the BIOCOR with significantly high negative keyness value when compared to the REFCOR.

4.1.2.3 High-frequency low-keyness words

It is not insignificant to take note of the fact that the BIOCOR encompasses great many frequently occurring lemmas that do not appear among the word families with high-keyness value (for en extensive list of words frequently applied in the BIOCOR turn to Section 4.1.1 on p. 118). This group of words, the set of high-frequency low-keyness items, show that the majority of the frequently used lexis of the BIOCOR is present in the REFCOR with a similar rate of frequency. Table 38 displays the collection of all these words, shows each item's frequency expressed in frequency bands, as well as the type of the lexical item (biology term, academic English or general English).

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

Table 38 High-frequency low-keyness words in the BIOCOR

It can clearly be seen that the group of high-frequency low-keyness words embraces nearly exclusively general English terms; only two instances of biology terms occur *('growth'* and *'reproduce')* and there are no academic terms at all.

The keyness characteristics of the BIOCOR provide revealing information about the register of the biology textbook for secondary school students. First, the results uncover that there is a nearly absolute scarcity of academic words among the key words. That is, the lexis of the BIOCOR can hardly be distinguished from that of the REFCOR on account of the use of academic English terms. This finding goes contrary to the expectations of the biology teachers and the students of the bilingual programme alike, who expressed their certainty about the biology texts being abundant in academic vocabulary, which makes the register of biology texts starkly different form other registers in their perception (Cserép, 1997). Second, the great majority of biology key words appear with high frequency in the BIOCOR. This indicates that the 10th grade students are expected to read a string of texts which contains recurrently repeated biology key words. In other words, the students' difficulty of processing the biology texts is hard to be accounted for by the students' unfamiliarity with the biology lexis due to the sporadic appearance of the specific lexis. Third, the extensive use of the proper noun *'John'* to such a great extent that it surprisingly appears among the key words of the BIOCOR demonstrates that the register intends to clarify the subject information it conveys more through practical examples than through highbrow, scholarly theoretical lines of thought. This tendency is in line with Shapiro's (2012) findings highlighting the fact that the register of science textbooks written for non-experts is more popularizing than academic; its language is less technical than that used in the discourse community of scientists. Fourth, the BIOCOR contains no lemmas with high negative keyness value at all, in other words, there are no lexical items which occur significantly less often in the BIOCOR than in the

REFCOR. That is, the register of the biology texts cannot be distinguished from the REFCOR from this respect, there are no significantly underrepresented general English lexical items. From the point of view of the ESL teacher, it signifies that the vocabulary of the general English reading tasks assigned in the 9th grade cannot be characterized by a superfluously expanded vocabulary in comparison with the lexis used in the biology texts. Finally, the BIOCOR can be characterized by a bounteous use of not register specific frequently occurring words, which are also frequently present in the REFCOR. This indicates that by the time students start pursuing their biology studies in the 10th grade they have already become familiar with a great part of the lexis of the BIOCOR through reading the texts of the REFCOR in the 9th grade. Considering all these five aspects of the keyness results, the lexis of the BIOCOR can hardly be described as challenging for 10th grade bilingual students. Let us then turn our attention from the BIOCOR's keyness characteristics to is lexical density features, a different lexical point of view which might give a reasonable explanation for the difficulties 10th grade students face when progressing biology texts.

4.1.3 Lexical density

Since the ratio of lexical items is not the same across registers (Halliday, 1985a), the distinctiveness of a register can be analysed in terms of lexical density. Written English applies significantly more lexical items in relation to grammatical tokens than spoken English (Halliday, 1985a; Ure, 1971). The watershed between them is expected to be at about 40%, written English scoring above that lexical density, while spoken English below 40% . Among the written registers, the ones that use formal language and whose information content is high tend to have an ever higher lexical density of approximately 60-70% (Kormos & Csölle, 2004). Written registers with thin lexical density (40-60%) at the same time reveal a low level of information packaging (Johansson, 2008). The present section explores if the academic

register of the biology texts for secondary students contains more lexical words than general English texts do.

Following Ure's (1971) formula, the lexical density of the BIOCOR and that of the REFCOR was counted as the ratio of the lexical words compared to the total number of words in the entirety of the two corpora respectively (see Table 39). The ratio of these figures in both corpora produces the percentage of open class words, the ones which convey information in both registers.

	Number of lexical words:	**Total number of words in the corpus:**	**Lexical density** (expressed in percentage)
The BIOCOR	3632	7,012	51.79692% = 52%
The REFCOR	3569	7098	50.41672% = 50%

Table 39 The lexical density of the BIOCOR and that of the REFCOR

The results uncover that the BIOCOR has a lexical density of 52% and the REFCOR has a nearly identical lexical density of 50%. Both results are consistent with the expected lexical density of written registers since both figures are above the 40% watershed (Kormos & Csölle, 2004). However, neither of the corpora shows traits of high lexical density (of about 60-70%), which is characteristic of registers that apply formal language and convey massive bulks of information. Despite the fact that informative, academic written texts are typically expected to have a higher lexical density than non-academic written registers of everyday topics (Johansson, 2008), this does not hold true for the academic BIOCOR. The 52% lexical density of the BIOCOR implies that the string of biology texts 10th graders read do not use formal language – either compared to other registers of formal language use in general (Kormos & Csölle, 2004) or contrasted with the REFCOR in particular. The value of the lexical density of the BIOCOR also suggests that the biology register is not more informative than the REFCOR. This lexical density result indicates that the 10th grade students' difficulty

of processing the BIOCOR cannot be caused by the high level of formality of the register, since it is not exceptionally high; moreover, it is fairly similar to that of the 9th grade's reading tasks. In a parallel manner, the difficulty of processing the BIOCOR can neither be explained by the texts' being too informatively packed with subject material, since its information content value is not dissimilar from that of the REFCOR. In other words, the register of the biology texts 10th grade students are exposed to read is not more difficult to comprehend from this respect than general English texts.

Although investigating the presence of particular lexical categories lay out of the direct focal point of the research, when summing up the tokens of the open class, autosemantic words, some revealing pieces of information naturally emerged about the characteristic language use of the BIOCOR. It is conspicuous from the data that some of the nominal lexical categories are underrepresented in the BIOCOR compared to the REFCOR, while other categories' frequency is heavier in the BIOCOR (see Table 40). The class of comparative adjectives abounds in the BIOCOR, lexical items which take the second degree of comparison appear over two times more frequently there than in the REFCOR of general English texts. Examining the functions of the excessive use of comparative adjectives uncovers that they are applied for four major reasons. Comparisons are typically used in definitions (e.g., *'smaller groups are called phyla'*), in explanations (e.g., *'viruses are simpler than any other organisms'*), in circumscriptions which avoid the of naming of certain objects in particular (e.g., *'the disease is caused by organisms which are smaller than bacteria' or 'the region where the skin is thicker than in other places'*) and in processes (e.g., *'gets larger, heavier'*) in the BIOCOR. In contrast, other lexical categories are used considerably more sparingly in the BIOCOR than in the REFCOR. Proper nouns, for example, are used nearly six times less frequently in the BIOCOR. This result can be explained by the fact that the

biology textbook is written for non-expert secondary students, that is, the target audience requires already established, widely-accepted pieces of knowledge in the field, which is not typically referenced by the names of scientists. Despite the fact that biology is an academic subject, its textbook at a pre-college level shows no signs of introducing current experiments and developments anchored by clear referencing in the field, which is otherwise distinctively represented in tertiary level academic writing (Jordan, 2002; Bailey, 2011; Chazal & Moore, 2013). Nouns that are neutral for number appear nearly two times less often in the BIOCOR than in the REFCOR. This dramatic difference might be the result of the REFCOR intentionally applying such nouns in an excessive number. Namely, the REFCOR is string of texts which was designed for ESOL students, who might be thought of in need of this particular lexico-grammatical practice, which is a problematic issue at level B2 (Vince & Emmerson, 2003). While the BIOCOR has no intention of providing the reader with a practice of nouns neutral for number in the form of reading input. Additionally, interjections, which appear in a relatively small number in the REFCOR, are not present in the BIOCOR at all. This result implies that the biology textbook avoids the use of emotional language and aims at neutral, objective ways of expressions. The frequency of other lexical categories, such as base adjectives, superlative adjectives, adverbs, singular and plural nouns shows balanced similarities between the two corpora (see Table 40).

Lexical category	Raw frequency in the BIOCOR (number of items)	Raw frequency in the REFCOR (number of items)
Comparative adjective	22	10
Proper noun	44	247
Noun neutral for number	36	63
Interjection	0	2
Base adjective	530	534
Superlative adjective	8	9
Adverb	346	356
Singular noun	1189	996
Plural noun	570	556

Table 40 The frequency of lexical categories in the BICOR and the REFCOR

Summing up the results of the three subsections of the lexical component of the analysis (frequently occurring words, keyness and lexical density), the BIOCOR is ready to be described as completely lacking any traits which could pose significant lexical challenges for 10th grade bilingual students. As the difficulty of processing the BIOCOR in the 10th grade cannot be explained satisfactorily at a lexical level, let us turn our attention to the sentence-level complexity of the corpus in order to see if those features give firmer grounds for the perceived difficulties of the texts.

4.2 Grammatical phenomena

The level of complexity of grammar structures can account for the difficulty of processing a text. This section of the dissertation discusses if the grammar used in the BIOCOR is significantly different from that of the REFCOR, and whether consequently the textbook abounds in grammatical phenomena which are challenging at B2 level. The results of this examination are provided in a thick, comparative description along the groups of grammatical phenomena of the grammatical component of the POTAI:

1. tenses and tense related structures;
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. modal verbs.

4.2.1 Tenses and tense related structures

Among the tense aspect of comparison, the frequency of fourteen linguistic features was examined. Considering the frequency of **present simple tense**, there is a significant difference (p=.016) for the BIOCOR (*M*=1.15) and the REFCOR (*M*=.83). The results show that there are significantly more instances of using the present simple tense in the BIOCOR than in the REFCOR. On average, present simple appears in every single sentence in biology texts, more precisely, there are 8 present simple verbs in 7 sentences; while general English texts contain fewer present simple items than sentences, as there are 5 present simple items in every 6 sentence.

The difference in the use of the **present continuous tense**, however, is not significant (p=.05) in the two registers. As the probability coefficient of present continuous tense is not lower than 5 per cent, the mean values of their frequency for the BIOCOR (*M*=.007) and for the REFCOR (*M*=.037) cannot be used for describing the sample in a generalizable way. That is, the fact that there are five times as many present continuous items in the REFCOR than in the biology texts (present continuous appearing once in every 143 sentence in the biology texts and once in every 27 sentence in the general English texts) should not be generalised and claimed to be a descriptive fact of the register of biology texts; however, it is true for the sample under investigation only.

As for the frequency of the **past simple tense**, there is a significant difference (p=.03) between the two registers. The BIOCOR (*M*=.091) tends to use the past simple three times more often than the REFCOR (*M*=.365). An item of past simple appears in each and every sentence in the BIOCOR, while it is only about every third sentence in the REFCOR that uses the past simple (to be more precise, every fourth in eleven sentences).

Examining the frequency of the use of the **past continuous tense**, it can be stated that their difference is not significant in the two registers (p=.19). Although there are no items in the past continuous in the BIOCOR, it cannot be generally claimed that biology texts apply no such tense, as the lack of significant difference prevents generalisations about the register. In a similar manner, neither can the frequency of past continuous tense be determined in general English texts based on the sample, where it is only every 111th sentence that contains the past continuous (*M*=.009).

The use of the **past perfect simple** in the BIOCOR (*M*=.006) and the REFCOR (*M*=.016) is not significantly different either (p=.29). Consequently, the fact that the BIOCOR uses three times fewer verbs in the past perfect simple than the REFCOR is an account true for the sample only; however, it cannot be given as a description of the register of biology texts in general.

The past perfect continuous is one of the four tenses that is not present in either of the corpora. As there is no appearance of the tense in the BIOCOR or in the REFCOR, obviously no significance and mean values could be counted. The lack of statistical computations prevents drawing generalizable conclusions about the register of biology texts from this respect, but it seems to be clear that there is no tendency of past perfect continuous being frequently used in either of the registers.

Giving an account of the frequency of the *'used to'* structure, the difference between the two registers is not significant (p=.34). Similarly to the lack of appearance of the past continuous tense in the BIOCOR, it can be claimed that no 'used to' items are present in the biology texts described; however, no generalisation can be made that the register of biology

texts does not contain the structure owing to its probability coefficient being far too high. Likewise, neither can the presence of the 'used to' structure in every 500th sentence in the REFCOR be generalised (*M*=.002).

Considering the frequency of the use of the **present perfect simple**, there is a significant difference in the two registers (p=.03). The BIOCOR applies nearly six times fewer present perfect items (*M*=.016) than the REFCOR (*M*=.093), which can be generally stated about the registers of the biology texts.

In contrast, the frequency of the use of the **present perfect continuous** bears no significance (p=.37) in the two registers. As the probability coefficient is not low enough to generalize the results of the sample, the fact that the BIOCOR (*M*=.001) uses seven times fewer present perfect continuous items than the REFCOR (*M*=.007) is a sample-specific characteristic feature. The reason behind might be that both registers use present perfect continuous sparingly (the BIOCOR in every 1000th sentence, the REFCOR in every 143 sentence), which makes the registers undistinguishably similar in this respect.

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

Just as the past perfect continuous is not represented in the texts by one single item, neither of the registers makes use of other 'will' future tenses besides future simple, that is,

the **future continuous, the future perfect simple**, and the **future perfect continuous** are not present in any of the texts. Evidently, the complete lack of their use in the corpora makes statistical operations impossible, which leaves no space for generalizable description in these respects. Correspondingly, the BIOCOR does not show tendencies with statistical certainty in these regards; however, it simply implies an extreme underuse of complex future forms.

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

4.2.2 Conditional structures

The second group of grammatical phenomena examined in the grammatical component of the POTAI was conditional structures, such as zero, first, second, third, and mixed conditionals. The range of these phenomena shows no significant difference between the two registers at all, since the probability coefficient, indicating the percent of coincidence in the two corpora, is way above 5% in all four cases. It is 15% (p=.15) in the case of zero conditionals, 32% (p=.32) in that of first conditionals, 33% (p=.33) for second conditionals, and 34% (p=.34) in the case of third conditionals. While the fifth type, the mixed conditional structure, is not present in any of the texts of the two corpora. Due to the lack of significant differences, the frequency results of the conditional structures are true for the BIOCOR and REFCOR samples only; however, they cannot be regarded as descriptive ratios of the register of biology texts for secondary students in general. **Zero conditional** structures appear two times more often in the BIOCOR (*M*=.03) than in the REFCOR (*M*=.012), there is a zero

conditional in ever 37th sentence in the BIOCOR, while it is only ever 81st sentence in a REFCOR that uses the structure. In contrast, **first conditional** structures are less typical of the BIOCOR (*M*=.005) than the REFCOR (*M*=.016). The first conditional appears more than three times less frequently in the BIOCOR, where the structure is used in ever 217th sentence, while it appears in the REFCOR in every 64th sentence. In a similar manner, the frequency of the **second conditional** structures is lower in the BIOCOR (*M*=.005) than in the REFCOR (*M*=.013). The sample uses the second conditional structures twice as rarely in the BICOR, in every 185th sentence, as in the REFCOR, where it appears in every 79th sentence. An even less frequently used hypothetical structure in the two corpora is the **third conditional**, which appears in every 417th sentence in the REFCOR (*M*=.002), while it is not represented in the BIOCOR at all. The absolute lack of **mixed conditional** structures in either of the REGISTERS indicates no salient importance of connecting hypothetical past and present in the samples.

4.2.3 Passive voice and causative structures

Subsequently, the passive voice, both with a direct and an indirect object, along with the causative structures such as *'have it done', 'get it done', 'needs doing',* and *'make somebody do something'* were examined in the two corpora. 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 registers 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 registers. Discussing the frequency of the **passive voice with a direct object** in the two corpora, it can be claimed that their probability coefficient (p=.053) is high, however slightly (.003), for the texts to be considered significantly different, consequently no generalisations can be made

about the register of the biology texts for secondary students based on this aspect of comparison. The corpora, however, indicate that the BIOCOR (*M*=.253) uses 1.5 times more passive items with a direct object than the REFCOR (*M*=.177). The BIOCOR applies the passive voice in every fourth sentence, while it is only every sixth sentence in the REFCOR that employs it.

4.2.4 Relative clauses

The group of grammatical phenomena labelled as relative clauses gave space to the comparison of the two registers 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. Considering the probability coefficient of the frequency of **defining relative clauses with a relative pronoun** (p=.065), it shall be noted that the two registers are not significantly different, even if the BIOCOR (*M*=.117) contains 1.5 times more such clauses than the REFCOR (*M*=.079). The BIOCOR applies defining relative clauses with a relative pronoun in every 8th sentence, while it is present in only every 13th sentence in the REFCOR. In contrast, the frequency of **defining relative clauses without a relative pronoun** shows a significant difference (p=.048) between the two registers. Therefore it can be generalised and asserted that the above grammatical item appears in the register of biology texts (*M*=.006) nearly six times less often than in general English texts (*M*=.034). The BIOCOR contains the grammatical item in every 167th sentence, while it appears more heavily, in every 29th sentence in REFCOR. Likewise, the probability coefficient of **non-defining relative clauses** indicates a significant difference (p=.03) between the two registers. It follows that the appearance of the grammatical item can be argued to be one of the characteristic features of the register of biology texts for secondary students. Non-defining relative clauses appear nearly seven times fewer in the BIOCOR

(*M*=.009) (one item in every 111th sentence) than in the REFCOR (*M*=.06) (one item in every 17th sentence). On the contrary, the significance of **progressive participle clauses** (p=.765) is far too high to be generalized, as a result, no general characteristics of the register of biology texts can be given in this respect. The fact that frequency of the progressive participle clauses is apparently the same for the BIOCOR (*M*=.0412) and for the REFCOR (*M*=.046) is a sample-specific description, not generalizable for larger corpora. In contrast, **progressive participles in the past** are not present in either of the registers, which makes statistical analysis impossible from this respect. Based on the two corpora, what can be stated with certainty is that progressive participles in the past are of no primary importance for the register of biology texts and general English texts alike. Similarly to progressive participle clauses, the frequency of **simple participle clauses** is not significantly different in the two registers (p=.956). Thus the fact that the texts contain nearly identical number of simple participle clauses, one in every 23rd sentence (*M*=.044 for the BIOCOR and *M*=.043 for the REFCOR), is a sample-specific observation, which cannot be generalised about the register of biology texts for secondary students. In contrast, **passive progressive participles**, either in the present or in the past, cannot be traced in any of the corpora, thus these reduced relative clauses cannot be analysed by statistical means. As a consequence, the frequency of their use in either of the registers cannot be demonstrated statistically; however, the lack of their importance in the two registers can be expected with high certainty.

4.2.5 Nominal relative clauses

Describing the BIOCOR from the point of view of the frequency of nominal relative clauses, it can be affirmed that there are no significant differences between the two registers, since all the probability coefficients are too high (above five per cent) to be generalizable. From this it follows that the description of the two corpora is sample specific in terms of

nominal relative clauses, that is, the differences found are not register specific. **Nominal relative clauses without a reporting verb without time shift** (p=.95) occur rather often in both corpora (*M*=4.5 for the BIOCOR and *M*=4.412 for the REFCOR). In contrast, **nominal relative clauses without a reporting verb with time shift** appear much less frequently (p=.12), particularly, approximately once about every second sentence in the REFCOR (*M*=.583), whereas there are no such appearances in the BIOCOR at all. On the contrary, **nominal relative clauses without a reporting verb with an infinite verb** are used with a modest frequency in the corpora (p=.77). Such clauses appear twice in each sentence in all the texts of the corpora on average (*M*=2 for the BIOCOR and *M*=2.375 for the REFCOR). Yet the frequency of **nominal relative clauses without a reporting verb with a preparatory 'it'** is considerably lower (p=.79), they tend to appear once in every eighth sentence in the BIOCOR (*M*=.125) and approximately once in every sentence in the REFCOR (*M*=.083). Reported speech without time shift, that is, **nominal relative clauses with a reporting verb without time shift,** appear in both corpora (p=.43), once in every third sentence in the REFCOR (*M*=.333) and nearly three times less frequently in the BIOCOR (*M*=.125). Whereas 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 corpora. 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 BIOCOR, however, they are used in the REFOCR (p=.34) with considerable frequency. It is present six times in five sentences (*M*=.083) in the general English texts. The frequency result 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 REFCOR (*M*=.25) as in the BIOCOR (*M*=.125). They are present in every fourth sentence in the REFCOR, while only in every eighth sentence in the BIOCOR. 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 corpora. It shall be noted, however, that all the frequency ratios of nominal relative clauses are characteristics of the BIOCOR, not those of the registers of biology texts for secondary students in general, as no significant differences could be traced in these respects.

4.2.6 Infinitives

The group of grammatical phenomena of infinitives comprises the analysis of the frequency of simple, progressive, active and passive forms of infinite verbs in the corpora. As the probability coefficient of **simple infinitive** (p=.46) shows no significant difference between the two registers, its frequency description does not give space for making generalisations. As a result, the fact that both registers contain an immense number of infinitives, 19 simple infinitives in each sentence in the BIOCOR (*M*=18.75), and 16 items in every sentence in the REFCOR (*M*=15.667), cannot be stated about the register of biology texts in general, however, shall be treated as a sample specific trait. On the other hand, analysing the results of the frequency of **passive infinitives** (p=.04) shows a significant difference between the two registers. The BIOCOR (*M*=3) applies three times more passive infinitives than that of REFCOR (*M*=1.08). Passive infinitives appear approximately three times in each sentence in the BIOCOR, while only once in a sentence in the REFCOR. In contrast, the use of **progressive infinitives** cannot be generalised for the registers, since the probability coefficient (p=.34) is too high to yield other than sample specific results. The BIOCOR contains no progressive infinitives at all; however, it appears in every fourth sentence in the REFCOR (*M*=.25). On the contrary, **progressive passive infinitives** cannot be identified in the REFOCR, while they are used in the BIOCOR in every fourth sentence (*M*=.25). However, the difference between the two registers is not significant (p=.23), that is, the fact that the use of progressive passive infinitives is considerably more frequent in the

BIOCOR is a sample specific description. Similarly to the frequency of progressive infinitives, **perfect infinitives** are not present in the BIOCOR at all, while they can be identified in every one and a half sentence in the REFCOR (*M*=.667). This difference is, however, not significantly different in the two registers (p=.074), thus it cannot be generalised about the register of biology texts for secondary students either. Based on the results, it can be argued that more complex infinitives do not typically appear in either of the registers, neither **perfect passive infinitives,** nor **perfect progressive infinitives** or **perfect progressive passive infinitives** can be identified in any of the corpora.

4.2.7 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 two corpora; however, none of them showed significant differences in the two registers. Moreover, none of these grammatical constructions are present in the BIOCOR; for which reason it can be claimed that the register of biology texts is highly unlikely to contain prepositions at the end of sentences. In contrast, prepositions tend to appear at the end of sentences in the REFOCR in every twelfth sentence **in questions** or **with an infinitive** (in both cases p=.34 and *M*=.083) and in every sixth sentence **in relative clauses** (p=.25 and *M*=.167). The difference between the two registers, however, cannot be generalised for the significantly lack of low probability coefficients.

4.2.8 Modal auxiliaries

Among the group of grammatical phenomena of modal verbs, forty-one different modalities were examined in the two corpora. 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 modal verbs, however, the probability coefficient is low enough (smaller than five per cent) to indicate a significant difference between the two registers. First, the frequency of the use of *'can' expressing ability in the present* is register specific for the BIOCOR (p=.013). It tends to appear extensively in the register of biology texts for secondary students, more than five times in each sentence on average (*M*=5.5), while its appearance in the REFCOR is half as massive as that. It is used approximately two times in each sentence in the REFCOR (*M*=5.5). Secondly, the frequency of *'may' expressing the level of certainty in the present* shows an account typical of the register of biology texts for secondary students (p=.038). This modal verb appears three times in two sentences on average in the BIOCOR (*M*=1.25), while far more scarcely in the REFCOR, where it appears only once in every second sentence (*M*=.5). Finally, it is the frequency of *obligation in the present expressed by 'must'* that differs significantly in the two registers (p=.028). The significant dissimilarity lies in the fact that the register of biology texts uses this modal auxiliary three times in two sentences (*M*=1.25), while it makes no appearance in the REFCOR at all.

Besides the above three modal verbs, no other modal auxiliary can be described as register-specific in the two corpora 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), which is twice as high in the BIOCOR (*M*=.375) than in the REFCOR (*M*=.167) (appearing once in every third and sixth sentences respectively), describes the BIOCOR, and

not the register of biology texts in general. In a similar manner, the frequency of the modal verb *'could' expressing ability in the past* (p=.66) describes the BIOCOR. It is present approximately twice in three sentences in the BIOCOR (*M*=.625), while appears twice in five sentences in the REFCOR (*M*=.417). In contrast, the modal auxiliary *'able to' expressing ability in the past* does not appear in the BIOCOR at all, while the REFCOR 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 BIOCOR, while they appear in the REFCOR 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). They appear in both corpora five times in six sentences (*M*=.833 for the BIOCOR and *M*=.875 for the REFCOR). In contrast, the auxiliary *'will" expressing the level of certainty in the present* cannot be found in the BIOCOR, while it is present once in every fourth sentence in the REFCOR (p=.34 and *M*=.25). Similarly, the modal verb *'should' expressing the level of certainty in the present* cannot be identified in the BIOCOR, however, it appears once in every second sentence in the REFCOR (p=.26 and *M*=.5). The modal auxiliaries *'might' expressing the level of certainty in the present* and *'should' expressing an obligation in the present* shows a four times more frequent use in the BIOCOR than in the REFCOR (p=.12). It appears once in every third sentence in the BIOCOR (*M*=.375), yet only once in every twelfth sentence in the REFCOR (*M*=.083). *Obligations in the present* and *in the past expressed by 'have to'* and *'had to'* are both less frequent in the BIOCOR. The first one is twice as scarce in the BIOCOR as in the REFCOR (p=.5, *M*=.125 and *M*=.25), it appears once in every eighth and fourth sentence correspondingly. The second one is used nearly four times less often in the corpora, it appears once in every eighth sentence in the

BIOCOR (p=.34, *M*=.125), while only once in about every second sentence in the REFCOR (*M*=.417). *Obligations expressed by 'to be to' in the present* and *in the past* are present in the REFCOR with the same frequency, appearing once in every sixth sentence (*M*=.167). However, they appear slightly more frequently in the BIOCOR in the present, once in every fourth sentence (p=.68, *M*=.25), while not at all in the past in the BIOCOR (p=.25). In contrast, the *obligation expressed by 'need' in the present* is three times more frequent in the BIOCOR than in the REFCOR (p=.38, *M*=.25). It appears once in every fourth sentence in the BIOCOR and only once in every twelfth sentence in the REFCOR (*M*=.083).

A considerable number of auxiliaries, 20 in particular, could not be identified in any of the corpora. The modals which are not use in either registers are as follows: 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.'*

4.2.9 Overview of the results of the analysis based on the grammatical component of the POTAI

Taking all the eight groups of grammatical phenomena of the grammatical component of the POTAI into consideration, the BIOCOR is far from trivial to be described as challenging for 10th graders to process with regard to grammar issues.

1) The BIOCOR is ready to be characterized by the lack of versatile use of tenses. There seems to be a stable preference for simple tenses to continuous tenses, which grammar trait renders the biology texts a smooth access of apprehension for Hungarian students. The mother tongue of the Hungarian bilingual students does not distinguish the simple versus continuous aspect of tenses, and it is the more complex continuous aspect which tends to be more problematic to process for low-achieving students in English, while simple tenses appear to pose fewer difficulties for ESL students at a B2 level. Among all the various tenses, present simple and past simple dominate the BIOCOR, whose simplicity (both in terms of formation and of the shades of meaning conveyed) make the register straightforward to process. The complete absence of several tenses (the past continuous, the past perfect continuous, the *'used to'* structure, the future continuous, the future perfect simple and continuous, and the *'going to'* structure) constitutes the grammatical plainness of the biology texts and thus fails to a account for the difficulty of processing the biology texts for 10th graders.

2) Among the four types of conditional structures, the zero conditional is the most prevalent in the BIOCOR. The zero conditional is used for explaining general laws of nature, where the result of the condition is always true. This explains why the zero conditional appears twice more frequently in the BIOCOR, which imparts knowledge on the laws of science, than in the general English REFCOR. The abundance of the zero conditional, compared to that of other conditionals, makes the biology register lucid for the target readers. The recognition of the zero conditional scarcely contains complications for students at a B2 level; on the other hand, comprehending situations of facts rather than hypothetical ones is less arduous. From the perspective of more complex hypothetical conditional structures (third

and mixed), the BIOCOR displays no signs of intricateness either, since they are not even present in the BIOCOR.

3) Scientific texts are expected to apply an impersonal and universal tone of language, for which reason the passive voice is anticipated to be used with high frequency in science texts (Wilkinson, 1992). In harmony with the expectations, the BIOCOR applies more instances of passive voice with a direct object than the REFCOR. However, it is important to note that statistically there is no significant difference between the frequencies of the passive voice in the two corpora. The lack of such a stark difference demonstrates the close similarity of the two registers, which might be explained by the fact that the biology textbook is written for non-scientific teenagers. In order to suit the needs of the target reader, the biology textbook tends to avoid the scientific, universal tone but amply applies situational examples, which incline to use the active voice. The apparent simplicity of the language use of the BIOCOR is not interfered with any 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'*) to any extent.

4) Informative, academic registers whose end is to impart knowledge apparently need to clarify concepts through definitions. The traditional Aristotelian type of 'genus – differentia specifica' definition linguistically relies on defining relative clauses. Accordingly, the BIOCOR uses defining relative clauses with a relative pronoun more frequently than the REFCOR. The student-friendly nature of the BIOCOR can be captured through the fact that it uses the less transparent linguistic structure applicable for giving definitions, the defining relative clauses without a relative pronoun, significantly fewer times than the REFCOR. Providing extra information, which is revealed by the frequency of the non-defining relative

clauses in the corpus, in not typical of the BIOCOR. This implies that the BIOCOR poses no serious difficulties for the reader in terms of entangling various pieces of information according to their order of importance and priority. The complete lack of progressive participles (both in the present and in the past) shows that the BIOCOR avoids concise and thus densely-packed linguistic constructions.

5) The intensive presence of nominal relative clauses (NRC) without a reporting verb without time shift or with an infinite verb in the BIOCOR reveals how little impersonal the register of biology textbook for secondary students is. The biology texts are bounteous of NRC sentences without a reporting verb which use personal tone, as if they were part of a dialogue in a spoken context, in order to clearly direct the attention of the reader, e.g., *'you will probably think,'* or *'we are not sure.'* NRC sentences with a reporting verb, however, do not regularly appear in the BIOCOR. The low-frequency of NRC sentences with a reporting verb implies that the register conveys widely-accepted laws of nature in the field of biology, which do not need reporting through scientific referencing. Although academic writing tends to introduce the results of experiments by referring to the researchers who conducted them (Bailey, 2011), the register of biology textbook written for secondary students fails to allow for such reporting as the target audience belongs to the community of teenagers, many of whom might not even wish to become scientists. The presence of reported open questions in the BIOCOR also indicates that the register is not void of personal tone. This is the result of the tendency to provide explanations through situations, where the reader is addressed directly.

6) The aspect of infinitives is the only perspective of the grammar part of the POTAI where the BIOCOR appears to be more challenging to process than the REFCOR. 9th grade

students trained on the REFCOR become familiar with simple infinitives, however, studying the BIOCOR requires the knowledge of handling passive infinitives as well, which is not as ample in the REFCOR as in the BIOCOR. Moreover, the BIOCOR also applies progressive passive infinitives, a grammar item which is completely absent in the REFCOR. Besides these two infinitives, the BIOCOR does not exploit the richness of more complex infinitives (such as perfect passive, perfect progressive, perfect progressive passive infinitives) to any extent, thus the register remains moderately challenging for 10th grade bilingual students from this respect.

7) Prepositions at the end of clauses is a grammar feature which is absolutely missing from the BIOCOR. Clause final prepositions is a grammar issue which can easily cause difficulties for Hungarian students at B2 level, whose mother tongue does not employ the structure (as Hungarian is an inflective language that does use prepositions at all). Thus the absence of occasionally student-puzzling clause final prepositions ensures that the BIOCOR is easily accessible for ESL students to study.

8) Finally, the BIOCOR appears to use modal auxiliaries in a frugal manner. Altogether 30 modals are not present in the biology texts at all, and it is only few of them which appear in the biology texts. The modal auxiliaries present in the corpus, however, are used lavishly in a repetitive fashion. Among the few modals, the functions of ability, that of the level of certainty and of obligation are also expressed in the BIOCOR. Ability in the present is conveyed through the use of the modal auxiliary *'can,'* which is one of the modals 9th graders first learn in the 'zero-year' intensive language course, thus it tends not to pose any challenge for them by the end of the academic year. The level of certainty is expressed by the modal auxiliary *'may'* in the BIOCOR, which is not among the ones which typically cause

difficulties of understanding a text for B2 level ESL students either. The function of obligation in the present is expressed by the modal *'must'* in the BIOCOR, which is surprisingly not present in the REFCOR to any extent. Despite the fact that the REFCOR avoids the use of 'must,' the auxiliary is not unfamiliar to Hungarian ESL students, whose mother tongue uses a similar sound modal (*'muszáj'*) with similar functions. The excessive presence of the modal *'must'* in the BIOCOR can be explained by the fact that the register regularly presents laws of nature with this modal rather than using the more impersonal passive voice.

From the results of the grammatical component of the POTAI it can be seen without serious uncertainties that the grammatical phenomena which describe the register do not make the biology texts demanding to comprehend for 10th grade bilingual students. Moreover, their grammar tends to be less advanced than that of the REFCOR. As the grammatical analysis of the corpus does not give sufficient explanations for the difficulty of processing the biology texts for 10th grade students, let us continue our inquiry with a different possible feature that can pose challenges for them: the sentence structure of the corpus.

4.3 Sentence complexity

The present section examines if the BIOCOR is accessible for 10th grade bilingual students without serious challenges with respect to the sentence complexity of the corpus. The complexity of sentences is studied from four aspects: the lengths of sentences, the length of packets, the relationship between the length of sentences and that of words, and finally the complexity of syntactic structures in the corpus.

4.3.1 Sentence length

One of the desirable features of an easily comprehensible text is simplicity, another is brevity (Woods et al., 1998). The length of sentences in the BIOCOR is analysed in order to uncover to what extent the register can be described as specific from this respect. The BIOCOR (of 7,021 words) comprises 567 sentences, while the REFCOR (of 7,098 words) contains 462 sentences. From these figures it can clearly be seen that the BIOCOR applies shorter sentences than the REFCOR. The average sentence length in the BIOCOR is 12.38 words, whereas that of the REFCOR is 15.38. The deduction can be made that sentences in the BIOCOR tend to be three words shorter than those in the REFCOR. Such a difference might not seem to be dramatic, however, it should be noted that even the REFCOR itself, which contains longer sentences than the BIOCOR, falls behind the expected sentence length in written English, that of 20 words on average (Harrison & Bakker, 1998). Comparing the average sentence length of the BIOCOR to this general baseline, it needs to be pinpointed that the BIOCOR uses 40% shorter sentences on average than it can be anticipated from an academic register. That is, the average sentence length of the BIOCOR gives no reason for 10^{th} graders to find the corpus difficult to process from this respect. On the contrary, the figure predicts a corpus that is broken up into numerous, relatively short conscious units, which are clearly perceptible due to the sentence final punctuation marks.

If it is not the average sentence length that poses challenges for 10^{th} grade bilingual students, it is to be investigated whether the uneven distribution of sentences of various lengths might cause difficulties of comprehension. Diagram 2 shows the distribution of sentences of different lengths in the BIOCOR.

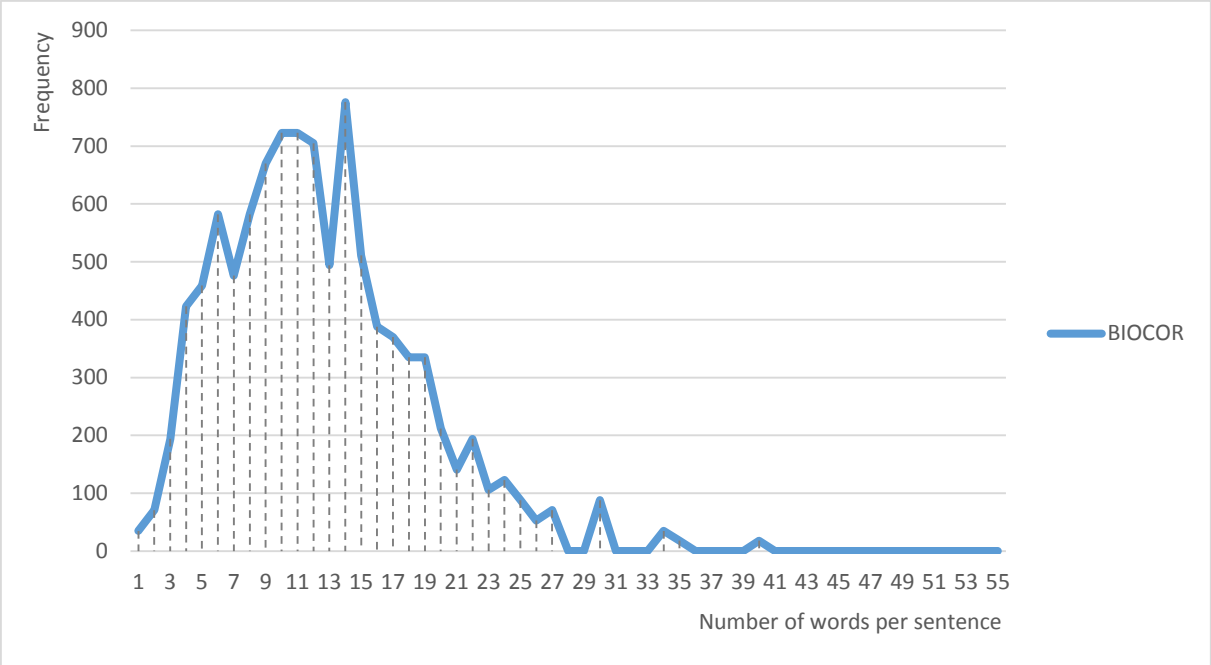

Diagram 2 The frequency of different sentence lengths in the BIOCOR

The peak of the line graph is at 14, summiting at 776, which reveals that the most frequently applied length of sentence consists of 14 words in the BIOCOR, and its presence is 7.76 percent among the entirety of all the sentences in the corpus. The line graph also displays that the most typical sentence length in the BIOCOR, i.e., the sentence length that the target readers most frequently face in the register, is within the range of seven to 15 words. Shorter sentences, whose range of length is between four to seven words, are numerous, too. Some of these strings of words are titles and headings; however, not all of them. The running text of the BIOCOR does contain crispy sentences of extreme brevity, ones that are shorter than seven words. The BIOCOR prefers using short sentences rather than combining them, for instance it writes *'Viruses have various shapes. Some are rod-shaped.'* in two separate sentences instead of mingling the information in one average-length sentence. Sentences of one single word or of two words are also present in the BIOCOR, but not commonly frequent. These elliptical sentences belong exclusively to titles and headings, none of them form part of the main text. Longer sentences with a length ranging from 16 to 22 words are not

overwhelmingly used in the BIOCOR but are modestly present. In contrast, ones that are longer than 22 words hardly appear. Among them, the longest sentence in the BIOCOR amounts to 40 words.

As a baseline of comparison, the distribution of sentences with different lengths in the REFCOR is shown in Diagram 3.

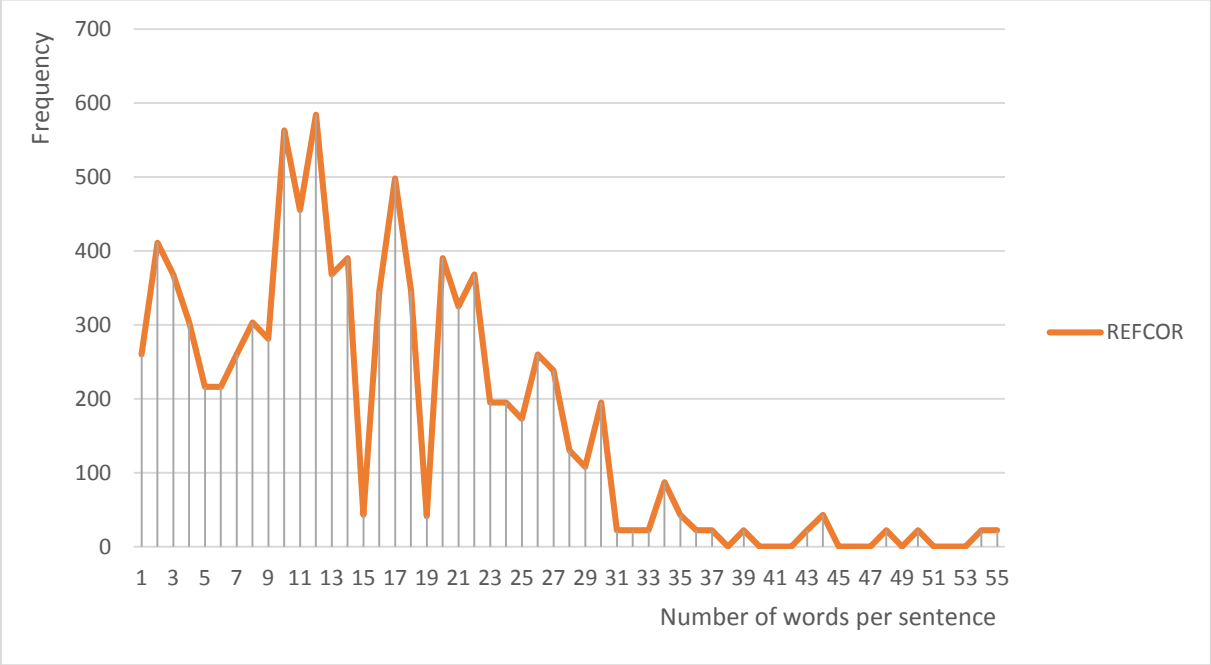

Diagram 3 The frequency of different sentence lengths in the REFCOR

The summit of the line graph is at 12, reaching 584, which indicates that the most frequently used sentence length in the REFCOR is 12 words, whose occurrence is 5.84 percent in the corpus. The most typical sentence length in the REFCOR, that is, the length of sentences that the target audience normally reads, is between eight to 24 words. It can be seen that the range of characteristic sentence length in the REFCOR is wider than in the BIOCOR. Even more notably, this range in the REFCOR contains sentences which are considerably longer than in the same range in the BIOCOR, in certain cases even three times longer. Brief sentences, ones that comprise four to eight words, are infrequent in the REFCOR, though

present. Some of these concise sentences are titles and headings; however, some are part of the main text. Extremely short sentences, ones that contain one to three words, are numerous in the REFCOR. These sentences of extreme brevity are exclusively titles and headings, none of them form mart of the main text. Longer sentences, ones which comprise 24 to 34 are less numerous in the REFCOR, however, still present. Exceedingly long sentences, which are longer than 35 words, appear sporadically in the REFCOR. The longest one among them contains no fewer than 55 words.

The similarities and differences of the distribution of various sentences lengths in the BIOCOR and that in the REFCOR can be observed in Diagram 4, which displays the frequencies of different sentence lengths in the two corpora comparatively.

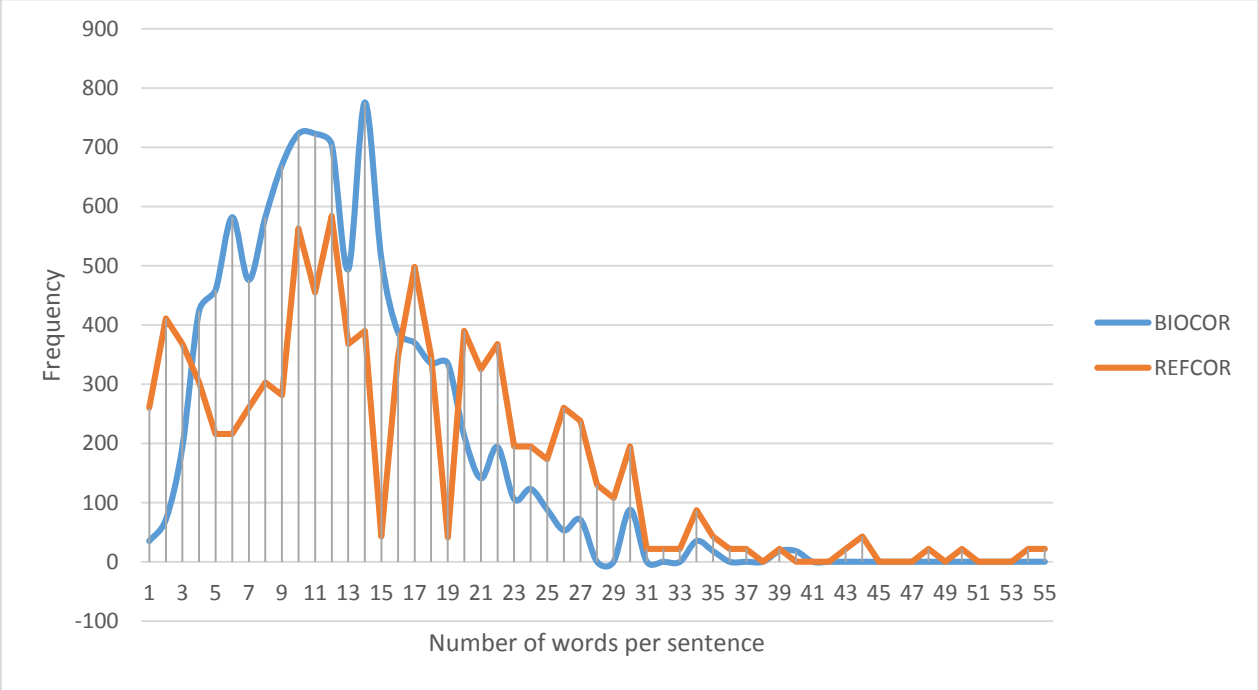

Diagram 4 The frequency of different sentence lengths in the two corpora: in the BIOCOR and in the REFCOR

The visual comparison clearly demonstrates that extremely short sentences, ones containing one to four words, are more numerous in the REFCOR, while are present in the BIOCOR in a modest manner. The reason for this difference lies in the distinct nature of titles and headings in the two corpora. The REFCOR tends to use one- or two-word-long labels, which capture the gist of the ensuing paragraphs, not infrequently applying a witty twist in the absolutely concise summary. In contrast, the BIOCOR can be characterized by longer text-organizing devices, which aim at focusing the reader's attention to the topic discussed and to the most important pieces of information of the following stretches of text. For instance, the heading *'Living things get rid of poisonous waste'* is not intended to be concise but rather informative. Headings in the BIOCOR also tend to be questions, such as *'How do bacteria survive bad conditions?'* or *'How is the tapeworm adapted to its parasitic life?,'* whose main purpose is to draw the readers' attention to the line of argumentation in the following paragraphs. Lengthier text-organizing devices in the BIOCOR can be explained by the objective of the register of the textbook: longer but clearer titles and headings serve educational purposes better than shorter and ambiguous or hazy ones. This feature of the BIOCOR gives no reason for 10th grade bilingual students to find the text challenging to process since the REFCOR appears to be more demanding and at times even puzzling from this respect. Diagram 3 also illustrates that the BIOCOR outweighs the REFCOR with sentences of four to 17 words, for which the reason is the prevalence of relatively short sentences in the BICOR. Since the BIOCOR appears to use shorter sentences dominantly and clearly more frequently than the REFCOR, the range of typical sentence length of the BIOCOR cannot explain the difficulties bilingual students perceive when attempting to comprehend the corpus either. Finally, longer sentences, ones which contain 17 words at least, are less abundant in the BIOCOR than in the REFCOR. There are two exceptions, two singular instances where the number of long sentences in the BIOCOR is larger than that in

the REFCOR (sentences of 19 and 40 words). However, the BIOCOR shows a tendency to use verbose sentences with great moderation, if at all. Again, similarly to the previous ones, this trait of the corpus cannot account for the difficulties 10th grade bilingual students experience to face. To conclude, the distribution of various sentence lengths in the BIOCOR reveals that the easily-perceivable logical units in the corpus, i.e., sentences, are structured in a less challenging and simpler manner than those in the REFCOR.

4.3.2 Packet length

The comprehension of a long sentence can be aided by punctuation marks, which unmistakably denote smaller units of the sentence and clearly signify different logical relationships among these subunits. The words in between two punctuation marks within a sentence are termed packets, whose length on average tend to be under eight words in a written English text (Harrison & Bakker, 1998). Computing the average packet length of the BIOCOR (containing 7,021 words) and that of the REFCOR (consisting of 7,098 words), it can be seen that both corpora correspond to the expect average. The BIOCOR levels at 8.052 words, while the REFCOR at 8.205 words. The packet length in the BIOCOR is slightly shorter than that in the REFCOR; however, the difference is not major enough to differentiate the two registers from this respect. Both the BIOCOR and the REFCOR average eight words in a packet as their usual rate.

In the case of uneven distribution, averages might be considerably different from the values which are most frequent in a corpus. In order to see to what extent packet length might cause problems for 10th grade bilingual students, the distribution of various packet lengths in the BIOCOR was also analysed, whose results are illustrated in Diagram 5. The most frequent packets, the ones the reader meets the most often when comprehending the text, contain six

words. The six-word-long packets constitute a great part, nearly 10 per cent (9.75%), of the corpus. The range of most frequently occurring packet lengths include packets that contain two to twelve words. The overwhelming majority of packets in this range is below the average packet length of the corpus. Thus nearly two thirds, 57.72 per cent, of the whole BIOCOR contains packets which do not exceed the average eight-word-length. The BIOCOR tends to use one-word-long packets moderately, which shows similarity to the use of longer packets containing 13 to 20 words. Packets of 21 to 23 words are still present in the BIOCOR, however, their appearance is sporadically insignificant. Packets which are even longer fail to appear more than once in the corpus. The longest packet among them, which contains two defining relative clauses, where no commas can be placed appropriately, comprises 29 words.

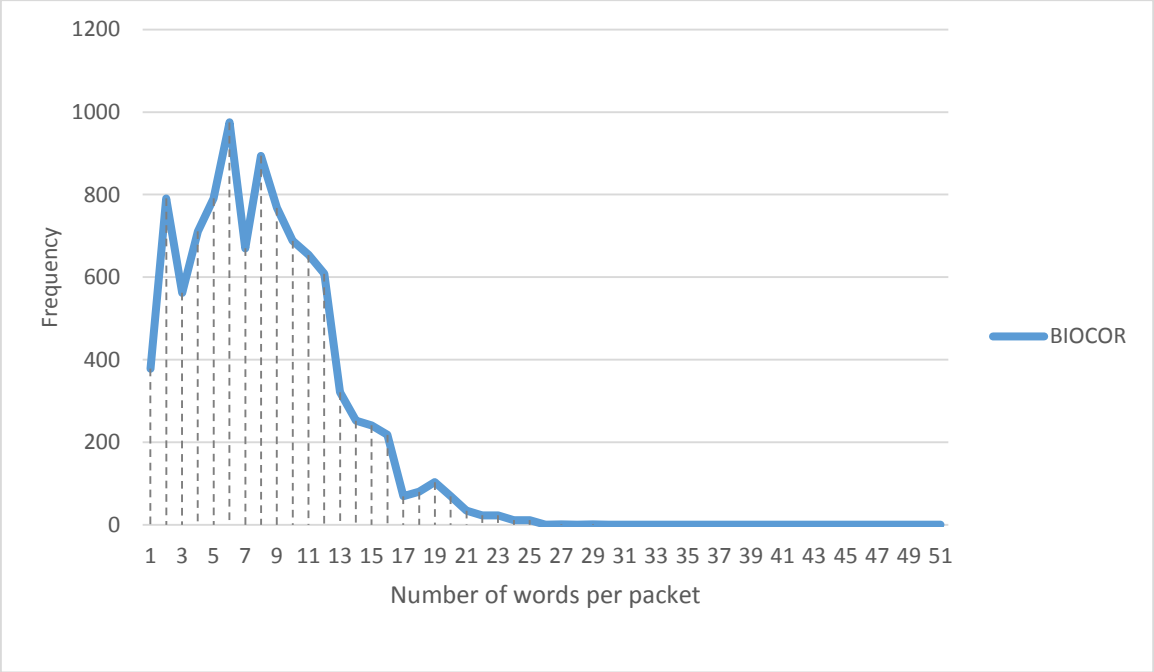

Diagram 5 The frequency of different packet lengths in the BIOCOR

The distribution of various packet lengths in the REFCOR shows a pattern which is similar to that of the BIOCOR in many ways (see Diagram 6). The line graph indicating the frequency of various packet lengths peaks at three, which reveals that the most numerous packet in the REFCOR contains three words. Similarly to the BIOCOR, the packet with the

most typical length builds up approximately 10 per cent (9.6%) of the REFCOR. Despite the similarity in the two patterns, there is a stark difference between the two registers. The most frequently occurring packet in the BIOCOR summits at six words, while the REFCOR only at the half of this figure, at three words. Paralleling with the BIOCOR, the range of most frequently occurring packet lengths in the REFCOR demonstrates an exact correspondence with that of the BIOCOR. In both cases the range of highest frequency tops at 12 words. Besides, the REFCOR also tends to use packets that do not exceed the average packet length of the corpus. With a nearly two-third presence (60.93%), packets that are not longer than eight words outnumber packets which are longer than the average length in the REFCOR, too. The range of mildly used packet lengths in the REFCOR coincides with that in the BIOCOR, the REFCOR also uses packets of 13 to 21 words moderately. Also, the lengths of packets that seldom occur at irregular intervals in the REFCOR, comprising 22-23 words, correspond to those in the BIOCOR. Likewise, packets of 24 words or of even longer length are exemplified only by single instances in both corpora. The longest packet in the REFCOR embraces 51 words, which is 150 per cent longer than the most extended packet in the BIOCOR.

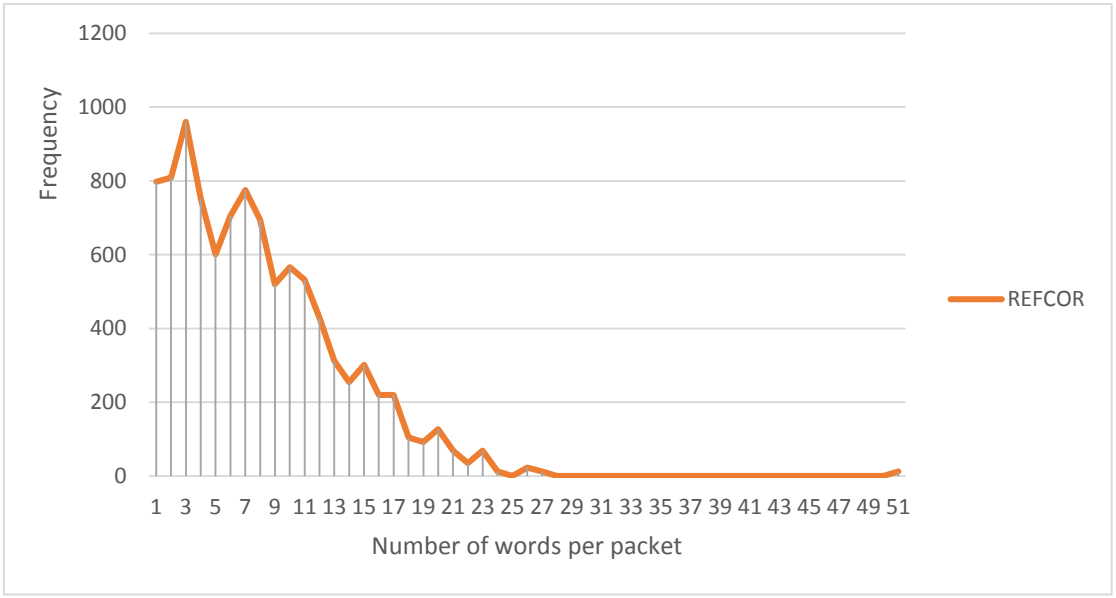

Diagram 6 The frequency of different packet lengths in the REFCOR

The results of the comparative analysis of the distribution of packets with various lengths in the BIOCOR and the REFCOR are displayed in Diagram 7. The great similarity in the shape of the two line graphs reveals a closeness in the pattern of the distribution of the different packet lengths in the two corpora.

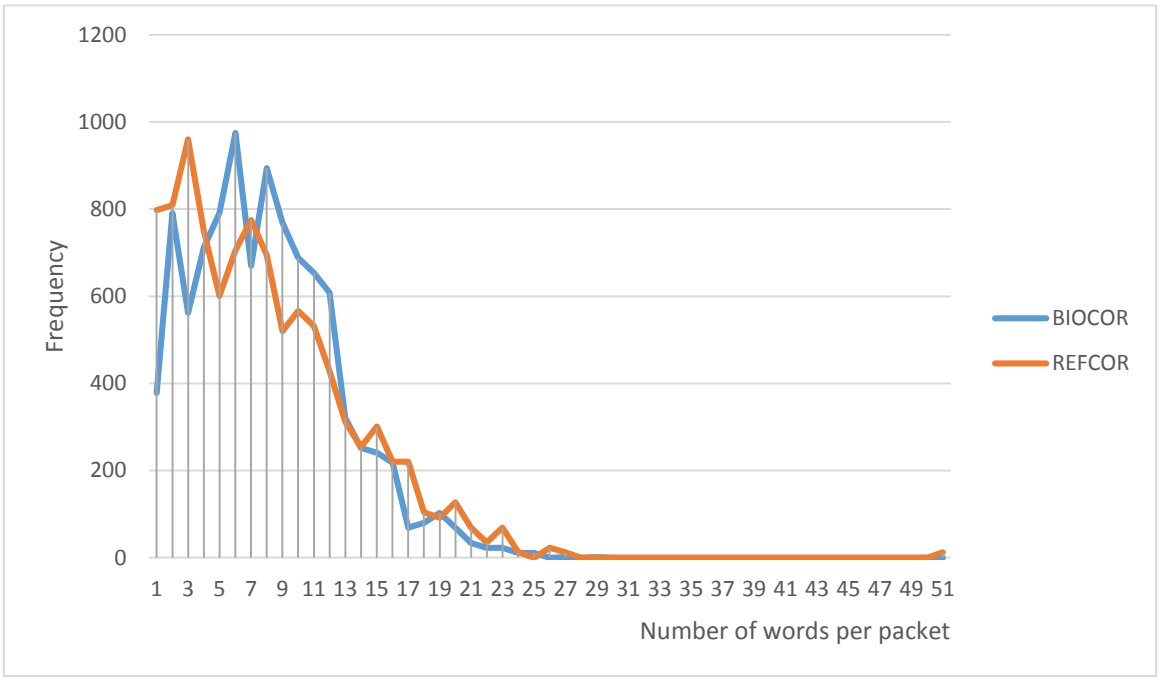

Diagram 7 The frequency of different packet lengths in the BIOCOR and in the REFCOR

The great bulk, nearly two thirds, of both corpora consists of packets whose lengths succeed in remaining crispy by not overstepping the eight-word boundary, which marks the average packet length in the corpora. This result reinforces the notion that processing the REFCOR prepares bilingual students appropriately for comprehending the BIOCOR in the 10th grade. With regard to the length of its clearly indicated sentential subunits, the majority of the whole BIOCOR poses no serious challenges for the target readers. Furthermore, the falling patterns of the two line graphs commence at the same point, at the packet length of 12 words. The correspondence is a sign of another similarity between the two corpora. Both the BIOCOR and the REFCOR incline to use packets whose lengths tend not to be longer than 12 words. The parallel that the range of most frequently occurring packet lengths goes up to a

dozen words at most both in the BIOCOR and the REFCOR indicates that BIOCOR is accessible from this point of view for 10th grade bilingual students, since they were trained on the REFCOR, which can be described by the same range of packet lengths in general. Packets of more extensive length, 13-20 words, appear with moderate frequencies both in the BIOCOR and in the REFCOR. It means that 9th grade bilingual students are exposed to processing such lengths of packets to an appropriate extent; that is, they cannot be claimed to be unfamiliar with lengthier packets when reading the BIOCOR in the 10th grade. In addition, packets of extreme length are untypical of the BIOCOR, their appearance is infrequently rare in the corpus. The same can be accounted about the REFCOR, thus 10th grade bilingual students encounter no novel challenges when processing the BIOCOR with regard to its exceptionally long packets. Besides the many similarities, there is one single difference between the BIOCOR and the REFCOR. The most often occurring packet length in the BIOCOR is six-word long, while that in the REFCOR amounts only to the half of it. This dissimilarity might be the only reason why 10th graders might find the BIOCOR difficult to process. However, taken into account the fact that the range of the most often applied packet lengths in the BIOCOR and the REFCOR coincide, the weighty impact of the difference in the peak values seems to dissipate. The influence of the different peak values on readability diminishes once it is kept in the foreground that the overlapping range of the most frequently used packet lengths in the two corpora is not insignificant in their amount as they embrace two-thirds of the whole string of texts. Altogether, the various lengths of packets in the BIOCOR gives no well-grounded explanation for the 10th grader bilingual students' perceived challenges. The clearly visible punctuation marks in the BIOCOR help the target reader unmistakeably recognize the sentential subunits of the text in the same manner as the ones in the REFCOR.

4.3.3 Readability indices

It needs to be recognized that no single measure on its own, neither sentence length nor packet length, reveals readability completely. It might be argued that short sentences or packets using longer than average words are more difficult to process than same-length sentences or packets using short (or average-length) words. In order to take this correlation into account, different readability measures have been developed. The current research examines the BIOCOR by a comparative analysis of five grade level readability indices (the automated readability index (ARI), the Coleman-Liau index, the Flesh-Kincaid index, the SMOG index, and the Gunning fog index) in order to discover if the corpus poses any serious difficulties for 10^{th} grade bilingual students from this respect. First the five mean value readability indices are compared and contrasted for the BIOCOR and the REFCOR. Then the distribution of the readability values of the BIOCOR is examined chapter by chapter in order to see if the biology texts which the 10^{th} grade bilingual students are assigned to read show the same level of difficulty all through the academic term.

The comparative bars of the mean readability values (see Diagram 8) noticeably reveal that the five readability indices imply five different grades for the smooth reading of the BIOCOR. The readability values of the BIOCOR predicted by the five indices range from grade 6 to grade 10, that is, a target reading group of teenagers from 11 to 16 years of age is suggested. The underlying reason for the difference among the prediction of the readability indices is the fact that all the five indices use different mathematical formulae (with different core values) to arrive at the predicted grade level (for the exact way of calculation see Section 3.3.3.3 on pp. 103-109).

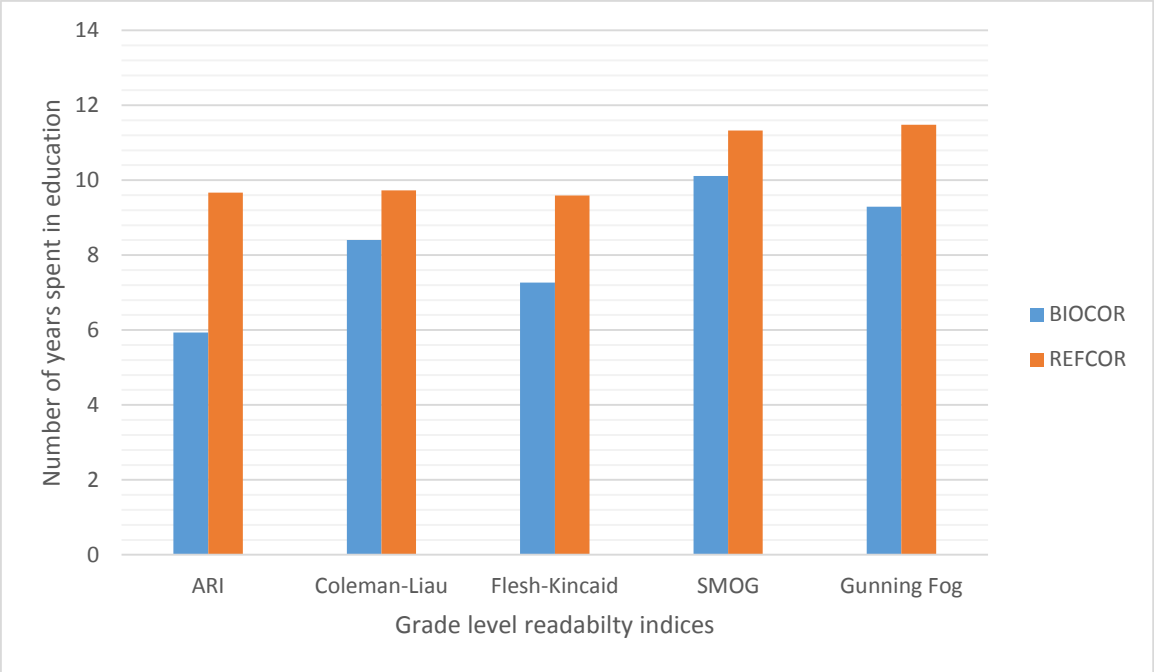

Diagram 8 The readability level of the BIOCOR and that of the REFCO

The automated readability index (ARI) predicts grade 6 for the BIOCOR (5.93), which suggests that the biology texts are intended for the 11-12 year-old age group of native English students, while ARI calculates the REFCOR to be of grade 10 difficulty (9.67), which covers students of 15-16 years old. Comparing the grades of the two corpora, the ARI computes a four-year education difference between them, where it is the REFCOR which needs four years more advanced studies than the BIOCOR. In other words, the BIOCOR requires a far smaller number of years of education than the REFCOR, thus the ARI values do not account for the difficulty 10th grade bilingual students face when processing the BIOCOR.

The Coleman-Liau index predicts grade 8 for the BIOCOR (8.4), which means the corpus is foreseen to be easily read by the age group of 13-14 year-old native English students. The same index predicts grade 10 for the REFCOR (9.73), which indicates an intended age group of 15-16 year-old pupils. The Coleman-Liau figures calculates that the comprehension of the REFCOR requires two more years of formal studies than that of the

BIOCOR. The difference of the reading difficulty of the two corpora is two years smaller in the case of the Coleman-Liau index than in that of the ARI, however, both readability indices predict that the REFCOR is years more difficult to process than the BIOCOR. Accordingly, the Coleman-Liau index does not predict any challenges of understanding the biology texts for the 10th grade bilingual students, either.

The Flesch-Kincaid grade level index shows that the BIOCOR corresponds to the difficulty of the readings of 7th graders (7.27), that is, it is understood effortlessly by 12-13 year-old native English students. Applying the same formula to the REFCOR gives grade level 10 (9.59), indicating an intended audience of age group 15-16, which is a three-year older target reading group than in the case of the BIOCOR. Again, the BIOCOR requires a considerably smaller number of years of formal education than the REFCOR, and thus the result of this formula cannot account for the difficulties 10th grade bilingual students perceive when processing the BIOCOR.

Among the five indices, the SMOG grade indicates the greatest number of years of formal education needed for comprehending the BIOCOR, grade 10 (10.11), which corresponds to age group 15-16. While the same index predicts grade level 11 for the REFCOR (11.33), which entails a target readership of 16-17 years old native students. This is the only index where the difference in the number of academic years needed to process the two corpora without serious difficulties is not more than one. Despite this relative closeness of the readability levels of the BIOCOR and the REFCOR, the SMOG grade also reveals that reading the REFCOR entails more years of studies than the BIOCOR. Similarly to the other readability indices, the SMOG index predicts no comprehension difficulties of the biology texts in the 10th grade.

Finally, the Gunning fog index anticipates that understanding the BIOCOR with ease needs nine years of education (9.29), it is predicted to be intended for 14-15 year-old native English students. Applying the same readability index to the REFCOR shows that it is understood without complications by two years older students, eleventh graders (11.48). In harmony with the other four grade level readability indices, the Gunning fog index also displays that the BIOCOR is less laborious to process than the REFCOR. This readability index gives no explanation for the struggles of processing the biology texts observed by 10th grade bilingual students.

Despite the failure of the five readability indices to predict the same number of academic years needed to process the BIOCOR with ease, they still detect a clear pattern of difficulty of the corpus. All the indices result in a grade level for the BIOCOR which is lower than that of the REFCOR. Although the various readability indices foresee diverse target reading age groups, all of them without exception reveal the tendency that the BIOCOR requires fewer years of formal education than that of the REFCOR. In the case of the SMOG grade, the difference is merely one year, however, all the other indices predict that comprehending the REFCOR presupposes several years more academic studies than understanding the BIOCOR. In the case of the Gunning fog and the Coleman-Liau indices it is two additional years, the Flesch-Kincaid index predicts three more years, and the automated readability index goes as far as anticipating four more academic years. Regardless of the type of the readability index, all five of them grade the BIOCOR to be comprehended at an earlier age than the REFCOR. This result disagrees with the pedagogic anticipations prevalent at the bilingual secondary school among the biology teachers (Cserép, 1997) and among the English teachers as well, who instruct 9th grade students to pass the FCE exam (from which materials

the REFCOR was compiled) in order to make them prepared for the 10th grade assignments (part of which is processing the BIOCOR).

None of the mean readability values give satisfactory justification for the perceived difficulties the 10th grade bilinguals face when reading the BIOCOR. On the contrary, all the indices imply the relative ease of the BIOCOR compared to that of the REFCOR. To see if readability accounts for the 10th graders challenges to any extent, it is worth examining whether the readability level of the BIOCOR is evenly distributed. Even distribution entails that the corpus poses the same level of difficulty all through its eight chapters. However, uneven distribution can imply that some parts of the BIOCOR exceed the mean readability values in an extreme manner and as a consequence might pose challenges for the target readers, which possible difficulties are evened out and thus not expressed in the mean values. The distribution of the ARI values of the BIOCOR, displayed in Diagram 9, shows that merely one of the chapters (Chapter 3) deviates mildly from the mean readability value of the BIOCOR (5.93). Reading Chapter 3 presupposes 8 years of formal education (7.55), which is two years more than the average of the BIOCOR. At other parts of the BIOCOR, however, the difference of the ARI values show an even distribution. Not more than three chapters are predicted to be slightly more difficult than the mean value: Chapter 2 (6.69), Chapter 4 (6.86) and Chapter 7 (6.84). None of the chapters of the BIOCOR peaks to reach close to the mean value of the REFCOR (9.67), even the most challenging chapter of the BIOCOR is predicted to be read with ease by a target group of two years younger than that of the REFCOR. For these reasons the distribution of the ARI values do not give a satisfactory account for the perceived difficulties of the 10th grade bilingual students.

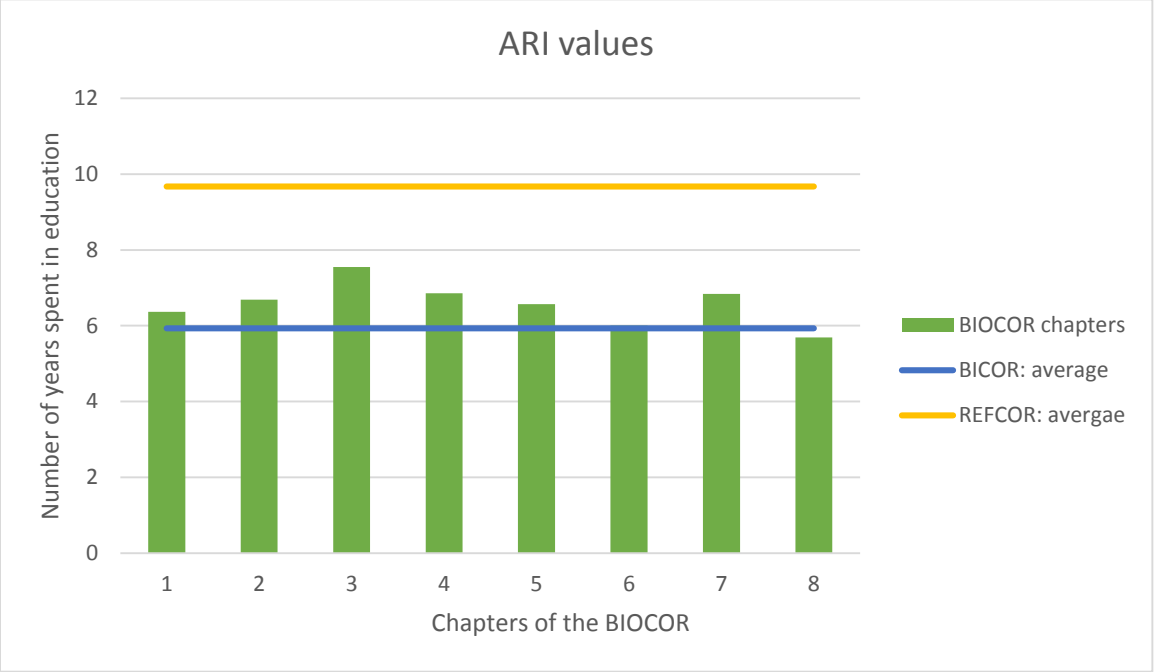

Diagram 9 The ARI values of the BIOCOR chapters compared to those of the averages of the BIOCOR and of the REFCOR

The distribution of the Coleman-Liau values is not entirely the same as that of the ARI values (see Diagram 10). The chapter that differs the most from the average Coleman-Liau readability value of the BIOCOR (8.4), Chapter 7 (9.67), is expected to be read by a two years

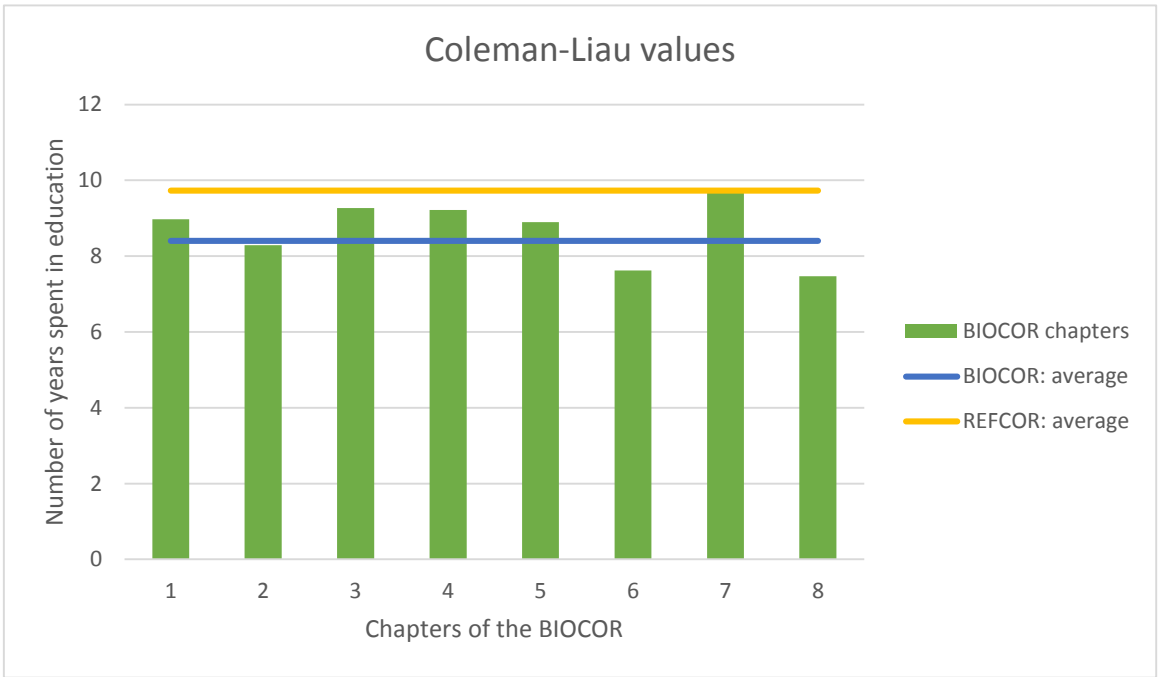

Diagram 10 The Coleman-Liau values of the BIOCOR chapters compared to those of the averages of the BIOCOR and of the REFCOR

older target group (grade 10) than the entirety of the corpus (grade 8). Besides that one single deviation, the difficulty of the biology chapters is evenly distributed. Four chapters show a slight readability difficulty compared to the mean readability value of the BIOCOR (Chapter 1 (8.97), Chapter 3 (9.27), Chapter 4 (9.22) and Chapter 5 (8.9)); nevertheless, the grade difference is confined to be one additional year. The most demanding chapter of the BIOCOR (Chapter 7; 9.67) is anticipated to be as challenging to process as the mean value of the REFCOR (grade 10; 9.73). Yet, the great majority of the BIOCOR, the other seven chapters, are not predicted to be as arduous to process as the average readability of the REFCOR. This way the distribution of the Coleman-Liau values do not explain the struggles 10th grade bilinguals face when processing the BIOCOR.

The distribution of the Flesh-Kincaid values of the BIOCOR also displays evenness (see Diagram 11). Among the eight chapters of the BIOCOR, it is only Chapter 4 (8.62; grade 9) which notably exceeds the expected mean readability value (7.27; grade 7) and consequently a two-year older target audience is predicted for this part of the corpus.

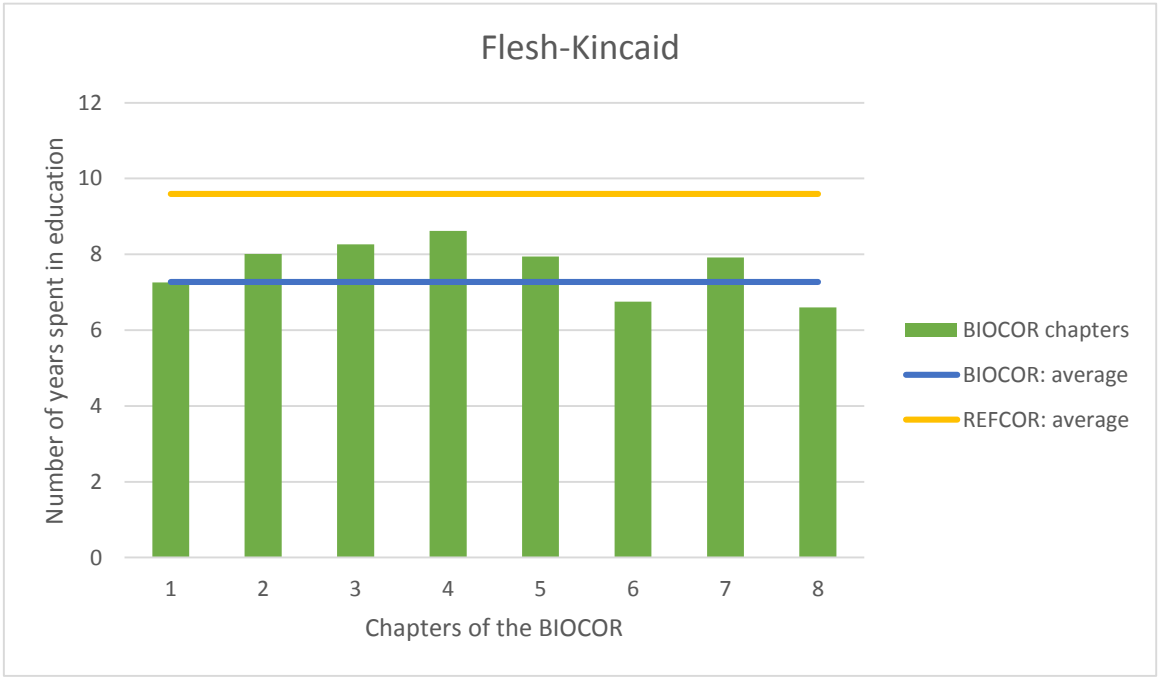

Diagram 11 The Flesh-Kincaid values of the BIOCOR chapters compared to those of the averages of the BIOCOR and of the REFCOR

The other seven chapters, however, do not show any major divergence from the mean readability value of the BIOCOR. A modest difference can be observed in the case of four chapters (Chapter 2 (8.01), Chapter 3 (8.26), Chapter 5 (7.94) and Chapter 7 (7.92)), where one additional year of education is prognosticated compared to the average readability value of the BIOCOR. Still, none of the BIOCOR chapters is expected to be as difficult to process as the mean readability of the REFCOR (9.59; grade 10). Apparently, the distribution of the Flesh-Kincaid values do not provide an explanation for the strenuous efforts 10th grade bilingual complain about when processing the BIOCOR.

The distribution of the SMOG values of the BIOCOR displays complete evenness throughout the entirety of the corpus (see Diagram 12). Although three of the chapters (Chapter 3 (11.03), Chapter 4 (10.99), Chapter 5 (10.62)) surpass the mean readability of the BIOCOR (10.11, grade10), the difference between the average readability and the more challenging chapters is not more than one grade. However, Chapter 6 (8.84) balances the difficulties; it is predicted to be read by 9th graders, a year younger target audience than that of the average of the corpus. The SMOG index differs from the previous three readability indices with regard to the relative mean difficulty of the two corpora, the BIOCOR and the REFCOR. It is the only index which anticipates the easiness of the BIOCOR to be measured by one single grade difference compared to the difficulty of the REFCOR. The predicted closeness of the mean readability of the BIOCOR and that of the REFCOR results in the fact that rounding the SMOG values of some chapters of the BIOCOR (3, 4, 5) are predicted to have the same difficult as that of the mean value of the REFCOR. This feature is unique in the readability pattern, the other indices predicted all chapters of the BIOCOR to fail to reach the highly challenging readability level of the REFCOR. At the same time, no parts of the BIOCOR are forecasted to be more difficult than the mean difficulty of the REFCOR, thus the challenges

10th grade bilingual students face when reading the BICORO cannot be justified by the distribution of the SMOG values, either.

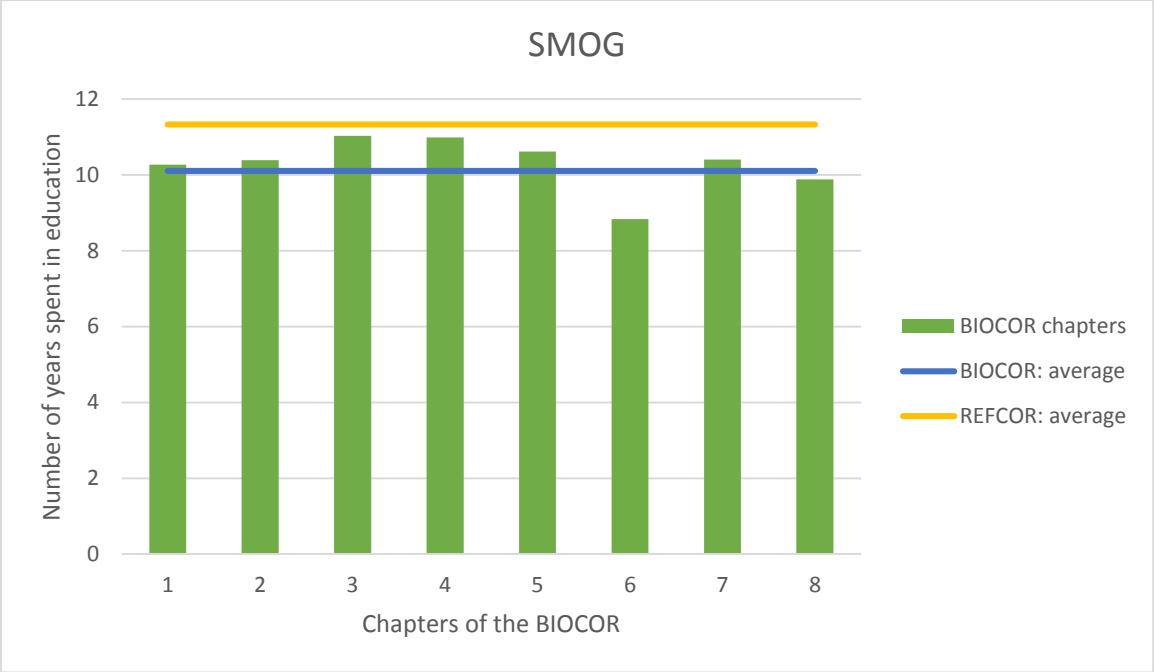

Diagram 12 The SMOG values of the BIOCOR chapters compared to those of the averages of the BIOCOR and of the REFCOR

Diagram 13 reveals that the distribution of the Gunning fog index values of the chapters of the BIOCOR show a modestly even distribution around the mean value of the BIOCOR (9.29, grade 9). There is one chapter which slightly exceeds the mean value by one grade (Chapter 2; 9.9) and two chapters which go above it by two grades (Chapter 3 (10.91) and Chapter 4 (10.56)). The latter two chapters are predicted to reach the difficulty of 11th grade, which coincides with the predicted grade level of REFCOR (11.48). The body of the BIOCOR, however, contains parts whose difficulty are anticipated to be a grade lower than the mean value of the corpus. The readability of Chapter 6 (8.3) is predicted to be one year bellow the mean value of the BIOCOR. Since none of the chapters of the BIOCOR are foreseen to have higher readability values than the mean value of the REFCOR, the Gunning fog values give no solid explanation for the hardships 10th grade bilingual students face when reading the BIOCOR.

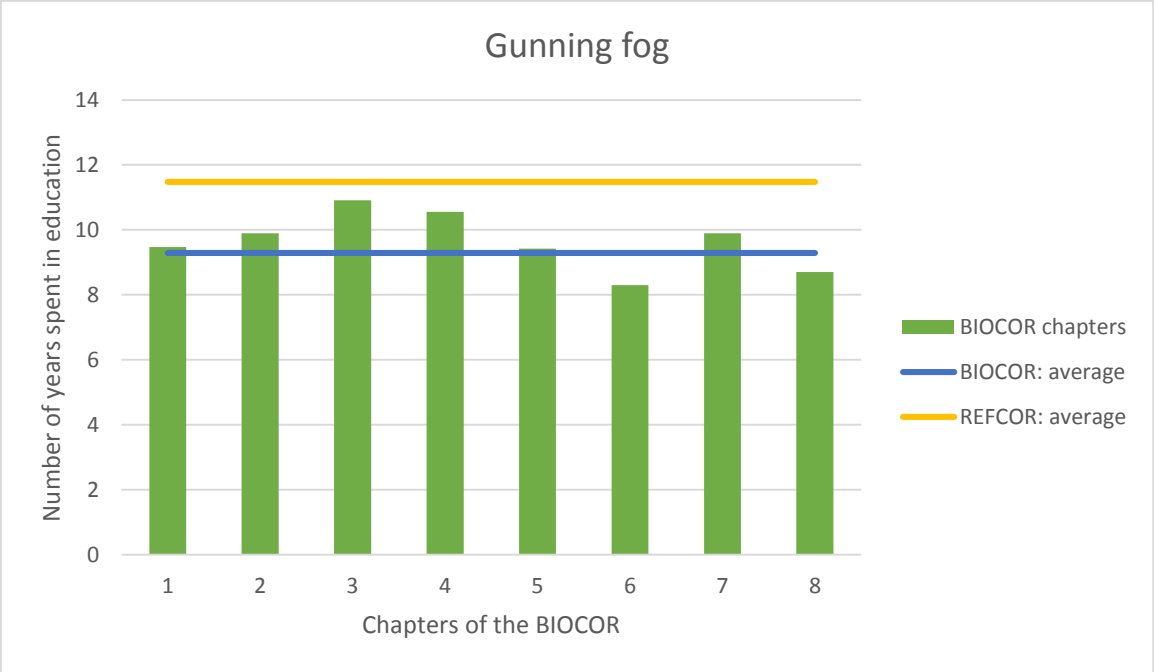

Diagram 13 The Gunning fog values of the BIOCOR chapters compared to those of the averages of the BIOCOR and of the REFCOR

The comparison the five readability indices in the BIOCOR makes the following pattern of readability clearly observable about the corpus.

1) The distribution of the readability level of the BIOCOR chapters reveals that the chapters in the corpus tend to be less challenging than the average readability level of the REFCOR: it is not typical for the BIOCOR chapters to reach the readability level of the REFCOR. There are readability indices (ARI and Flesh-Kincaid) by whose calculations none of the BIOCOR chapters are as difficult as the mean readability value of the REFCOR. One readability index (Coleman-Liau) predicts one single chapter to be as difficult as the mean value of the REFCOR, while others foresee two (Gunning fog) or three (SMOG) chapters out of the eight BIOCOR chapters to be no less challenging than the average of the REFCOR. Even the readability index which yields the highest values for the BIOCOR (SMOG) anticipates less than half of the corpus to be as difficult as the REFCOR, and predicts the other half to be easier to comprehend than the mean difficulty of the REFCOR. The other four

readability indices predict this proportion to be dramatically lopsided in favour of the ease of the BIOCOR compared to the difficulty of the REFCOR.

2) The difference between the mean readability values of the BIOCOR and those of the REFCOR are great enough to be measured in grade differences. The range of the difference between the expected numbers of years spent in education varies between one and four (SMOG 1, Coleman-Liau and Gunning fog 2, Flesh-Kincaid 3, and ARI 4). Regardless of which readability index is considered, it is always the REFCOR that scores years above the BIOCOR in the difficulty of readability.

3) Strikingly, none if the indices predicts any of the chapters of the BIOCOR to be more difficult than the mean value of the REFCOR. Irrespective of the type of calculation of the index, no parts of the BIOCOR can be described as more difficult to process than the mean value of the REFCOR.

This pattern of readability, with regard to the mean values and the distributions as well, allows for no explanation for the hardships 10th grade bilinguals perceive when processing the BIOCOR. In no parts of the pattern does the BIOCOR appear to be more taxing to understand than the REFCOR. Although there are few signs of higher readability values in the BIOCOR, which are characteristic of an extremely small proportion of the corpus, they reach no further than the average difficulty of the REFCOR. In this manner, the readability values can provide no clear justification for the obstacles of reading the BIOCOR by a target audience, who has successfully processed the REFCOR, which requires more strenuous efforts to process according to the readability values.

4.3.4 Syntactic structure

Readability indices describe the level of difficulty of a corpus based on different variables of the text (the length of sentences and words, the number of syllables and characters). Although these measures take into consideration the length of a sentence, readability indices fail to consider the syntactic structure of sentences. For this reason, the grading of a text that contains long and simple sentences might not be radically different from the one that applies sentences that have approximately the same length but contain many dependent clauses (supposing that the length of words is the same in the two texts). Consequently, it is worth examining if the BIOCOR shows any traits of challenging complexity in its syntactic structure, which could account for the difficulty 10th grade bilingual students face when processing the biology texts.

Exploring the sheer number of clauses within the sentences of the BIOCOR (see Diagram 14), it can be claimed that the majority of the sentences corpus (54%) includes one-clause-long simple sentences. Another great bulk of the BIOCOR, one-third of the sentences of the BIOCOR (33%), includes two-clause-long sentences. From a syntactic point of view, the two shortest types of structures build up nearly the whole corpus (87%), that is, approximately nine out of ten sentences of the BIOCOR. The frequency of longer sentences containing three clauses in the BIOCOR drops drastically: merely every 10th sentence tends to be longer than two clauses. Longer sentences than these, ones that use four clauses (3%) or five clauses (1 single instance, which does not amount to 1%) are insignificantly present 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.

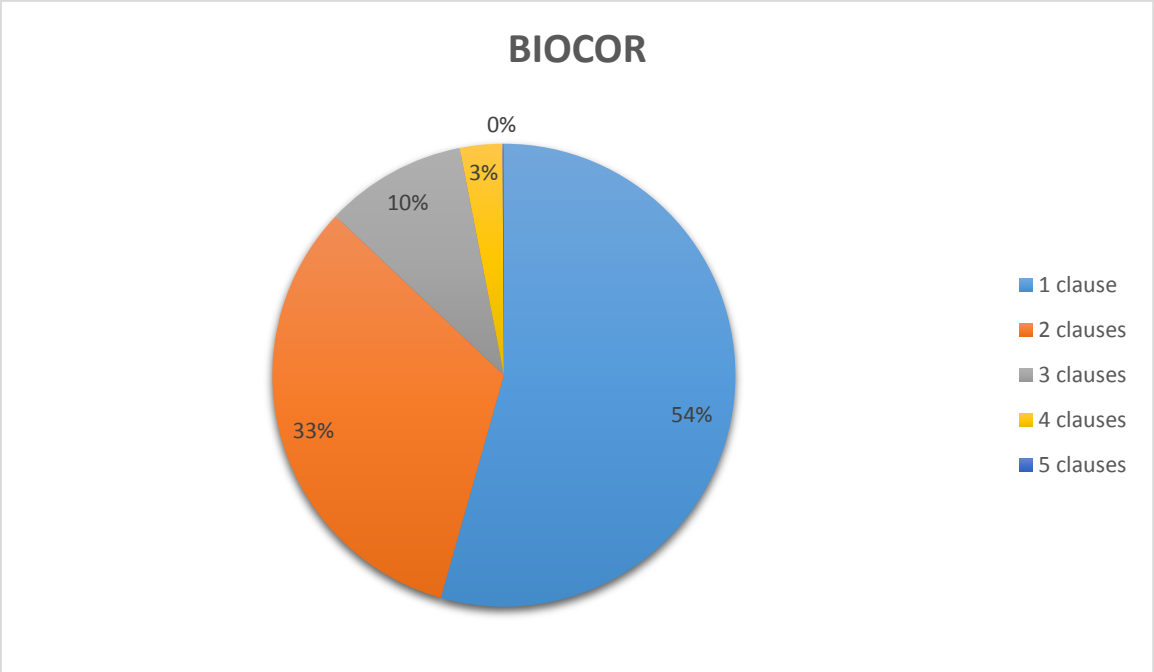

Diagram 14 The frequency of sentences with different numbers of clauses in the BIOCOR

As a baseline of comparison (see Diagram 15), the REFCOR applies one-clause-long simple sentences to a smaller extent (39%) than the BIOCOR (54%). The REFCOR uses the syntactically simplest structure 15% fewer times than the BIOCOR. 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 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. The presence of one- or two-clause-long sentences in the REFCOR does not even reach the two-third of the entirety of the sentences of 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.

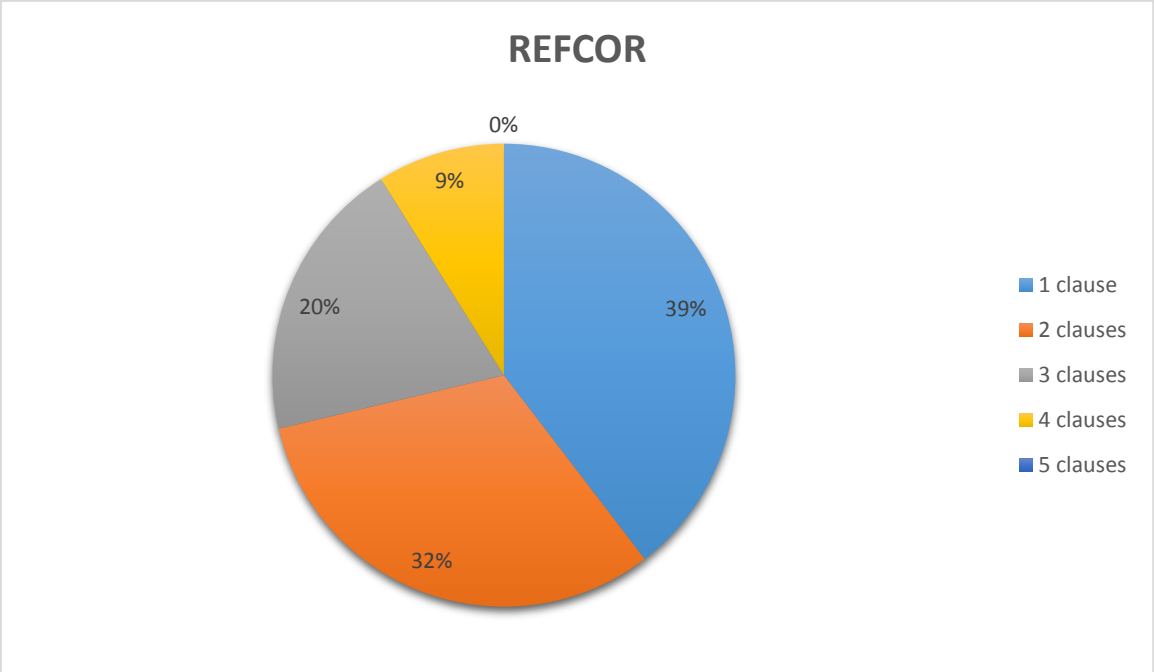

Diagram 15 The frequency of sentences with different numbers of clauses in the REFCOR

For a better understanding of the syntactic nature of the register, Diagram 16 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 Section 3.3.3.4, Table 13). As it was already revealed by Diagram 14, 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, it is barely 7%. Compounding three

independent clauses (Code 3) in the BIOCOR is insignificantly trifle, it amounts merely to 1%. The presence of complex sentences with one single dependent clause (Code 4) in the BIOCOR is nearly half as numerous (26%) as that of simple sentences in the corpus (54%). Longer complex sentences with two dependent clauses (Code 5), however, appear five times fewer (5%) than the shortest type of complex sentences. 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 type show a similarly small rate of appearance in the BIOCOR. Code 5 (complex sentences with two dependent clauses) and Code 7 (compound-complex sentences with two independent clauses and one dependent clause) both appear to a rather limited extent, 5% and 4% in the corpus. Four-clause-long compound-complex sentences (comprising two independent clauses and two dependent clauses, Code 8) are applied half as rarely, their rate amounts to 2% in the BIOCOR. However, four-clause-long compound-complex

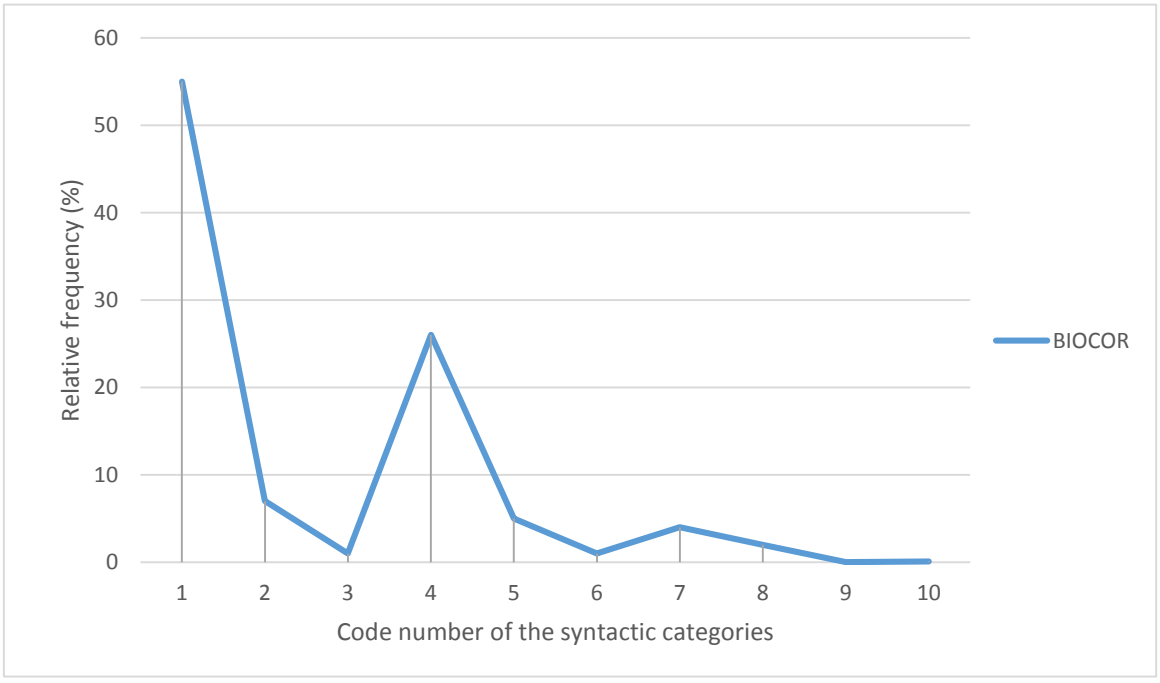

Diagram 16 The frequency of the ten types of syntactic structures in the BIOCOR

sentences of different syntactic nature (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, its frequency does not even reach 1%.

The distribution of various syntactic structures in the REFCOR displays a different pattern than that of the BIOCOR (see Diagram 17). Simple sentences (Code 1) are less

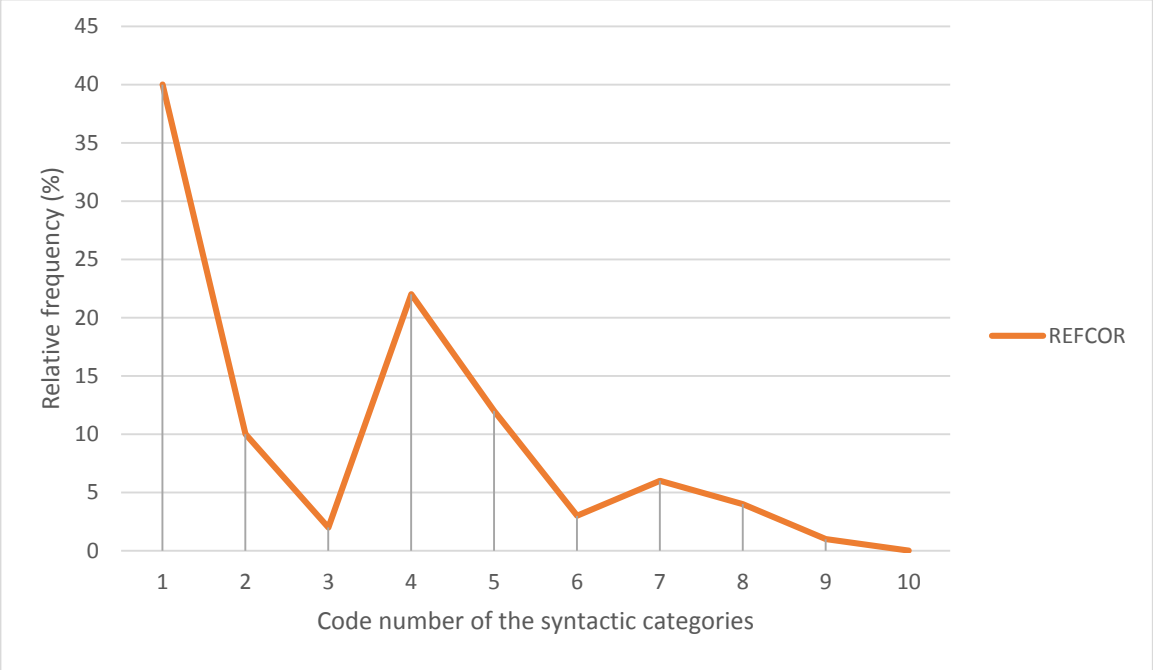

Diagram 17 The frequency of the ten types of syntactic structures in the REFCOR

extensively used in the REFCOR (40%) than in the BIOCOR (54%). However, the frequency of compound sentences containing two independent clauses (Code 2) is notably more numerous in the REFCOR (10%) than in the BIOCOR (7%). Compound sentences with three independent clauses (Code 3) are not considerably present in the REFCOR (2%), but still, 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 (10%). Three-clause-long complex sentences, ones 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 twice as repeated in this corpus as in the BIOCOR (5%). Even longer complex sentences, ones 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 dependent clause (Code 7) or with two independent clauses and two dependent clauses (Code 8) appear with similarly modest frequency in the REFCOR (6% and 5%). The unique presence of four-clause-long compound-complex sentences, ones with three independent clauses and one dependent clause (Code 9), is negligibly small (1%) in the REFCOR. Longer compound-complex sentences, ones with three independent clauses and two dependent clauses (Code 10) are completely absent from the REFROC.

Diagram 18 displays the comparison and contrast 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 structures' code numbers. Based on 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 places are the peak-points of the line graphs, at Code 1 ($p=.001$) and 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 allows the implication that the BIOCOR is more accessible to process than the REFCOR to be evident. 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 numerous as in the BIOCOR. Again, this result demonstrates 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. This abundance of complex sentences might signify that the BICOR requires more effort to be accessible from its readers. However, the presence of complex sentences containing two dependent clauses is more than twice as heavy in the REFCOR as in the BIOCOR. Furthermore, complex sentences with three dependent clauses are used three times more repeatedly 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.

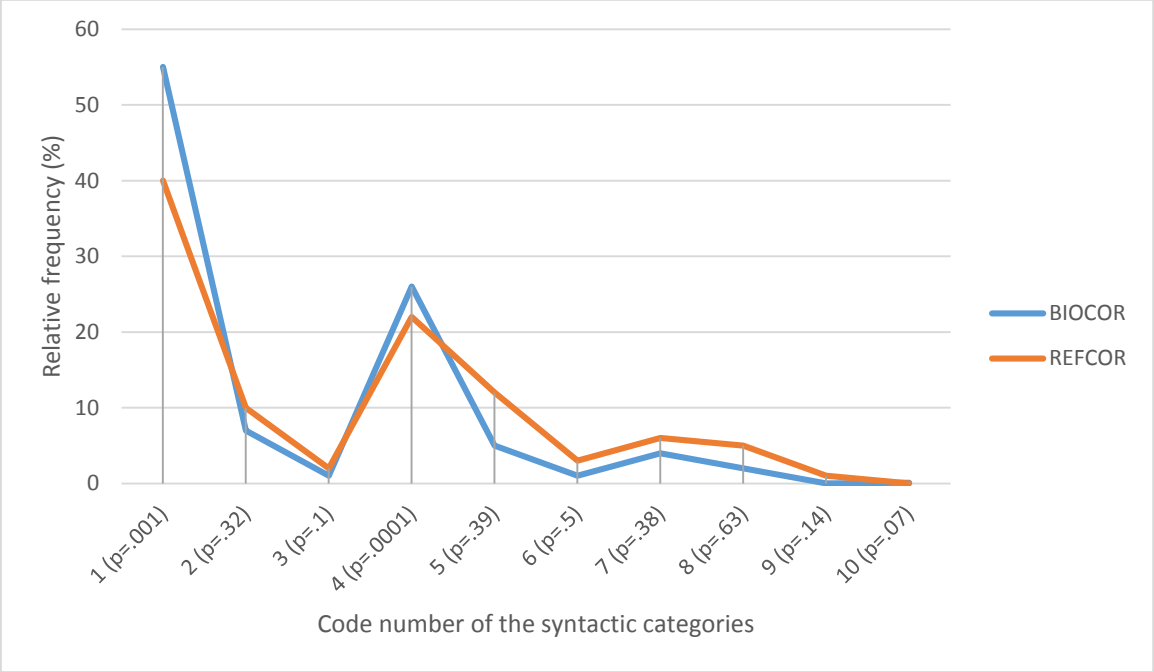

Diagram 18 The frequency of the ten type of syntactic structures in the BIOCOR and in the REFCOR

For this reason, 9th grade bilingual students trained on the REFCOR should hardly find the relatively simple use of complex sentences in the BIOCOR demanding to process. The extremely mild presence of complex-compound sentences in the BIOCOR also makes the corpus more uncomplicated for its readers than the REFCOR. In conclusion, the examination of all the various syntactic structures used in the BIOCOR gives less than solid explanation for the perceived difficulties of 10th grade bilingual students when processing the relevant chapters of the biology textbook.

In this chapter sentence complexity of the BIOCOR and that of the REFCOR has been examined from four different points of view: sentence length, packet length, readability indices and syntactic structure. Apparently, neither of these aspects have given satisfactory explanations for the difficulties 10th grade bilingual students face when they process the biology textbook. Thus it is reasonable to progress with the discovery of the register in a slightly different direction. Let us now see, how markedly the different logical relationships in the content of the biology textbook are signposted for the reader through the overt use of textual metadiscourse markers.

4.4 Textual metadiscourse (TMD)

Textual metadiscourse (TMD) is the collection of linguistic devices that overtly reveal the cohesiveness of the text through explicating the text's organization and displaying the logical flow of its ideas. TMD directs the reading process by way of underlying certain logical relationships, by clearly indicating discourse organization and by clarifying the connections of propositional content. The rate of recurrence of TMD devices, which guide the readers' understanding of the text, influences the readability of the text. The frequent use of TMD markers creates a visibly more cohesive text, which is easier for the reader to process and

interpret than a text with fewer TMD markers, which keeps the text's argumentation linguistically more covert (Gosden, 1992). The comprehension of long, syntactically complicated sentences can be improved if the logical relationship between the clauses are overtly expressed (Selzer, 1983). With this view in the forground, the TMD practice of the BIOCOR was investigated to see the extent to which it guides the target readers to an easier comprehension of the textbook's biology content. Then the frequency and quality of TMD devices in the BIOCOR were compared to those of the REFCOR to disclose if the BIOCOR poses challenges for the 10th grade bilinguals through the texts' spare use of TMD markers.

The ratio of sentences containing TMD markers and those which go without TMD devices both in the BIOCOR and in the REFCOR are displayed in Diagram 19. The bar charts evidently indicate that the two corpora show no substantial difference in this respect. The BIOCOR applies sentences which contain a TMD marker up to the 80% (79.82) of the entirety of the corpus, while it avoids using any TMD markers in only 20% (20.18) of its sentences. Exactly the same way, it is 81% (80.89) of the REFCOR which contains sentences with TMD devices and merely 20% (19.11) of the corpus uses sentences without any TMD markers. The same level of TMD markers in the BIOCOR and in the REFCOR pinpoint that the BIOCOR is as overtly structured as the REFCOR. The heavy preponderance of TMD markers in the two corpora is striking, as Hyland (1998b) found a much lower density both for academic writing (53%) and for biology research articles (60%), in particular[4]. One of the reasons why Hyland's (1998b) data reveal a much lower density of TMD markers in the use

[4] Hyland (1998b) measured TMD against the number of words in the text, for which reason the percentages in his research mean how many TMD markers occur in a stretch of 100 words. The current research, however, uses the sentence as the point of reference since numerous TMD markers are multi-word combinations, which still express one single logical relationship. For example, the contrastive TMD marker 'on the other hand' contains four words, nonetheless, it creates one single relationship in the propositional content of text. For this reason, the present analysis measures the frequency of TMD markers against the number of sentences rather than that of words. Consequently, Hyland's (1998b) results (frequency against the number of words) were converted (into the frequency against the number of sentences) so that the figures became comparable.

of academic English is that his taxonomy was not as extensive as the one used in the present research, consequently a handful of TMD markers might have been collected here, which were not gathered in the previous research. Secondly, Hyland (1998b) discovered the TMD practice of academic English use through investigating research articles written for scientists specialized in the specific field. In contrast, the target audience of the two corpora under investigation is different: the BIOCOR was written for secondary students, that is, for not-yet-specialists of the field; while the REFCOR for was designed for language learners at B2 level. The results point out that pre-college level academic English apply TMD markers to a much greater extent than tertiary academic English. The overabundance of TMD devices in the BIOCOR can be explained by the aim of the corpus: to impart knowledge for teenagers. This aim is best reached through making the logical flow of the texts linguistically as overt as possible. The results of the high TMD ratio in the BIOCOR, however, gives no explanation for the challenges 10th grade bilingual students meet when attempting to comprehend the corpus.

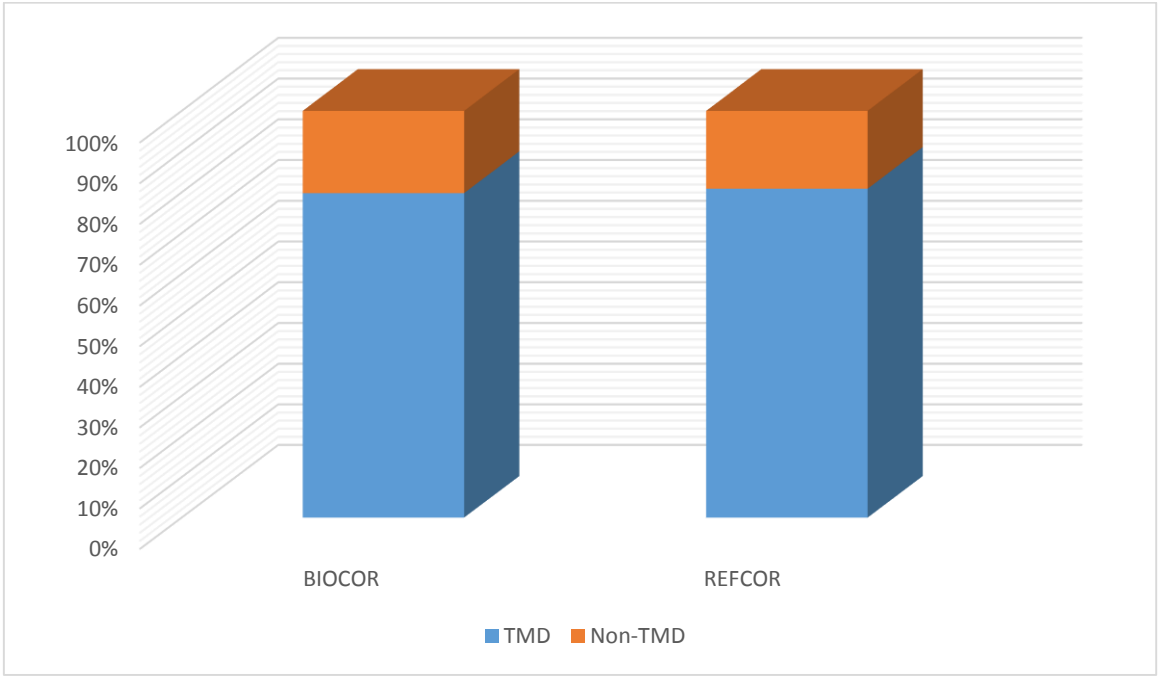

Diagram 19 The ratio of TMD and non-TMD sentences in the two corpora

After unveiling the extremely high frequency of TMD markers in the BIOCOR, now let us turn our attention to the quality of TMD devices through examining the appearance of the eleven different types of TMD functions in the corpus (see Diagram 20). The most numerously applied TMD function in the BIOCOR is addition. It appears in 36% of the sentences of the corpus, which means that more than every third sentences on average contains an additive TMD marker. The function of contrast, on the other hand, is less frequently used in the BIOCOR (6% of the sentences); it appears in every 17th sentence of the corpus. In a similar manner, the function of purpose is overtly expressed in only 5% of the sentences of the BIOCOR. A TMD marker indicating purpose is used in every 20th sentence in the corpus. In an even smaller proportion (1% of the sentences) does the TMD marker conveying the logical relationship of giving reason appear in the BIOCOR. The TMD marker for denoting the result of a logical relationship is used even more sparingly, only 0.55% of the sentences of the BIOCOR applies this linguistic device. A more widely appearing TMD

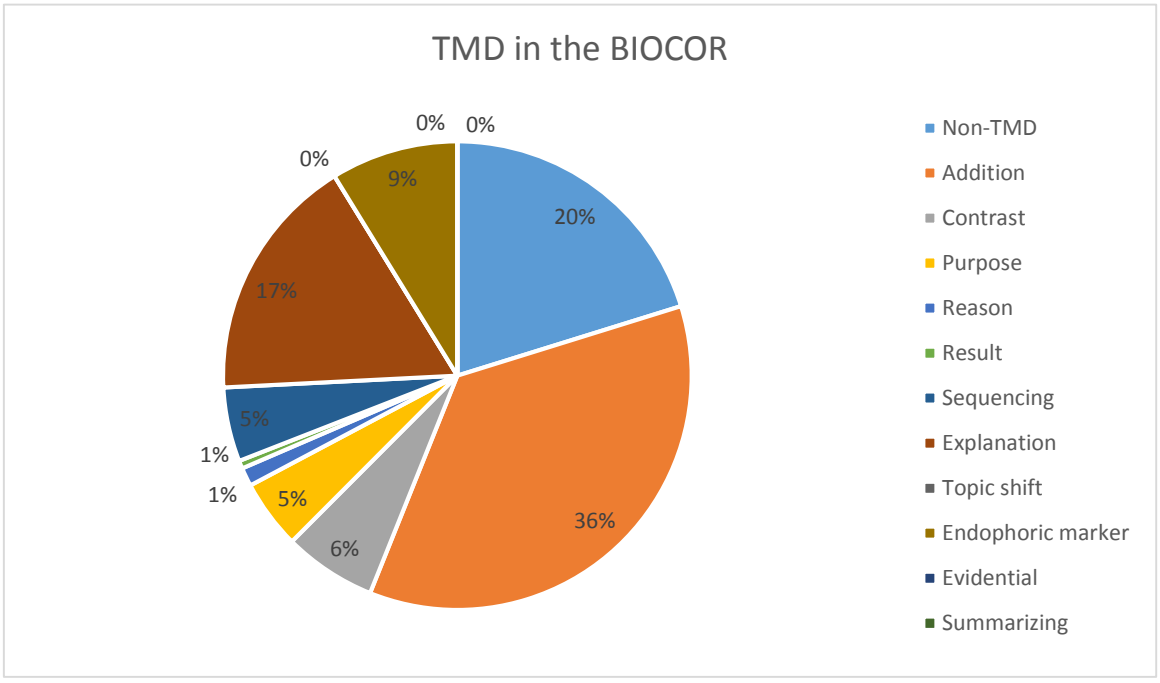

Diagram 20 The frequency of TMD functions in the BIOCOR

function is that of sequencing, which is used in 5% of the sentences of the BIOCOR. In other words, every 20th sentence in the corpus overtly displays a sequential logical relationship. The function of explanation is even more frequently expressed through TMD markers, 17% of the corpus, or ever sixth sentence explicitly explicates the propositional content of the biology texts. In contrast, topic shifts are not transparent through TMD markers in the BIOCOR to any extend (0% of the sentences contains such linguistic devices). Reference within the corpus through endophoric TMD markers is more regularly apparent in the BIOCOR, however. Every 11th sentence (9% of the sentences) bears a linguistically visible relationship with other parts of the corpus. Nonetheless, the functions of providing evidentials and summarizing propositional content are not carried out through TMD marker (0% of the sentences in the BIOCOR apply them).

A slightly less evenly distributed appearance of the eleven different types of TMD functions can be observed in the REFCOR (see Diagram 21). The overwhelming majority of

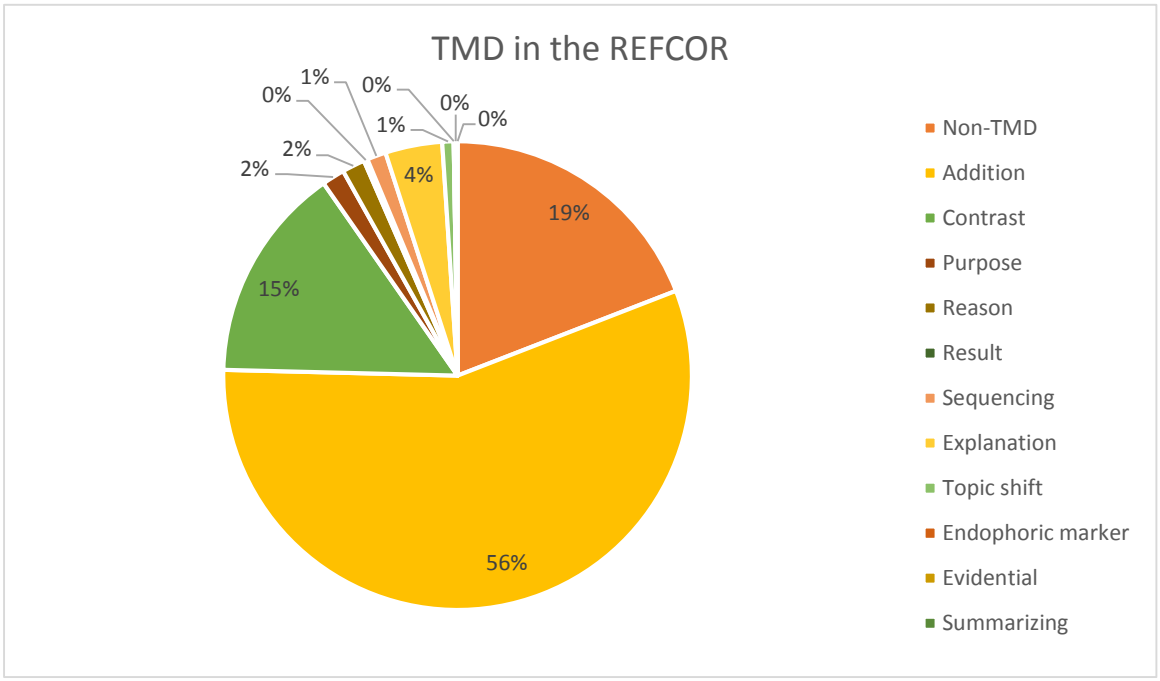

Diagram 21 The frequency of TMD functions in the REFCOR

TMD markers in the REFCOR belongs to addition. 56% of the sentences in the corpus, that is, more than every second sentence contains a TMD marker which overtly expresses an additive meaning. The second most typical TMD function in the REFCOR is that of contrasting, which appears in 15% of the sentences of the corpus. No fewer than nearly every sixth sentence clearly articulates the logical relationship of contrast in the REFCOR. The TMD functions of giving purpose and reason are much more modestly applied, however. These two functions are overtly represented in only 2-2% of the sentences of the REFCOR. This covers a humble proportion of a mere appearance in every 50th sentence in the corpus. Even more underrepresented is the TMD function of showing results, which fails to be used in the REFCOR to any extent. The presence of sequencing through TMD markers is tremendously timid, too. Linguistic devices of sequencing appear in the REFCOR barely up to 1% of the sentences. The function of providing explanation expressly is more favoured in the REFCOR. Every 25th sentence (4% of the sentences) contains a TMD marker that makes explication linguistically visible. Topic shifting, however, is moderately indicated through TMD devices in the REFCOR. Only 1% of the sentences contains a TMD marker which signposts a change in the flow of topics in the corpus. Three of the TMD functions completely lack representation in the REFCOR: neither endophoric markers, nor evidentials, nor summarizing TMD markers help the understanding of the corpus.

Finally, the comparison of the distribution of the eleven TMD functions (see Diagram 22) and that of the particular TMD markers across the two registers (the BIOCOR and the REFCOR) give a better understanding to what extent the BIOCOR might pose challenges for 10th grade bilingual students due to the TMD practice of the biology corpus.

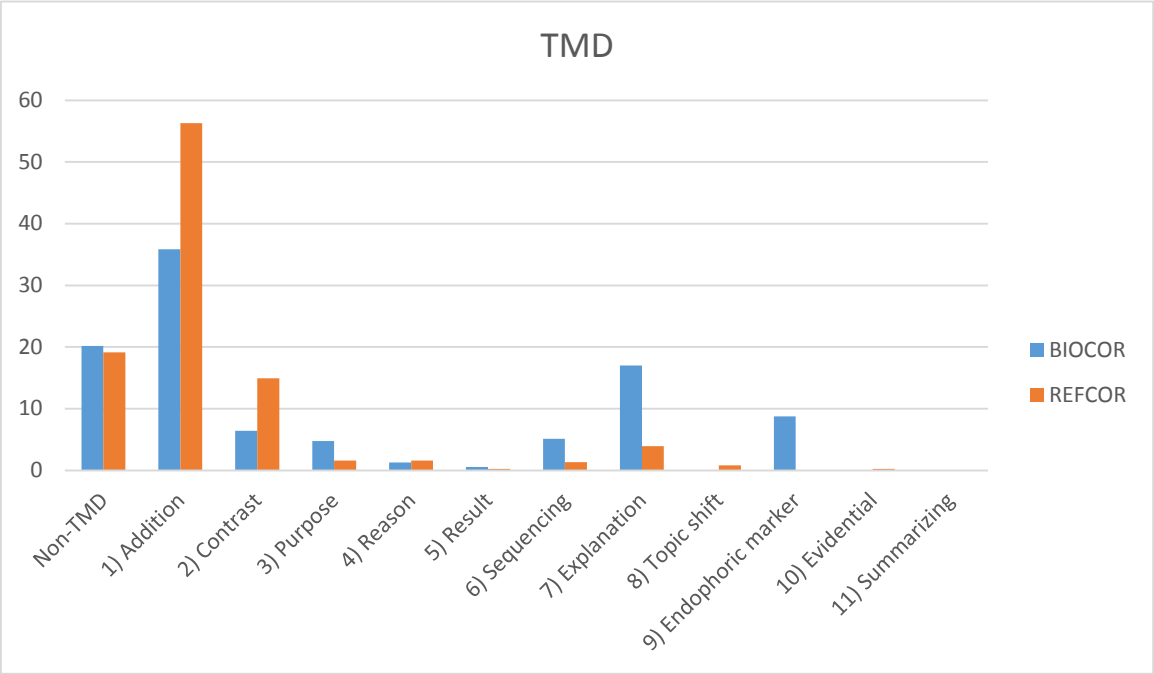

Diagram 22 Comparison of the frequency of TMD functions in the two corpora

1) TMD markers with an additive function are considerably less prevalent in the BIOCOR (36%) than in the REFCOR (52%). The versatility of the additive TMD markers, however, is nearly twice greater in the BIOCOR than in the REFCOR. The use of additive TMD markers in the latter corpus is reduced to 'and,' plus a few instances of 'also,' and 'too.' In contrast, the BIOCOR relies on more TMD markers expressing addition, besides the above mentioned ones, it also uses 'in addition,' and 'as well.' The wider variety of TMD markers with the function of addition in the BIOCOR might pose a meagre challenge for 10th grade bilinguals who are only prepared to process a limited number of TMD markers of that kind. The difficulty does not seem to be insurmountable, though, as the additional TMD markers in the BIOCOR ('in addition,' and 'as well') appear to convey their functions of addition rather visibly.

2) The function of contrast is expressed nearly three times less frequently through TMD markers in the BIOCOR (6%) than in the REFCOR (15%). The range of contrastive TMD

markers in the BIOCOR is noticeably narrower than in the REFCOR, which applies nine different types of contrastive linguistic devices (‘but,’ ‘although,’ ‘though,’ ‘even though,’ ‘while,’ ‘however,’ ‘nevertheless,’ ‘in spite of,’ and ‘on the contrary’). The BIOCOR expresses contrast with an overlap of TMD markers, which are, however, smaller in number (‘but,’ ‘although,’ ‘though,’ ‘however,’ ‘on the other hand,’ and ‘in contrast’). The greater predictability of the contrastive TMD practice of the BIOCOR makes the corpus easier to process.

3) TMD markers expressing purpose are applied more than twice as frequently in the BIOCOR (5%) as in the REFCOR (2%). The types of TMD markers which convey the function of giving purpose are extremely confined in the REFCOR, where they are reduced to one single marker (‘so’). In contrast, the same function is expressed much more multifariously in the BIOCOR, where four different types occur (‘so,’ ‘in order to,’ ‘so as to,’ and the infinitive of purpose). The greater variety of purpose-expressing TMD markers in the BIOCOR might challenge those 10th grader bilinguals who are used to processing only one of them. Two of these TMD markers which are used in the BIOCOR but fail to appear in the REFCOR seem to be more demanding. In the case of ‘so as to’ the student needs to recognize that the three words functionally belong together, which is not definitely obvious. The infinitive of purpose is a different kind of challenge, since here the reader is required to distinguish in which function and consequently with what meaning the infinitive is used. As the REFCOR, which completely lacks these TMD markers, provides no such practice for 9th grade students, bilingual students who are lower-achievers in ESL can struggle with the BIOCOR from this respect.

4) The moderate use of the TMD markers giving reason displays a minuscule difference between the BIOCOR (1%) and the REFCOR (2%). The BIOCOR relies on only one such TMD marker ('because'), while the REFCOR is more abundant in reason-giving TMD markers and uses various others ('because (of),' 'as,' 'since,' and 'due to'), too. From this point of view, the BIOCOR is less challenging to understand than the REFCOR, and thus should pose no challenges for 10th graders to process.

5) The function of showing result is expressed through TMD markers on a small scale in both corpora, 1% in the BIOCOR and 0.3% in the REFCOR. Though the density of such TMD markers is broadly similar in the two registers, there is no overlap in the occurrence of the particular TMD markers within the same functional category (the BIOCOR uses 'thus,' and 'hence,' while the REFCOR applies 'as a result'). Although it might be argued that the small rate of resultative TMD markers makes the logical structure of the BIOCOR more opaque and thus more demanding to comprehend, it is important to note that the REFCOR applies these TMD markers even more sparingly. Consequently, 10th grade bilingual students, who were to read through the REFCOR in the 9th grade became impervious to such challenges, are well-trained to tackle such difficulties.

6) The frequency of TMD markers expressing sequencing displays a notable difference between the two registers, the BIOCOR (5%) applies this TMD function five times more often than the REFCOR (1%). The most prevalent sequencing TMD marker ('then') appears in both registers recurrently. Besides, the BIOCOR applies a sequence starter TMD marker ('first') intermittently. Since the logical progression of the content information in the BIOCOR is conspicuously more transparent through the abundant use of sequencing TMD

markers than in the REFCOR, 10th grade bilingual students meet no challenges from this respect when processing the BIOCOR.

7) The use of explanatory TMD markers in the two corpora is dissimilar to an enormous extent. The BIOCOR (17%) applies more than four times as many TMD markers expressing explanation than the REFCOR (4%). Both corpora uses the markers 'such as,' 'in other words,' and 'this / which means,' while the markers 'for instance,' and 'defined as' appear intermittently only in the REFCOR. The BIOCOR, on the other hand, tends to use the explanatory markers 'known as,' and 'so,' along with the numerously repeated 'for example,' and 'called' TMD markers. Since the BIOCOR is exceedingly more transparent with regards to linguistically revealing explanatory relationships between ideas, it cannot be described as more difficult than the REFCOR from this aspect either.

8) Explicitly showing topic shifts through TMD markers is not characteristic of either of the registers. The BIOCOR uses no such markers (0%), while the REFCOR applies them in an extremely minor portion of its sentences (1%). Although the BIOCOR avoids using TMD markers expressing topic shift, processing the register can hardly be labelled as challenging from this respect for the 10th grade students. On the one hand, the REFCOR, which prepares the bilingual students for academic reading, fails to contain considerably more topic shift TMD markers (only few instances of 'well,' and 'now' are used), thus the 10th graders are accustomed to this meagre practice. On the other hand, the BIOCOR crystal clearly reveals its topic shifts through using other markers, namely applying headings to its paragraphs.

9) The function of referring to different parts of the text is broadly fulfilled in the BIOCOR through TMD markers. Nearly every 10th sentence of the corpus (9%) uses endophoric

markers. In contrast, such markers tend to be completely missing from the REFCOR (0%). The fact that the BIOCOR gives clear references with its wide variety of referential TMD markers ('see,' 'later,' 'page X,' 'Table X,' and the extensively used 'Figure X') as to where the reader can find different pieces of information within the text reduces the level of difficulty of the register.

10) Both registers have a tendency to be void of evidential TMD markers. The BIOCOR applies no such markers at all, while the REFCOR relies on them to a trifle extent (0.3%), where the markers 'Z claims' appears. The fact that evidential TMD markers are not typically used in the BIOCOR sheds light on a characteristic trait of the register: pre-college academic writing does not strive to support its claims through the results of other findings but tends to impart widely accepted knowledge.

11) A similar pattern emerges when we rank TMD markers which express the function of providing a summary. Neither the BICOR, nor the REFCOR applies any such markers. Although the BIOCOR might be stated to be more straightforward for its target readers to be understood once it contained explicit summaries, this shortcoming cannot be seriously treated as a possible challenge for the 10th grade bilinguals since they were expected and tested to process texts of similar nature from this respect.

Summarizing the results of the comparative TMD analysis of the BIOCOR and the REFCOR, it can be stated that the BIOCOR is barely in position of posing serious challenges for 10th grade bilingual students. On the one hand, the BIOCOR shows an extremely high rate of TMD markers (80%), which indicates that the flow of ideas in the register is clearly transparent and the logical relationships are heavily signposted for the reader. On the other

hand, the various types of TMD functions are more evenly distributed in the BIOCOR than in the REFCOR. As a result, the BIOCOR provides visible guidance through linguistic devices in more types of logical relationships. Finally, the BIOCOR has a tendency of greater predictability with regard to TMD markers as the register applies fewer types of TMD markers within the same function than the REFCOR. This characteristics of the BIOCOR makes the register less demanding to be processed than the REFCOR. In those cases where the BIOCOR uses more types of TMD markers within the same function than the REFCOR, that is, where the TMD signposting of the BIOCOR is less predictable than that of the REFCOR, the BIOCOR still remains to be easily understood as most of the greater variety of TMD markers show their function clear.

Applying all the components of the POTAI to the BIOCOR, the results of the analysis clearly indicate that the biology texts show no clear signs of being challenging for the 10th grade bilingual students. Moreover, the BIOCOR appears to be more simplistic to be processed from multifarious linguistic aspects than the REFCOR. Thus it can be concluded that the perceived difficulties these bilingual students face do not stem from the texts themselves but originate from some different source.

5 Pedagogical implications: a checklist for ESL and biology ESP teachers

In order to help ESL and biology ESP teachers choose finely tuned texts which can be used for preparing bilingual students for their biology studies in English, the characteristic traits of the BIOCOR are summarized in Table 41. The register of the biology textbook for pre-college students is described here from all the aspects analysed in the current research, thus the summative chart provides an extensive overview of the distinguishing qualities of the register. The linguistic information about the register is reviewed in a tabular form (Table 41), so that the boxes' straightforward categories and briefly-worded descriptions readily support the educators' decision of selecting texts which are at an appropriate level of English for being used in ESL or ESP courses preparing secondary students for biology studies. Table 41 also gives room for remarks which show various further directions in developing teaching materials for biology ESP specifications. Reference to relevant sections of the dissertation, where more detailed information about the results is available, is also collected in Table 41. The point of reference in providing the level of difficulty of the register was that of the REFCOR, that is, the B2 level.

Component of the POTAI		Level of difficulty	Remark
Lexis	Frequently occurring words (Section 4.1.1)	Ranges from A1 to B2 (one single exception: *organism* (C1))	**ESP vocabulary:** **Biology terms:** *parasite, cell, bacteria, virus, growth, amoeba, reproduce, malaria, blood, tapeworm* **Academic English:** there is no instance of academic English in the register
	Keyness (Section 4.1.2)	Ranges from A1 to B2	**ESP vocabulary:** **Biology terms:** *intestine, agar, gut, genus* **Academic English:** *process*
	Lexical density (Section 4.1.3)	B2	not packed with more information than general English texts

Grammatical phenomena	Tenses and tense related structures (Section 4.2.1)	below B2	**The most dominant tenses:** the present simple and the past simple **Completely absent tenses:** the past continuous, the past perfect continuous, the *'used to'* structure, the future continuous, the future perfect simple and continuous, and the *'going to'* structure
	Conditional structures (Section 4.2.2)	below B2	**The most typical structure:** zero conditional **Completely absent structures:** third and mixed conditionals
	Passive voice and causative structures (Section 4.2.3)	below B2	**Most typical:** passive voice with a direct object **Completely absent structures:** passive voice with an indirect object, causative structures such as *'have it done,' 'get it done,' 'needs doing,'* and *'make somebody do something'*
	Relative clauses (Section 4.2.4)	below B2	**Most common:** defining relative clauses with a relative pronoun **Less common:** defining relative clauses without a relative pronoun **Not typical:** non-defining relative clauses Completely absent: progressive participles (both in the present and in the past)
	Nominal relative clauses (NRC) (Section 4.2.5)	below B2	**Most common:** NRC without a reporting verb without time shift or with an infinite verb **Less common:** reported open questions **Not typical:** NRC sentences with a reporting verb

	Infinitives (Section 4.2.6)	above B2	**Most typical:** simple infinitives, passive infinitives **Less common:** progressive passive infinitives **Completely absent:** perfect passive, perfect progressive, perfect progressive passive infinitives
	Prepositions at the end of sentences (Section 4.2.7)	below B2	Completely absent
	Modal verbs (Section 4.2.8)	below B2	**Most typical:** ability in the present (*'can'*), level of certainty (*'may'*), obligation in the present (*'must'*)
Sentence complexity	Sentence length (Section 4.3.1)	below B2	**Most typical:** relatively short sentences (4-17 words)
	Packet length (Section 4.3.2)	B2	**Most typical:** 8-12 words (typical of written English) **Less frequent:** 13-20 words
	Readability indices (Section 4.3.3)	below B2	one to four years easier than B2
	Syntactic structure (Section 4.3.4)	below B2	**Most typical:** one-clause-long simple sentence **Less frequent:** two-clause-long sentence **Not typical:** three-clause-long sentence
Textual metadiscourse	Addition	less overt than B2	versatile TMD markers
	Contrast	less overt than B2	narrow range of TMD markers
	Purpose	more overt than B2	versatile TMD markers
	Reason	less overt than B2	narrow range of TMD markers
	Result	more overt than B2	narrow range of TMD markers
	Sequencing	more overt than B2	narrow range of TMD markers
	Explanation	more overt than B2	versatile TMD markers
	Topic shift	less overt than B2	no such instances

	Endophoric markers	more overt than B2	versatile TMD markers
	Evidentials	less overt than B2	no such instances
	Summarizing	B2	no such instances

Table 41 The characteristic traits of the biology textbook register

With regard to defining the CEFR level of the lexis in focus, an online software developed by the Lifelong Learning Programme of two departments of the University of Cambridge (Cambridge University Press and Cambridge English Language Assessment, http://vocabulary.englishprofile.org) was applied. The difficulty of general lexis in the biology textbook register, both in the case of frequently occurring lexis and in that of high keyness words, does not tend to go beyond the B2 level: it ranges from A1 to B2. The frequently occurring lexis in the entirety of the BIOCOR contains no more than one single instance of above-B2 vocabulary (*'organism'* (C1)), ten instances of biology terms ('*parasite,' 'cell,' 'bacteria,' 'virus,' 'growth,' 'amoeba,' 'reproduce,' 'malaria,' 'blood,'* and *'tapeworm'*), and no instances of academic English at all. The key vocabulary of the register, which consistently, without any exception, avoids ranging above the B2 level, comprises four biology terms ('*intestine*,' '*agar*,' '*gut*,' and '*genus'*) and one single instance of academic English (*process*). In harmony with its lexical plainness, the lexical density of the biology textbook register does not reach the expected level of informative, academic written prose, but has the same value as that of general English texts. That is, the biology textbook register fails to contain more information than general English texts at the B2 level.

Compared to the CEFR B2 level, the grammatical phenomena which are typical of the biology textbook register reveal great simplicity. The most representative tenses are the simple ones (the present simple and the past simple in particular), and a large number of more complex tenses are either underrepresented or totally absent from the register. Among the

conditional structures the simplest one, the zero conditional, is the most common; and the more complex ones, the third and the mixed conditionals, are absolutely absent from the biology textbook register. The occurrence of passive voice tends to be numerous, however, statistically there is no significant difference in its frequency in the biology textbook register and that of general English texts. The use of relative clauses displays that the biology textbook register avoids linguistically complicated sophistication: the most straightforward way of offering definitions (through the use of defining relative clauses with a relative pronoun) is typical, while progressive participles are not applied at all. Nominal relative clauses are used in order to address the target readers directly, which increases the personal pitch of the biology textbook register, and reduces its academic tone. The relatively frequent use of passive infinitives is one of the few traits of the biology textbook register which poses challenges for the non-native independent user of English (B2). In contrast, the complete lack of prepositions at the end of sentences lends an easy accessibility to the biology textbook registers. The range of modal auxiliaries in the biology textbook register is limited: besides the three frequently used ones (*'can,' 'may,'* and *'must'*), the great variety of modal auxiliaries typical at B2 level is absent from the register.

The length of sentences indicates that the biology textbook register does not aim to use verbose sentences but strives to apply shorter ones than typical at B2 level. Considering the length of packets in the biology textbook register, similarities with general English texts at B2 level can be observed. Sentence complexity revealed through various grade level readability indices also points towards the fact the biology textbook register tends to be bellow B2 level. This characteristics is in harmony with the most typical syntactic structure of the biology textbook register, the simplest, one-clause-long sentence.

The biology textbook register displays the presence of an extremely high rate of TMD markers, which shows the wide-ranging overtness of the register: the flow of ideas is clearly transparent and the logical relationships are heavily signposted for the reader. In comparison with general English texts at a B2 level, the biology textbook register reveals more types of logical relationships through TMD markers, and at the same time relies on fewer types of TMD markers. Both of these traits allow the register to be described as bellow B2 level.

Overviewing the various linguistic aspects of the biology textbook register, the appropriate level of the texts which prepare bilingual students for their biology studies in English is below the CEFR B2.

6 Conclusion

The closing chapter of the dissertation concludes the findings, briefly summarizes the novelty of the research project, and points out possible future areas of research.

6.1 Summary of the results

The current theoretically and pedagogically motivated study aimed at finding possible means to describe the prevailing register features of the biology texts used at an English-Hungarian bilingual secondary school. The analytical tool, which was developed from the perspective of the ESL teacher, was applied to this corpus in order to measure its level of difficulty with the intention of revealing to what extent the general English reading texts assigned in the intensive language preparatory course at the bilingual secondary school enable 9th grade students to handle the biology texts used in the subsequent term.

As one single linguistic trait cannot fully describe a register, a comprehensive analytical tool (POTAI) was developed to tap the register-specific traits which are of relevance for ESL teachers. Each component of the POTAI was examined separately (in line with Research Question 1) to yield reliable and valid data concerning the determination of dominant register features (see Section 3.3 on pp. 72-117). Since all the components proved to be capable of providing such data, their synergy is a reliable instrument that produces data which can describe written registers with a high rate of validity. It is important to note that the more components of the POTAI are applied to a particular register, the higher rate of validity can be achieved. The reason behind this is that none of the components of the POTAI is sufficient in itself to identify traits that completely map a register; however, their complexity involves various perspectives, which can display a detailed description of the register.

Addressing the second research question, the in-depth text analysis yielded sufficient data that reveal the characteristic linguistic features of the BIOCOR in comparison with those of the REFCOR. The diversity of the various components of the POTAI produced data which clearly converge in one direction (see Chapter 4 on pp. 118-143 or Chapter 5 on pp. 203-208). All the components of the POTAI (see Section 3.3.5 on p. 116) generated results which demonstrate that the BIOCOR fails to be more complex linguistically than the REFCOR, the BIOCOR does not surpass the REFCOR in difficulty. Moreover, the BIOCOR has a strong tendency to be more simplistic from this regard than the REFCOR. Although science textbooks are expected to use academic English (Cserép, 1997), science textbooks for secondary students should be distinguished from those written for tertiary studies. The findings of the present research reveal that a pre-college textbook bears the traits of popularizing literature rather than those of academic prose, which result harmoniously nests within Shapiro's (2012) model arguing for secondary textbooks not being academic literature. The results which unanimously demonstrate the lack of linguistic complexity in the biology textbook register confirm that the general reading texts assigned in the 9th grade prepare bilingual students well for their academic studies in English the following year. What is more, processing the REFCOR provides a firm linguistic grounding for the bilingual students, which is above the linguistic level minimally needed for comprehending the biology textbook in the 10th grade. The prevailing linguistically straightforward character of the BIOCOR, however, suggests that the perceived challenges 10th grade bilingual students face during their studies in English is not explicable in terms of the language of their textbook, that is, it stems from a different source.

Relying on the knowledge gained from mapping out the characteristic linguistic features of the biology textbook register in detail, recommendations can be formulated for

educators (Research Question 3). To select finely-tuned texts which are at the appropriate level for preparing secondary students for their biology studies in English, ESL and biology ESP teachers need to be cognizant of the fact that the linguistic level of the register is below the CEFR B2 (for further details see Chapter 5 on pp. 203-209). Despite the fact that the bilingual immersion programme aims at preparing students for passing a Cambridge B2 exam (FCE), the readings bilingual students are eventually assigned to comprehend during their studies do not require such an excellent command of English.

6.2 Novelty of the research

The results of the present research contribute to four main areas of study. Within the field of corpus linguistics, scant attention has been paid to the development of a comprehensive text-analytical instrument that produces relevant data for ESL teachers (see Section 2.5 on pp. 33-37). Thus the design of a novel instrument which yields relevant information for educators (ESL and ESP teachers) breaks new ground by filling this theoretical and pedagogical lacuna. Secondly, the application of the text-analytical instrument provides valuable new results for the field of register analysis as well. The exploratory, in-depth linguistic analysis of the register of biology textbooks for secondary school students brings new knowledge in a yet uncharted area, which can be used for developing ESL and biology ESP teaching materials. Thirdly, the results of the current study also enhance the field of ESP research. The fine-grained analysis of the register of biology textbooks for secondary students offers useful insights into what is required linguistically from the target readers to successfully comprehend texts in the register. This knowledge enriches our understanding of what needs to be involved both in compiling biology ESP teaching materials and in developing biology ESP course syllabi. Fourthly, the findings of the research project may be integrated into the field of bilingual education in a novel manner, too. Awareness of the

linguistic needs bilingual students should master in order to process their textbooks can feed into the improvement of the intensive language course of the bilingual immersion programme, which was established in Hungary nearly three decades ago and whose linguistic aims have not yet been revised at the school ever since. Besides the local application, the results of the study call for the attention of other international bilingual schools too, where ESOL courses are provided for the students.

6.3 Areas for future research

The results of the register analysis do not give solid ground for the explication of the challenges which the 10th grade bilingual students face when processing the biology textbook; on the contrary, there is no room for doubt that the findings illustrate the relative straightforwardness of the register compared to general English texts. For this reason, further contextualization is planned to be carried out, which can lead to detecting different possible sources for the difficulties the students perceive. One of the future steps of the research involves classroom observations, where the regular practice of biology classes can be discovered. The on-sight observations can directly provide various types of information, e.g., the skills which students are required to use in class, the styles of teaching and learning in class, or the various ways of assessments. Another crucial step of the extension of the present research is an interview with students, which can advance our understanding of the possible reasons of the challenges 10th grade students face when processing their English-language textbooks, through exploring several different perspectives in order to gain insights into the heart of the problem. An interview study might give the chance for students to verbalize what they recognize to be the main educational difference between the 9th and the 10th grade (revealing what students find problematically difficult), to reflect on their motivation (which

immensely influences comprehension (DuBay, 2004)), or to shed light on the study techniques they use in class or at home.

In order to cooperate with the secondary school and to bring research knowledge into teaching practice, a group interview session is also planned with the subject teachers instructing in the 10th grade and ESL teachers working in the intensive language programme. The group interview is intended to facilitate communication between the different departments of the school, where subject teachers and ESL teachers can discuss what steps they need to make for their smooth working together. With the view of revising the intensive language course of the bilingual immersion programme, the linguistic characteristics of the textbooks of other subjects taught in English are also planned to be mapped out through the application of the POTAI.

Although several more perspectives could be explored in order to fully understand the reasons why 10th grade bilingual students find processing their textbooks in English challenging, the current study accomplished its outlined aims and objectives to design a pedagogically oriented text-analytical tool which is capable of gaining linguistic data applicable for ESL and ESP teachers. The other intention of this doctoral dissertation, to describe the register of the biology textbook for secondary school students through applying the newly-developed instrument, was also fulfilled. The experience gained from conducting the present research project shall guide the researcher in her future research activities.

References

Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. Amsterdam: Benjamins.

Aijmer, K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55-76). Amsterdam: Bemjamins.

Akar, D., & Louhiala-Salminen, L. (1999). Towards a new genre: A comparative study of business faxes. In F. Bargiela-Chiappini, & C. Nickerson (Eds.), *Writing business: Genres, media and discourses* (pp. 207-226). London: Longman.

Alderson, J. C., & Urquhart, A. H. (1984). *Reading in a foreign language*. London: Longman.

Allen, J. P., & Widdowson, H. G. (1974). Teaching the communicative use of English. *International Review of Applied Linguistics*, *12*(1). 1-21.

Aston, G. 1995. Corpora in language pedagogy: Matching theory and practice. In G. Cook, & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics* (pp.257-270). Oxford: Oxford University Press.

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

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

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

Bailey, S. (2011). *Academic writing: A handbook for international students.* New York: Routledge.

Baker, C., & Jones, S. P. (1998). *Encyclopedia of Bilingualism and Bilingual Education*. Clevedon: Multilingual Matters.

Barber, C (1985). Some Measurable Characteristics of Modern English Prose. In J. M. Swales (Ed.), *Contributions to English syntax and phonology*. *Episodes in ESP* (pp. 1-21). Oxford: Pergamon.

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

Beaugrande, R. d. (1997). *New foundations for a science of text and discourse: Cognition, communication, and the freedom of access to knowledge and society*. Norwood, N.J: Ablex.

Beaugrande, R. d. (2001). Large Corpora, Small Corpora, and the Learning of Language: In M. Ghadessy (Ed.), *Small Corpus Studies and ELT. Theory and Practice* (pp. 3-28). Philadelphia, PS: John Benjamins.

Beaugrande, R. d., & Dressler, W. U. (1981). *Introduction to text linguistics*. London: Longman.

Beauvais, P. (1989). A speech-act theory of metadiscourse. Written Communication, 61, 11-30.

Bednarek, M. (2006). *Evaluation in media discourse: analysis of a newspaper corpus*. London: Continuum.

Bell, A. (1991). *The language of news media*. Oxford: Blackwell.

Berkenkotter, C., & Huckin, T. N. (1993). Rethinking genre from a sociocognitive perspective. *Written Communication, 4,* 475-509.

Bernardini, S. (2000). Systematising serendipity: proposals for concordancing large corpora with language learners. In L. Burnard, & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 225-234). Frankfurt: Peter Lang.

Bernardini, S. (2004). Corpora in the classroom: an overview and some reflections on future developments. In J. Sinclair (Ed.), *How to use corpora in language teaching* (pp. 15-38). Amsterdam: Benjamins.

Bhathia, V. K. (1993). *Analysing genre: language use in professional settings*. London: Longman.

Bhatia, V. K. (1998). Generic patterns in fundraising discourse. *New directions for philanthropic fundraising, 22,* 95-110.

Bhatia, V. K. (2002). A generic view of academic discourse. In J. Flowerdew (Ed.), *Academic discourse* (pp. 21-39). Harlow: Longman.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, 62, 384-414.

Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

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

Biber, D. (1991). Oral and literate characteristics of selected primary school reading materials. *Text, 11,* 79-96.

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

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

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

Biber, D. (2006). *University language: a corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.

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

Biber, D., Connor, U., & Thomas, A. (2007). *Discourse on the move: using corpus analysis to describe discourse structure*. Amsterdam: John Benjamins.

Biber, D., Conrad, S., & Reppen, R. (1994). Corpus-based issues in applied linguistics. *Applied Linguistics, 15*, 169-189.

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

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

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

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

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

Biber, D., & Finegan, E. (1997). Diachronic relations among speech-based and written registers in English. In T. Nevalainen, & L. Kahlas-Tarkka (Eds.), *To explain the present: studies in the changing English language* (pp. 253-275). Helsinki: Societe Neophilologique.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. London: Longman.

Blakesley, D. & Hoogeveen, J. L. (2011). *Writing: A manual for the digital age.* Boston, MA: Wadsworth.

Bloomfield, L. (1933). *Language*. New York: Holt.

Bognár, A. (2000). A két tanítási nyelvű oktatás 12 „nem tucat" éve a magyar közoktatásban. *Modern Nyelvoktatás, 6*(1), 56-63.

Bormouth, J. R. (1966). Readability: a new approach. *Reading Research Quarterly, 1*, 79-132.

Boyle, M., & Warwick, L. (2014). *Skillful reading and writing. Student's book 4.* London: Macmillan.

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

Bruce, B., Rubin, A., & Starr, K. (1981). Why readability formulas fail. *IEEE Transactions on Professional Communication*, *24*, 50-52.

Bruthiaux, P. (1994). Me Tarzan, you Jane: linguistic simplification in "personal ads" register. In D. Biber, & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 136-154). New York: Oxford University Press.

Bruthiaux, P. (1996). *The discourse of classified advertising: exploring the nature of linguistic simplicity*. Oxford: Oxford University Press.

Bunton, D. (1999). The use of higher level metatext in PhD theses. *English for Specific Purposes, 18*(Supplement 1), S41-S56.

Bunton, D. (2002). Generic moves in Ph.D. thesis introductions. In J. Flowerdew, (Ed.), *Academic discourse* (pp. 57-75). Harlow: Longman.

Bunton, D. (2005). The structure of PhD conclusion chapters. *Journal of English for Academic Purposes, 4*, 207-224.

Camiciottoli, B. C. (2007). *The language of business studies: A corpus-assisted analysis*. The Netherlands: John Benjamins Publishing Company.

Campbell, P. (1975). The personae of scientific discourse. *Quarterly Journal of Speech, 61*, 391-405.

Candlin, C. N., Bruton, C. J., & Leather, J. H. (1976). Doctors in Casuality: Specialist course design from a data base. *International Review of Applied Linguistics*, 14(*3*), 245-273.

Carmines, E.G., & Zeller, R.A. (1991). *Reliability and validity assessment*. Newbury Park: SAGE Publications.

Carter, R. (2004). *Language and creativity: the art of common talk*. London: Routledge.

Carver, R. P. (1990). Predicting accuracy of comprehension from the relative difficulty of material. *Learning and Individual Differences*, *2*, 405-422.

Chall, J. S. (1958). *Readability: An appraisal of research and application*. Columbus, OH: Ohio State University Press.

Chall, J. S. (1984) Readability and prose comprehension. Continuities and discontinuities. In J. Flood (Ed.), *Understanding reading comprehension: cognition, language, and the structure of prose*. Neward, DE: International Reading Association.

Chall, J. S., & Conrad, S. S. (1991). *Should textbooks challenge students? The case for easier or harder textbooks*. New York: Teachers College Press.

Chall, J. S., & Dale, E. (1995). *Readability revisited, the new Dale-Chall readability formula*. Cambridge, MA: Brookline Books.

Chazal, E., & Moore, J. (2013). *Oxford EAP. Advanced*. Oxford: Oxford University Press.

Chen, Q., & Donin, J. (1997). Discourse processing of first and second language biology texts. Effects of language proficiency and domain-specific knowledge. *The Modern Language Journal*, *81*(2), 209-227.

Cheng, X., & Steffensen, M. (1996). Metadiscourse: A technique for improving student writing. *Research in the Teaching of English, 30(*2), 149-181.

Cheryl, A. E. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing, 4*(2), 139-155.

Christie, F. (1992). The "received" tradition of literacy teaching: The decline of rhetoric and corruption of grammar. In B. Green (Ed.), *The insistence of the letter: Literary studies and curriculum theorizing* (pp. 75-106). London: Falmer Press.

Christie, F. (2002). *Classroom discourse analysis*. London: Continuum.

CIDE (1995). *Cambridge international dictionary of English*. Cambridge: Cambridge University Press.

Cindy, C., & James, P. (2007). *The psychological functions of function words*. New York: Psychology Press.

COBUILD (1990). *Collins COBUILD English grammar*. London: Harper Collins.

COBUILD (1995). *Collins COBUILD English dictionary*. London: Harper Collins.

Cohen, A., Glasman, H., Rosenbaum-Cohen, P. R., Ferrara, J., & Fine, J. (1988). Reading English for specialised purposes: Discourse analysis and the use of standard informants. In. P. Carrell, J. Devine, & D. Eskey (Eds.), *Interactive approaches to second language reading* (pp. 162-167). Cambridge: Cambridge University Press.

Cohen, C., & Mannion, L. (1980). *Research methods in education*. Croom Helm.

Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology, 60*(2), 283-284.

Collins, P. (1991). *Cleft and pseudo-cleft construction in English*. London: Routledge.

Connor, U. (1996). *Contrastive rhetoric: cross-cultural aspects of second-language writing*. Cambridge: Cambridge University Press.

Connor, U., & Mauranen, A. (1999). Linguistic analysis of grant proposals. European Union research grant. *English for Specific Purposes, 18*(1), 47-62.

Connor, U., Precht, K., & Upton, T. (2002). Business English: Learner data from Belgium, Finland, and the U.S. In. S. Granger, J. Hung, & Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 175-194). Amsterdam: John Benjamins.

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

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

Conrad, S., & Biber, D. (2000). Adverbial marking of stance in speech and writing. In S. Hunston, & G. Thompson (Eds.), *Evaluation in text: authorial stance and the construction of discourse*. Oxford: Oxford University Press.

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

Cope, B., & Kalantzis, M. (1993). *The powers of literacy: A genre approach to teaching writing*. Philadelphia, PA: University of Pittsburgh Press.

Coulthard, M. (1985). *An introduction to discourse analysis*. London: Longman.

Coupland, N. (1978). Is readability real? *Communication of scientific and technical information, 35*, 15-17.

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

Crismore, A. (1989). *Talking with readers: Metadiscourse as rhetorical act*. New York: Peter Lang.

Crismore, A., & Farnsworth, R. (1990). Metadiscourse in popular and professional science discourse. In W. Nash (Ed.), *The writing Scholar. Studies in Academic Discourse* (pp. 119-136). Newbury Park, CA: Sage.

Crismore, A., Markkannen, R., & Steffensen, M. (1993). Metadiscourse in persuasive writing: A study of texts written by American and Finnish university students. *Written Communication, 10*(1), 39-71.

Crossley, S., Greenfield, J. & McNamara, D.S. (2008). Assessing text readability using cognitively based indices. *TESOL Quarterly 42*(3), 475-493.

Crystal, D. (2001). *Language and the internet*. Cambridge: CUP.

Crystal, D., & Davy, D. (1969). *Investigating English style*. London: Longman.

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

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

Cummins, J. (1980). The cross-lingual dimensions of language proficiency: Implications for bilingual education and the optimal age issue. *TESOL Quarterly, 14*(2), 175-188.

Cummins, J. (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).

Dale, E. (1967). *Can you give the public what it wants?* New York: World Book Encyclopedia.

Davison, A. (1984). Readability formulas and comprehension. In G. G. Duffy, L. R. Rochler, & J. Mason (Eds.), *Comprehension instruction: Perspectives and suggestions*. New York: Longman.

del-Teso-Craviotto, M. (2006). Language and sexuality in Spanish and English dating chats. *Journal of Sociolinguistics, 10*(4), 460-480.

DeMarco, C. (1986). The role of register analysis in an English for Special (sic!) Purposes (ESP) curriculum. *TESOL TEIS Newsletter, 2*(2). Retrieved March 10, 2013, from http://www.tesol.org/read-and-publish/newsletters-other-publications/interest-section-newsletters/teis-newsletter/2011/10/31/the-role-of-register-analysis-in-an-english-for-special-purposes-(esp)-curriculum-(from-winter-1986-vol.-2-no.-2).

Dias, P. (1994). Initiating students into the genres of discipline-based reading and writing. In A. Freedman, & Medway (Eds.), *Learning and teaching genre* (pp. 193-206). Portsmouth, NH: Boynton Cook.

Dolch, E. W. (1939). Fact burden and reading difficulty. *Elementary English Review, 16,* 135-138.

Drobnic, K. (1978). Mistakes and modification in course design. In T. Trimble, M. Trimble, & L. Drobnic (Eds.), *English for specific purposes: Science and Technology*. Corvallis, Oregon: Oregon State University Press.

DuBay, W. H. 2004. *The principles of readability*. Cost Mesa, CA. Impact Information. Retrieved March 10, 2012 from http://files.eric.ed.gov/fulltext/ED490073.pdf.

Dudley- Evans, T., & Henderson, W. (Eds.). (1990). *The language of economics: the analysis of economics discourse*. London: Modern English Publications.

Duffy, T. M. (1985). Readability formulas: What's the use? In T. M. Duffy, & R. Waller (Eds.), *Designing Usable Texts* (pp. 113-143). London: Academic Press.

Duffy, T. M., & Kabance, P. (1981). Testing a readable writing approach to text revision. *Journal of Educational Psychology, 74*(5), 733-748.

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

Edwards, J. (2006). Foundations of bilingualism. In V. K. Bhatia, & W. C. Ritchie (Eds.), *The handbook of bilingualism* (pp. 7-31). Oxford: Blackwell.

Eggins, S., Martin, J. R. (1997). Genres and registers of discourse. In T. van Dijk (Ed.), *Discourse as structure and process* (pp. 230 – 256). London: Sage.

Ellis, J. N., & Ure, J. 1969. Language varieties: Registers, In A. R. Meetham (Ed.), *Encyclopaedia of Linguistics, Information and Control* (pp. 251-259), Oxford: Pergamon.

Ervin-Tripp, S. (1972). On sociolinguistic rules: alternation and co-occurrence. In J. Gumperz, & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp. 213-250). New York: Holt.

Ewer, J. R., & Latorre, G. A. (1969). *A course in basic scientific English*. London: Longman.

Ewer, J. R., & Huges-Davies, E. (1971). Further notes on developing an English Programme for students of science and technology. In J. M. Swales (Ed.), *Episodes in ESP*. Oxford: Pergamon.

Fahnestock, J. (1993). Genre and rhetorical craft. *Research in the Teaching of English*, *27*, 265-271.

Farr, J. N., Jenkins, J. J., & Paterson, D. G. (1951). Simplification of the Flesch reading ease formula. *Journal of Applied Psychology*, *35*(5), 333-357.

Feez, S. (2001). Heritage and innovation in second language education. In A. M. Johns (Ed.), *Genre in the classroom* (pp. 47-68). Mahwah, NJ: Lawrence Erlbaum.

Ferguson, C. (1983). Sports announcer talk: syntactic aspects of register variation. *Language in Society*, *12*, 153-172.

Flowerdew, J. (1993). An educational, or process, approach to the teaching of professional genres. *ELT Journal*, *27*, 11-15.

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

Flowerdew, J., & Dudley-Evans, T. (2002). Genre analysis of editorial letters to international journal contributors. *Applied Linguistics*, *23*(4), 463-489.

Flowerdew, L. (2002). Corpus-based analysis in EAP. In J. Flowerdew (Ed.), *Academic discourse* (pp. 95-114). Harlow: Longman.

Flowerdew, L. (2005). An integration of corpus-based and genre-based approaches to text analysis in EAP/ESP: countering criticisms against corpus-based methodologies. *English for Specific Purposes*, *24*, 321-332.

Ford, C. E., Fox, B. A., & Thompson, A. A. (Eds.). (2001). *The language of turn and sequence*. Oxford: Oxford University Press.

Fox., A., Butakto, D., Hallahan, M., & Crawford, M. (2007). The medium makes a difference: gender similarities and differences in instant messaging. *Journal of Language and Social Psychology, 26(*4), 389 – 397.

Francis, G., Hunston, S., & Manning, E. (1996). *Grammar patterns*. London: Harper Collins.

Fredericksen, C. H., & Donin, J. (1991). Constructing multiple semantic representations in comprehending and producing discourse. In G. Denhiere, & J. P. Rossi (Eds.), *Texts and text processing* (pp. 19-44). Amsterdam: North-Holland.

Freedman, A. (1993). Show and tell? The role of explicit teaching in the learning of new genres. *Research in the Teaching of English*, *27*, 222-251.

Freedman, A. (1994). "Do as I say": The relationship between teaching and learning new genres. In A. Freedman, & P. Medway (Eds.), *Genre and the new rhetoric* (pp. 191-210). London: Taylor & Francis.

Freedman, A., & Medway, P. (1994). Introduction: New views of genre and their implications for education. In A. Freedman, & P. Medway, (Eds.), *Learning and teaching genre* (pp. 1-22). Portsmouth, NH: Boynton Cook.

Fry, E. B. (1963). *Teaching faster reading*. London: Cambridge University Press.

Fry, E. B. (1988). Writeability: the principles of writing for increased comprehension. In B. L. Sakaluk, & S. J. Samuels (Eds.), *Readability: Its past, present, and future*. Newark, DE: International Reading Association.

Fry, E. B. (2002). Readability versus leveling. *The Reading Teacher, 56*, 286-291.

Gains, J. (1999). Electronic mail – a new style of communication or just a new medium? An investigation into the text features of email. *English for Specific Purposes, 18*(1), 81-101.

Geisler, C. (1995). *Relative infinitives in English*. Uppsala: University of Uppsala.

Ghadessy, M. (1988). The language of written sports commentary: soccer – a description. In Ghadessy, M. (Ed.), *Registers of written English: Situational factors and linguistic features* (pp. 17-51). London: Pinter.

Ghadessy, M. (Ed.). (1988). *Registers of written English: Situational factors and linguistic features*. London: Pinter.

Ghadessy, M (1993). On the nature of written business communication. In Ghadessy, M. (Ed.), *Register analysis: Theory and practice* (pp. 149-164). London and New York: Pinter Publishers.

Giddens, A. (1979). *Central problems in social theory*. Berkeley, CA: University of California Press.

Gilliland, J. (1972). *Readability*. London: Hodder and Stoughton.

Gledhill, C. (2000). The discourse function of collocation in research article introductions. *English for Specific Purposes*, *19*(2), 115-135.

Gosden, H. (1992). Discourse functions of marked theme in scientific research articles. *English for Specific Purposes*, *11*, 207-224.

Grabe, W. & Kaplan, R. B. (2006). Applied linguistics in North America. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (pp. 363 - 369). Amsterdam. Elsevier.

Granger, S. (1983). *The be + past participle construction in spoken English with special emphasis on the passive*. Amsterdam: Elsevier Science Publications.

Grellet, F. (1981). *Developing reading skills: A practical guide to reading comprehension exercises*. Cambridge: Cambridge University Press.

Gunawardena, C. N. (1989). The present perfect in the rhetorical divisions of biology and biochemistry journal articles. *English for Specific Purposes*, *8*, 265-273.

Gunning, R. (1952). *The technique of clear writing*. New York: McGraw-Hill.

de Haan, P. (1989). *Postmodifying clauses in the English noun phrase. A corpus-based study*. Amsterdam: Rodopi.

Haberlandt, K. F., & Graesser, A.C. (1985). Component processes in text comprehension and some of their interactions. *Journal of Experimental Psychology*, *114*, 357-374.

Halliday, M. A. K. (1978). *Language as a social semiotic: The social interpretation of language and meaning*. London: Arnold.

Halliday, M. A. K. (1985a). *An introduction to functional grammar*. London: Edward Arnold.

Halliday, M. A. K. (1985b). *Spoken and written language*. Oxford: Oxford University Press.

Halliday, M. A. K. (1985c). Systemic background. In J. D. Benson, & W. S. Greaves (Eds.), *Systemic Perspectives on Discourse, Volume 1. Selected Theoretical Papers from the 9th International Systemic Workshop* (pp. 1 15). Norwood, NJ: Ablex Publishing Corporation.

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

Halliday, M. A. K. (1989). *Spoken and written language.* Oxford: Oxford University Press.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M. A. K., & Hasan, R. (1985). *Language context and text: Aspects of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halliday, M. A. K., & Hasan, R. (1989). *Language, context and text: aspect of language in a social-semiotic perspective*. Oxford: Oxford University Press.

Halliday, M. A. K., & Hasan, R. (1990). *Language, context and text: Aspects of language in social-semiotic perspective. Oxford*: Oxford University Press.

Halliday, M. A. K., McIntosh, A., & Strevens, P. (1964). *The linguistic sciences and language teaching*. London: Longman.

Hamp-Lyons, L., & Thompson, P. (2006). Introduction to special issue. Academic English in secondary schools. *Journal of English for Academic Purposes*, *5*, 251-253.

Hargis, G., Hernandez, A. K., Hughes, P., Ramaker, J., Rouiller, S., & Wilde, E. (1998). *Developing quality technical information: A handbook for writers and editors*. Upper Saddle River, NJ: Prentice Hall.

Harris, R. A. (1991). Rhetoric of science. *College English 53*(3), 282-307.

Harrison, S., & Bakker, P. (1998). Two new readability predicators for the professional writer. *Journal of Research in Reading*, *21*(2), 121-138.

Hart. J. R. (2007). *A writer's coach: The complete guide to writing strategies that work*. New York: Anchor Books.

Heath, S. B., & Langman, J. (1994). Shared thinking and the register of coaching. In D. Biber, & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 82-105). New York: Oxford University Press.

Henrichs, L. F. (2010). *Academic language in early childhood interactions: A longitudinal study of 3- to 6-year-old Dutch monolingual children*. Amsterdam Center for Language and Communication: Amsterdam.

Henry, A., & Roseberry, R. (2001). A narrow-angled corpus analysis of moves and strategies of the genre: letter of application. *English for Specific Purposes, 20*, 153-167.

Herbert, A. J. (1965). *The structure of technical English*. London: Longman.

Herring, S. C. (1996). *Computer-mediated communication: Linguistics, social and cross-cultural perspectives*. Amsterdam: John Benjamins.

Herring, S. C., & Paolillo, C. (2006). Gender and genre variation in weblogs. *Journal of Sociolinguistics, 10*(4), 493 – 459.

Hogue, A. (2008). *First steps in academic writing.* New York: Longman.

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

Holmes, J. (1986). The Brazilian National ESP Project. *Communication skills training in bilateral aid projects,* 190-191. London: British Council.

Hopkins, A., & Dudley-Evans, T. (1988). A genre-based investigation of the discussion sections in articles and dissertations. *English for Specific Purposes, 7*(2), 113-121.

Horner, B. (1997). Students, authorship, and the work of composition. *College English*, *59*(5), 505-529.

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

Huddleston, R. (1984). *Introduction to the grammar of English*. Cambridge: Cambridge University Press.

Hundt, M. (2004). The passival and the progressive passive: a case study of layering in the English aspect and voice system. In H. Lindquist, & C. Mari (Eds.), *Corpus approaches to grammaticaliziation in English* (pp. 79-120). Amsterdam: Benjamins.

Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.

Hunston, S. (2006) Corpus Linguistics. In K. Brown (Ed.), *Encyclopedia of Language and Linguistics* (pp. 234 - 250). Amsterdam: Elsevier.

Hunston S., & Thompson, G. (2001). *Evaluation in text*. Oxford: Oxford University Press.

Hutchinson, T., Waters, A., & Breen, M. P. (1979). An English language curriculum for technical students. *Practical Papers in English Language Education*, *2,* 146-171.

Hutchinson, T., & Waters, A. (1987). *English for specific purposes: A learning-centred approach*. Cambridge: Cambridge University Press.

Hyland, K. (1998a). *Hedging in scientific research articles*. Amsterdam: John Benjamins.

Hyland, K. (1998b). Persuasion and context: The pragmatics of academic metadiscourse. *Journal of Pragmatics, 30,* 437-455.

Hyland, K. (1998c). Exploring corporate rhetoric: Metadiscourse in the CEO's letter. *Journal of Business Communication, 35*(2), 224-245.

Hyland, K. (1999). Talking to students: metadiscourse in introductory course books. *English for Specific Purposes, 18*(1), 3-26.

Hyland, K. (2000). *Disciplinary discourses. Social interactions in academic writing.* London: Longman.

Hyland, K. (2001). Humble servant of the discipline? Self-reference in research articles. *English for Specific Purposes, 20*, 207-226.

Hyland, K. (2002a). *Teaching and researching writing.* Harlow: Pearson Education.

Hyland, K. (2002b). Genre: Language, context, and literacy. *Annual Review of Applied Linguistics, 22*, 113-135.

Hyland, K. (2005). *Metadiscourse*. London: Continuum.

Hyland, K. (2010). Metadiscourse: Mapping interactions in academic writing. *Nordic Journal of English Studies, 9*(2), 125-143.

Hyland, K., & Tse, P. (2004). Metadiscourse in academic writing: a reappraisal. *Applied Linguistics, 25*(2), 156-177.

Hyland, K., & Tse, P. (2005). Hooking the reader: a corpus study of evaluative *that* in abstracts. *English for Specific Purposes, 24*, 123-139.

Hymes, D. (1974). *Foundations in sociolinguistics: An ethnographic approach*. Philadelphia: University of Pennsylvania Press.

Hymes, D. (1984). Sociolinguistics: stability and consolidation. *International Journal of the Sociology of Language, 45*, 39-45.

Hyon, S. (1996). Genre in three traditions: Implications for ESL. *TESOL Quarterly, 30*(4), 693-722.

Intraprawat, P., & Steffensen, M. (1995). The use of metadiscourse in good and poor ESL essays. *Journal of Second Language Writing, 4*(3), 253-272.

Jacobson, J. M. (1998). *Content area reading: Integration with the language arts*. New York: Delmar.

Jalilifar A., & Alipour, M. (2007). How explicit instruction makes a difference: Metadiscourse markers and EFL learners' reading comprehension skill. *Journal of College Reading and Learning, 38*(1), 35-52.

Janda, R. (1985). Note-taking English as a simplified register. *Discourse Processes, 8*, 437-454.

Janni, G. (2000). Mi szakmai elitet képzünk. *Educatio, 4,* 799-810.

Johansson, C. (1995). *The relativizers 'whose' and 'of which' in present-day English. Description and theory.* Uppsala: University of Uppsala.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *University of Lund Working Papers in Linguistics, 53*, 61-79.

Johns, T. (1991). Should you be persuaded: Two samples of data-driven learning materials. In T. Johns, & P. King (Eds.), *Classroom concordancing: English Language Research Journal 4 (1-16).* Birmingham: University of Birmingham, Centre for English Language Studies.

Johns, A. (1995). Genre and pedagogical purposes. *Journal of Second Language Writing, 4*(2), 181-190.

Jordan, R. R. (2002). *Academic writing course*. Harlow: Longman.

Just, M.A., & Carpenter, P.A. (1980). A theory of reading: from eye fixations to comprehension. *Psychological Review 87*(4), 329-354.

Kamberelis, G. (1995). Genre as institutionally informed social practice. *Journal of Contemporary Legal Issues, 6*, 115-170.

Károly, K. (2007). *Szövegtan és fordítás*. [Text analysis and translation]. Budapest: Akadémiai Kiadó.

Kemper, S. (1983). Measuring the interference load of a text. *Journal of Educational Psychology, 75*(3), 391-401.

Kennedy, C. (2012). ESP projects, English as a global language, and the challenge of change. *Ibérica*, *24*, 43-54.

Kern, R. P. (1979). *Usefulness of readability formulas for achieving army readability objective*. Fort Benjamin Harrison, ID: Technical Advisory Service, U.S. Army Research Institute.

Khairi, I. A. (2001). English for specific purposes in Malaysia: international influence. *Journal of Southeast Asian Education, 2*(2), 345-361.

Khine, M. S. (2013). *Critical analysis of science textbooks: Evaluating instructional effectiveness*. Dordrecht: Springer.

Kincaid, J. P., & Delionbah, L. J. (1973). Validation of the automated readability index: A follow-up. *Human Factors, 15*, 17-20.

Kintsch, W., & Miller, J. R. (1981). Readability: A view from cognitive psychology. In J. Flood (Ed.), *Understanding reading comprehension: cognition, language, and the structure of prose* (pp. 220-232). Neward, DE: International Reading Association.

Klare, G. R. (1963). *The measurement of readability*. Ames, IA: Iowa State University Press.

Klare, G. R. (1968). The role of word frequency in readability. *Elementary English*, 45, 12-22.

Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly, 10*, 62-102.

Klare, G. R. (1976). A second look at the validity of the readability formulas. *Journal of Reading Behavior, 8*, 159-162.

Klare, G. R. (1982). Readability. In *Encyclopedia of educational research* (Vol. 3, pp. 1520-1531). New York: The Free Press.

Klare, G. R. (1984). Readability. In P.D. Pearson (Ed.), *Handbook of reading research* (pp. 681-744). New York: Longman.

Klare, G. R., & Buck, B. (1954). *Know your reader, the scientific approach to readability*. New York: Hermitage House.

Kocsány, P. (2002). *Szöveg, szövegtípus, jelentés: A mondás mint szövegtípus*. [Text, text type and meaning: Saying as a text type]. Budapest: Akadémiai Kiadó.

Koda, K. (2005). *Insights into second language reading*. New York: Cambridge University Press.

Koester, A. J. (2006). *Investigating workplace discourse*. London: Routledge.

Kong, K. C. C. (2006). Property transaction report: news, advertisement or a new genre? *Discourse Studies, 8(6)*, 771-796.

Kormos, J., & Csölle, A. (2004). *Topics in applied linguistics*. Budapest: ELTE Eötvös Kiadó.

Krashen, S. (1985). *The input hypothesis: issues and implications*. London: Longman.

Kress, G., & Hodge, R. (1979). *Language as ideology*. Boston, MA: Routledge and Kegan Paul.

Leckie-Tarry, H. (1993). The specification of a text: Register, genre and language teaching. In M. Ghadessy (Ed.), *Register Analysis*: *Theory and Practice* (pp. 26-42). London: Pinter Publishers.

Lackstrom, J. E., Selinker, L., & Trimble, L. P. (1970). Grammar and technical English. In. R. C. Lugten (Ed.), *English as a second language. Current issues* (pp. 101-133). Philadelphia: CCD Chilton.

Lackstrom, L., Selinker, L., & Trimble, L. (1973). Technical rhetorical principles and grammatical choice. *TESOL Quarterly, 7*(2).

Lambert, W. E., & Tucker, G. R. (1972). *Bilingual education of children: the St. Lambert experiment*. Rowley, MA: Newbury House.

Language Policy Unit of the Council of Europe. (1996). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Strasbourg: Retrieved October 15, 2010 from www.coe.int/lang-CEFR.

Laurén, U. (2002). Some lexical features of immersion pupils' oral and written narration. *University of Lund Working Papers in Linguistics, 50*, 63-78.

LDOCE (1995). Longman dictionary of contemporary English. London: Longman.

Le, T., Yue, Y., & Le, Q. (2011). *Linguistic complexity and its relation to language and literacy education.* New York: Nova Science Publishers.

Lee, D. (2001). Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language, learning, and technology, 5*(3), 37-72.

Lee, J. J., & Subtirelu, N. C. (2015). Metadiscourse in the classroom: A comparative analysis of EAP lessons and university lectures. *English for Specific Purposes, 37*, 52-62.

Leech, G., Rayson, R., & Wilson, A. (2001). *Word frequencies in written and spoken English*. London: Longman.

Lewis, M. (1998). *Implementing the lexical approach*. Hove: Language Teaching Publishers.

Linnarud, M., & Thoursie, S. (2008). English and German in Swedish classrooms: Writing in the two languages compared. *Nordic Journal of English Studies, 7*(2), 75-98.

Louhiala-Salminen, L. (1999). *From business correspondence to message exchange: The notion of genre in business communication*. Jyvaskyla: University of Jyvaskyla.

Love, A. M. (1991). Process and product in geology: An investigation of some discourse features of two introductory textbooks. *English for Specific Purposes, 10*, 89-109.

Love, A. M. (2002). Introductory concepts and "cutting edge" theories: can the genre of the textbook accommodate both? In J. Flowerdew (Ed.), *Academic discourse* (pp. 76-91). Harlow: Longman.

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

MacDonald, S. P. (2005). The language of journalism in treatments of hormone replacement news. *Written Communication, 22*(3), 275-297.

Mackay, R. (1978). Identifying the nature of learner's needs. In R. Mackay, & A. J. Mountford (Eds.), *English for specific purposes*: *A case-study approach*. London: Longman.

Macken-Horarik, M. (2002). 'Something to shoot for': A systemic functional approach to teaching genre in secondary school science. In A. M. Johns (Ed.), *Genre in the classroom* (pp. 21-46). Mahwah, NJ: Lawrence Erlbaum.

Mair, C. (1990). *Infinitival complement clauses in English*. Cambridge: Cambridge University Press.

Mann, M., & Taylore-Knowles, S. (2007). *Destinations: C1 & C2*. Oxford: Macmillan.

Manzo, A. (1970). Readability: a postscript. *Elementary English*, *47*, 962-965.

Marco, M. J. (2000). Collocation frameworks in medical research papers: A genre-based study. *English for Specific Purposes*, *19*(1), 63-86.

Marshall, H. (1987). Quantity surveying reports. *ELR Journal*, *1*, 117-155.

Martin, J. R. (1985). *Factual writing: Exploring and challenging social reality*. Geelong: Deakin University Press.

Martin, J. R. (1997). Analysing genre: functional parameters. In F. Christie, & J. R. Martin (Eds.), *Genre and institutions: social processes in the workplace and school* (3-39). London: Continuum. 1997.

Martin, J. R. (2000). Design and practice: Enacting functional linguistics. *Annual Review of Applied Linguistics*, *20*, 116-126.

Martin, J. R. (2001a). Language, register and genre. In A. Burns, & C. Coffin (Eds.), *Analysing English in a global context* (pp. 149-166). London: Routledge.

Martin, J. R. (2001b). Beyond exchange: appraisal system in English . In S. Hunston, & G. Thompson (Eds.), *Evaluation in Text* (pp. 142-175). Oxford: Oxford University Press.

Martin, J. R., Christie, F., & Rotherty, J. (1987). Social processes in education: A reply to Sawyer and Watson. In I. Reid (Ed.), *The place of genre in learning: Current debates* (pp. 46-57). Geelong, Australia: Deakin University Press.

Matthiessen, C. M. I. M. (1993). Register in the round: diversity in a unified theory of register analysis. In M. Ghadessy (Ed.), *Register analysis: theory and practice* (pp. 221-292). London: Pinter Publishers.

Mauranen, A. (1993a). Contrastive ESP rhetoric: Metatext in Finnish-English economics texts. *English for Specific Purposes*, *12*, 3-22.

Mauranen, A. (1993b*). Cultural differences in academic rhetoric*. Frankfort: Peter Lang.

Maxwell, M. (1978). Readability: Have we gone too far? *Journal of reading*, *21*, 525-530.

McCarthy, M. (2000). Mutually captive audiences: small talk and the genre of close-contact service encounters. In J. Coupland (Ed.), *Small talk* (pp. 84-109). Harlow: Longman.

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

McLaughlin, G. H. (1969). SMOG grading: a new readability formula. *Journal of Reading*, *22*, 639-646.

Medgyes, P. (2011). *Aranykor. Nyelvoktatásunk két évtizede. 1989-2009*; Budapest: Nemzeti Tankönyvkiadó.

Meyer, C. (1992). *Apposition in Contemporary English*. Cambridge: Cambridge University Press.

Meyer, C. (2002). *English corpus linguistics: an introduction*. Cambridge: Cambridge University Press.

Miller, C. (1984). Genre as social action. *Quarterly Journal of Speech, 70*, 151-167.

Ministry of Education and Culture, Department for EU Relations. (2008). *Education in Hungary: Past, Present, Future – An overview.* Budapest: Pátria Nyomda. Retrieved 8 March, 2012 from www.okm.gov.hu/letolt/english/education_in_hungary_080805.pdf.

Munby, J. (1978). *Communicative syllabus design: A sociolinguistic model for defining the content of purpose-specific language programmes*. Cambridge: Cambridge University Press.

Művelődési és Közoktatási Minisztérium [Hungarian Ministry of Culture and National Education]. (1997). *Két tanítási nyelvű iskolai oktatás irányelvei*. (26/1997. VII. 10. MKM rendelet) [Educational principles of bilingual schools, Regulation No 26/1997]. Budapest: CompLex Jogtár.

Myers, G. (1989). The pragmatics of politeness in scientific texts. *Applied Linguistics*, *4*, 1-35.

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

Narelle, F. S., Leigh, N. W., Roslyn, K. G., & Gillian, P. (1994). *Analysis of student performance in statistics*. Sydney: University of Technology.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: working with the British component of the International Corpus of English*. Amsterdam: Benjamins.

Nesi, H. (2005). A corpus-based analysis of academic lectures across disciplines In J. Cotterill, & A. Ife (Eds.), *Language across boundaries* (pp. 201-218). London: BAAL / Continuum Press.

Nuttall, C. E. (1982). *Teaching reading skills in a foreign language*. London: Heinemann Educational Books.

Nwogu, K. N. (1991). Structure of science popularizations: A genre analysis approach to the schema of popularized medical texts. *English for Specific Purposes*, *10*, 111-123.

OALD (1995). *Oxford advanced learner's dictionary*. Oxford: Oxford University Press.

O'Loughlin, K. (1995). Lexical density in candidate output on two versions of an oral proficiency test. *Language Testing, 12*(2), 217-237.

O'Keffee, A., & McCarthy, M. (2010). *The Routledge handbook of corpus linguistics*. London: Routledge.

Oster, S. (1981). The use of tenses in reporting past literature in EST. In L. Selinker, E. Tarone, & V. Haneli (Eds.), *English for academic and technical purposes. Studies in honor of Luis Trimble* (pp. 76-90). Rowley, MA: Newbury House.

Ota, A. (1963). *Tense and aspect of present-day American English*. Tokyo: Kenkyusha.

Patridge, B. (2002). Thesis and dissertation writing: An examination of published advice and actual practice. *English for Specific Purposes, 21*,125-143.

Perfetti, C.A., Landi, N., & Oakhill , J. (2005), The acquisition of reading comprehension skill. In M. J. Snowling. & C. Hulme (Eds.), *The Science of Reading* (pp. 227-247). Oxford: Blackwell.

Pickett, D. (1986). Business English: Falling between two styles. *COMLON*, *26*, 16-21.

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

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1989). *The comprehensive grammar of the English language*. Harlow: Longman.

Raimes, A. (1991). Out of the woods: Emerging traditions in the teaching of writing. *TESOL Quarterly*, *25*, 407-430.

Ramanathan, V., & Kaplan, R. B. (2000). Genres, authors, discourse communities: Theory and application for (L1 and) L2 writing instructors. *Journal of Second Language Writing, 9*(2), 171-191.

Rayner, K. & Pollatsek, A. (1989). *The Psychology of Reading*. Englewood Cliffs, UK: Prentice Hall.

Read, J., & Nation, P. (2006). *An investigation of the lexical dimension of the IELTS speaking test. IELTS Research Reports 6.* Sydney: IELTS Australia and British Council.

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

Redish, J. C., & Selzer, J. (1985). The place of readability formulas in technical communication.*Technical Communication*, *32*(4), 46-52.

Reid, I. (1987). A generic frame for debates about genre. In L. Reid (Ed.), *The place of genre in learning: Current debates* (pp. 1-8). Geelong, Australia: Deakin University, Centre for Studies in Literacy Education.

Reid, T. B. W. (1956). Linguistics, structuralism, philology. *Archivum Linguisticum, 8*, 28-37.

Richards, J. (1974). Word list: Problems and prospects. *RELC Journal*, *5*(2), 69-84.

Richterich, R. (1984). A European credit system for modern language learning by adults. In J. A. van Ek, & J. L. M. Trim (Eds.), *Across the threshold level*. Oxford: Pergamon.

Richterich, R., & Chancerel, J. L. (1980). *Identifying the needs of adults learning a foreign language*. Oxford: Pergamon.

Rittman, R. J. (2007). *Automatic discrimination of genres: The role of adjectives and adverbs as suggested by linguistics and psychology*. Unpublished doctoral dissertation, New Brunswick Rutgers University, New Jersey.

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

Rodgers, C. (1969). *Freedom to learn*. Columbus, OH: Merrill.

Ruddell, M. (2005). *Teaching content reading and writing Hoboken.* Hoboken, NJ: John Wiley.

Ruiz-Garrido, M. F., Palmer-Silveira, J. C., & Fortanet-Gómez, I. (2010). *English for professional and academic purposes.* Amsterdam: Rodopi.

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

Salager-Meyer, F. (1990). Discoursal flaws in medical English abstracts: A genre analysis per research- and text-type. *Text, 10*, 365-384.

Samraj, B. (2002). Introduction in research articles: variation across disciplines. *English for Specific Purposes*, *21*, 1-17.

Sanders, T., & Sanders, J. (2006). Text and text analysis. In K. Brown (Ed.), *Encyclopedia of language & linguistics* (pp. 597 - 607). Amsterdam: Elsevier.

Schiffrin, D. (1980). Meta-talk: Organisational and evaluative brackets in discourse. *Sociological Inquiry: Language and social interaction, 50*(3-4), 199-236.

Schiffrin, D., Tannen, D., & Hamilton, H. F. (2001). *The handbook of discourse analysis*. Oxford: Blackwell.

Schriver, K. (2000). Readability formulas in the new millennium: What's the use? *ACM journal of computer documentation, 24*(3), 138-140.

Scott, M. (1999). *WordSmith tools*. Oxford: Oxford University Press.

Scott, M., & Tribble, C. (2006). *Textual pattern. Key words and corpus analysis in language education*. Amsterdam: John Benjamins Pub.

Selinker, L., Todd-Trimble, M., & Trimble, L. (1976). Presuppositional rhetorical information in EST discourse. *TESOL Quarterly, 10*(3), 281-290.

Selinker, L., & Trimble, L. (1976). Scientific and technical writing: The choice of tense. *English Teaching Forum, 14*(4).

Selzer, J. (1981). Readability is a four-letter word. *Journal of Business Communication,18*(4), 23-34.

Selzer, J. (1983). What constitutes a 'readable' technical style? In P. V. Anderson, R. J. Brockmann, & C. R. Miller (Eds.), *New essays in technical and scientific communication: research, theory, practice* (pp. 71-89). Famingdale, NY: Baywood.

Semino, E., & Short, M. (2004). *Corpus stylistics: speech, writing and thought presentation in a corpus of English writing*. London: Routledge.

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

Sherman, L. A. (1893). *Analytics of literature: A manual for the objective study of English prose and poetry*. Boston: Ginn & Co.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (2004). *Trust the text: language, corpus and discourse*. London: Routledge.

Smith, E. A., & Kincaid, J. P. (1970). Derivation and validation of the automated readability index for use with technical materials. *Human Factors, 12,* 457-464.

Smith, E. A., & Senter, R. J. (1967). *Automated Readability Index*. Wright Patterson AFB, Ohio: Aerospace Medical Division.

Stegen, O. (2005). Editing Rangi narratives: A pilot study in literature production. *Edinburgh Working Papers in Applied Linguistics, 14*, 68-98.

Stegen, O. (2007). Lexical density in oral versus written Rangi texts. *SOAS Working Papers in Linguistics, 15*, 173-184.

Stotesbury, H. (2003). Evaluation in research article abstracts in the narrative and hard sciences. *Journal of English for Academic Purposes, 2*, 327-241.

Stubbs, M. (2004). Language Corpora. In A. Davies, & C. Elder (Eds.), *The handbook of applied linguistics* (pp. 106 - 132). Malden, MA: Blackwell publishing.

Swain, M., & Johnson, R. K. (1997). Immersion education: a category within bilingual education. In R. K. Johnson, & M. Swain (Eds.), *Immersion education: international perspectives* (pp. 11-16). Cambridge: Cambridge University Press.

Swales, J. M. (1971). *Writing scientific English: A textbook of English as a foreign language for students of physical and engineering sciences*. London: Nelson.

Swales, J. M. (1981). *Aspects of article introductions*. Birmingham: University of Aston.

Swales, J. M. (1986). A genre-based approach to language across the curriculum. In M. L. Tickoo (Ed.), *Language across the curriculum* (pp. 10-22). Singapore: Regional English Language Centre.

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

Swales, J. M. (2004). *Research genres: explorations and applications*. Cambridge: Cambridge University Press.

Swales, J. M., & Feak, C. B. (2012). *Academic writing for graduate students: Essential tasks and skills.* Ann Arbor: University of Michigan.

Taavitsainen, I. (1999). Metadiscursive practices and the evaluation of early English medical writing (1375-1550). In J. M. Kirk (Ed.), *Corpora Galore: Analyses and techniques in describing English* (pp. 191-207). Amsterdam: Rodopi.

Taavitsainen, I., & Pahta, P. (2004). *Medical and scientific writing in late Medieval English*. Cambridge: Cambridge University Press.

del-Teso-Craviotto, M. (2006). Language and sexuality in Spanish and English dating chats. *Journal of Sociolinguistics, 10*(4), 460-480.

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

Thompson, S. (1994). Frameworks and contexts: A genre-based approach to analysing lecture introductions. *English for Specific Purposes, 13,* 171-186.

Thompson, G. (2001). Interaction in academic writing: Learning to argue with the reader. *Applied Linguistics, 22*(1), 58-78.

Thompson, P. (2000). Citation practices in PhD theses. In L. Burnard, & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 91-101). Frankfurt am main: Peter Lang Publishers.

Threadgold, T. (1988). The genre debate. *Southern Review, 21,* 315-330.

Thurlow, C. (2003). Generation Txt? The sociolinguistics of young people's text-messaging. *Discourse Analysis Online, 1*. Retrieved March 10, 2013, from http://extra.shu.ac.uk/daol/articles/v1/n1/a3/thurlow2002003.html.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam: Benjamins.

Tognini-Bonelli, E. & Camiciotti, G. L. (2005). *Strategies in academic discourse*. Amsterdam and Philadelphia: John Benjamins.

Tottie, G. (1991). *Negation in English speech and writing. A study in variation*. San Diego: Academic Press.

Trask, R. L. (1999). *Key concepts in language and linguistics*. London: Routledge.

Tribble, C. (1999). *Writing difficult texts*. Unpublished doctoral dissertation, Lancaster University.

Tribble, C. (2002). Corpora and corpus analysis: New windows on academic writing. In J. Flowerdew (Ed.), *Academic discourse* (pp.131-149). London: Longman, Person Education.

Trimble, L. (1985). *EST. A discourse approach*. Cambridge University Press.

Ulusoy, M. (2006). Readability approaches: Implications for Turkey. *International Education Journal, 7*(3), 323-332.

Upton, T. (2002). Understanding direct mail letters as a genre. *International Journal of Corpus Linguistics, 7*(1), 65-85.

Upton, T., & Connor, U. (2001). Using computerised corpus analysis to investigate the textlinguistic discourse moves of a genre. *English for Specific Purposes, 20*(4), 313-329.

Ure, J. (1968). Practical registers. *English Language Teaching, 23*, 107-215.

Ure, J. (1971). Lexical density and register differentiation. In J.E. Perren, & J.L.M. Trim (Eds.), *Applications of linguistics* (pp. 443-452). Cambridge: Cambridge University Press.

Ure, J. (1982). Introduction: approaches to the study of register range. *International Journal of the Sociology of Language, 35*, 5-23.

Ure, J., & Ellis, J. (1977). Register in descriptive linguistics and linguistic sociology. In O. Uribe-Villegas (Ed.), *Issues in sociolinguistics (*pp. 197-243*)*. The Hague: Mouton.

Valdes, G., Barrera, R., & Cardenas, M. (1984). Constructing matching texts in two languages: The application of propositional analysis. *The Journal for the National Association of Bilingual Education, 9*, 3-19.

Valero-Garces, C. (1996). Contrastive ESP rhetoric: Metatext in Spanish-English economics texts. *English for Specific Purposes, 15*(4), 279-294.

Vande Kopple, W. J. (1985). Some exploratory discourse on metadiscourse. *College Composition and Communication, 36,* 82-93.

Vande Kopple, W. J. (1998). Relative clauses in spectroscopic articles in the Physical Review, beginnings and 1980. *Written Communication, 15*(2). 170-202.

Verantola, K. (1984). *On noun phrase structures in engineering English*. Turku: University of Turku.

Vilha, M. (1999). *Medical writing: Modality in focus*. Amsterdam: Rodopi.

Vince, M., & Emmerson, P. (2003). *First certificate language practice: English grammar and vocabulary*. Oxford: Macmillan.

Vinh, T., Si, F., & Damon, T. (2013). Lexical Density and readability: A case study of English textbooks. *Language, Society and Culture, 37*, 61-71.

Walker, R. H. (1967). Teaching the present perfect tenses. *TESOL Quarterly, 1*, 17-30.

Webber, P. (1994). The function of questions in different medical journal genres. *English for Specific Purposes, 13*(3), 257-268.

Weissberg, B. (1993). The graduate seminar: Another research-process genre. *English for Specific Purposes, 12*, 23-35.

West, M. (1953). *A general service list of English words*. London: Longman.

White, S. (2011) Understanding Adult Functional Literacy: Connecting Text Features, Task Demands, and Respondent Skills. New York: Routledge.

Widdowson, H. G. (1978). *Teaching language as communication*. Oxford: Oxford University Press.

Widdowson H. G. (1981). *Learning purpose and language use*. Oxford: Oxford University Press.

Widdowson, H. G. (1996). *Linguistics*. Oxford: Oxford University Press.

Widdowson, H. G. (1998). Context, community and authentic language. *TESOL Quarterly, 32*(4), 705-716.

Widdowson, H. G. (July, 2002). *Corpora and language teaching tomorrow*. Keynote lecture presented at 5th Teaching and Language Corpora Conference, Bertinoro, Italy.

Wilkinson, A. M. (1992). Jargon and the Passive Voice: Prescriptions and Proscriptions for Scientific Writing. *Journal of Technical Writing and Communication*, 22, 319-325.

Williams, J. M. (1995). *Style: Toward clarity and grace*. Chicago: University of Chicago Press.

Williams, J. M., & Colomb, G. G. (1993). The case for explicit teaching: Why what you don't know won't help you. *Research in the Teaching of English, 27*, 252-264.

Williams, R. (1985). Teaching vocabulary recognition strategies in ESP reading. *English for Specific Purposes, 4*(2), 121-131.

Willis, D. (2003). *Rules, patterns and words: grammar and lexis in English language teaching*. Cambridge: Cambridge University Press.

Woods, B., Moscardo, G., & Greenwood, T. (1998). A critical review of readability and comprehensibility tests. *The Journal of Tourism Studies, 9*(2), 49-61.

Worthington, D., & Nation, I. S. P. (1996). Using texts to sequence the introduction of new vocabulary in an EAP course. *RELC Journal, 27*(2), 1-11.

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

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

Yu, G. (2007). Lexical diversity in MELAB writing and speaking task performances. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 5*, 79-116.

Zamel, V. (1984). The author responds. *TESOL Quarterly, 18*, 154-157.

Websites:
www.fulbright.org.uk.
http://vocabulary.englishprofile.org.

Appendices

Appendix A

List of biology terms and their frequency in Bands 4-10 in the BIOCOR

(raw frequency and frequency expressed in percentages)

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

Appendix B

Dimensions of the multidimensional analysis (Biber, 2001)

1. involved – informational
2. narrative – nonnarrative
3. elaborated reference – situation dependent reference
4. overt expression of argumentation
5. abstract style nonabstract style
6. Online informational elaboration marking stance
7. Academic hedging

Appendix C

UCREL CLAWS7 Tag Set

Number of the code in alphabetical order	Part-of-speech code	Meaning of the part-of-speech code (with examples)
1	APPGE	possessive pronoun, pre-nominal (e.g., *my, your, our*)
2	AT	article (e.g., *the, no*)
3	AT1	singular article (e.g., *a, an, every*)
4	BCL	before-clause marker (e.g., *in order (that), in order (to)*)
5	CC	coordinating conjunction (e.g., *and, or*)
6	CCB	adversative coordinating conjunction (*but*)
7	CS	subordinating conjunction (e.g., *if, because, unless, so, for*)
8	CSA	*as* (as conjunction)
9	CSN	*than* (as conjunction)
10	CST	*that* (as conjunction)
11	CSW	*whether* (as conjunction)
12	DA	after-determiner or post-determiner capable of pronominal function (e.g., *such, former, same*)
13	DA1	singular after-determiner (e.g., *little, much*)
14	DA2	plural after-determiner (e.g., *few, several, many*)
15	DAR	comparative after-determiner (e.g., *more, less, fewer*)
16	DAT	superlative after-determiner (e.g., *most, least, fewest*)

17	DB	before determiner or pre-determiner capable of pronominal function (*all, half*)
18	DB2	plural before-determiner (*both*)
19	DD	determiner (capable of pronominal function) (e.g. *any, some*)
20	DD1	singular determiner (e.g., *this, that, another*)
21	DD2	plural determiner (*these, those*)
22	DDQ	wh-determiner (*which, what*)
23	DDQGE	wh-determiner, genitive (*whose*)
24	DDQV	wh-ever determiner (*whichever, whatever*)
25	EX	existential *there*
26	FO	formula
27	FU	unclassified word
28	FW	foreign word
29	GE	Germanic genitive marker - (*'* or *'s*)
30	IF	*for* (as preposition)
31	II	general preposition
32	IO	*of* (as preposition)
33	IW	*with, without* (as prepositions)
34	JJ	general adjective
35	JJR	general comparative adjective (e.g., *older, better, stronger*)
36	JJT	general superlative adjective (e.g., *oldest, best, strongest*)
37	JK	catenative adjective (*able* in *be able to, willing* in *be willing to*)

38	MC	cardinal number, neutral for number (*two, three..)*
39	MC1	singular cardinal number (*one*)
40	MC2	plural cardinal number (e.g., *sixes, sevens*)
41	MCGE	genitive cardinal number, neutral for number (*two's, 100's*)
42	MCMC	hyphenated number (*40-50, 1770-1827*)
43	MD	ordinal number (e.g., *first, second, next, last*)
44	MF	fraction, neutral for number (e.g., *quarters, two-thirds*)
45	ND1	singular noun of direction (e.g., *north, southeast*)
46	NN	common noun, neutral for number (e.g., *sheep, cod, headquarters*)
47	NN1	singular common noun (e.g., *book, girl*)
48	NN2	plural common noun (e.g., *books, girls*)
49	NNA	following noun of title (e.g., *M.A.*)
50	NNB	preceding noun of title (e.g., *Mr., Prof.*)
51	NNL1	singular locative noun (e.g., *Island, Street*)
52	NNL2	plural locative noun (e.g., *Islands, Streets*)
53	NNO	numeral noun, neutral for number (e.g., *dozen, hundred*)
54	NNO2	numeral noun, plural (e.g., *hundreds, thousands*)
55	NNT1	temporal noun, singular (e.g., *day, week, year*)
56	NNT2	temporal noun, plural (e.g., *days, weeks, years*)
57	NNU	unit of measurement, neutral for number (e.g., *in, cc*)
58	NNU1	singular unit of measurement (e.g., *inch, centimetre*)

59	NNU2	plural unit of measurement (e.g., *ins., feet*)
60	NP	proper noun, neutral for number (e.g., *IBM, Andes*)
61	NP1	singular proper noun (e.g., *London, Jane, Frederick*)
62	NP2	plural proper noun (e.g., *Browns, Reagans, Koreas*)
63	NPD1	singular weekday noun (e.g., *Sunday*)
64	NPD2	plural weekday noun (e.g., *Sundays*)
65	NPM1	singular month noun (e.g., *October*)
66	NPM2	plural month noun (e.g., *Octobers*)
67	PN	indefinite pronoun, neutral for number (*none*)
68	PN1	indefinite pronoun, singular (e.g., *anyone, everything, nobody, one*)
69	PNQO	objective wh-pronoun (*whom*)
70	PNQS	subjective wh-pronoun (*who*)
71	PNQV	wh-ever pronoun (*whoever*)
72	PNX1	reflexive indefinite pronoun (*oneself*)
73	PPGE	nominal possessive personal pronoun (e.g., *mine, yours*)
74	PPH1	3rd person sing. neuter personal pronoun (*it*)
75	PPHO1	3rd person sing. objective personal pronoun (*him, her*)
76	PPHO2	3rd person plural objective personal pronoun (*them*)
77	PPHS1	3rd person sing. subjective personal pronoun (*he, she*)
78	PPHS2	3rd person plural subjective personal pronoun (*they*)
79	PPIO1	1st person sing. objective personal pronoun (*me*)

80	PPIO2	1st person plural objective personal pronoun (*us*)
81	PPIS1	1st person sing. subjective personal pronoun (*I*)
82	PPIS2	1st person plural subjective personal pronoun (*we*)
83	PPX1	singular reflexive personal pronoun (e.g., *yourself, itself*)
84	PPX2	plural reflexive personal pronoun (e.g., *yourselves, themselves*)
85	PPY	2nd person personal pronoun (*you*)
86	RA	adverb, after nominal head (e.g., *else, galore*)
87	REX	adverb introducing appositional constructions (*namely, e.g.*)
88	RG	degree adverb (*very, so, too*)
89	RGQ	wh- degree adverb (*how*)
90	RGQV	wh-ever degree adverb (*however*)
91	RGR	comparative degree adverb (*more, less*)
92	RGT	superlative degree adverb (*most, least*)
93	RL	locative adverb (e.g., *alongside, forward*)
94	RP	prep. adverb, particle (e.g., *about, in*)
95	RPK	prep. adv., catenative (*about* in *be about to*)
96	RR	general adverb
97	RRQ	wh- general adverb (*where, when, why, how*)
98	RRQV	wh-ever general adverb (*wherever, whenever*)
99	RRR	comparative general adverb (e.g., *better, longer*)
100	RRT	superlative general adverb (e.g., *best, longest*)
101	RT	quasi-nominal adverb of time (e.g., *now, tomorrow*)

102	TO	infinitive marker (*to*)
103	UH	interjection (e.g., *oh*, *yes*, *um*)
104	VB0	*be*, base form (finite i.e., imperative, subjunctive)
105	VBDR	*were*
106	VBDZ	*was*
107	VBG	*being*
108	VBI	*be*, infinitive (*To be or not..., It will be...)*
109	VBM	*am*
110	VBN	*been*
111	VBR	*are*
112	VBZ	*is*
113	VD0	*do*, base form (finite)
114	VDD	*did*
115	VDG	*doing*
116	VDI	*do*, infinitive (*I may do..., To do...)*
117	VDN	*done*
118	VDZ	*does*
119	VH0	*have*, base form (finite)
120	VHD	*had* (past tense)
121	VHG	*having*
122	VHI	*have*, infinitive
123	VHN	*had* (past participle)

124	VHZ	*has*
125	VM	modal auxiliary (*can, will, would,* etc.)
126	VMK	modal catenative (*ought, used*)
127	VV0	base form of lexical verb (e.g., *give, work*)
128	VVD	past tense of lexical verb (e.g., *gave, worked*)
129	VVG	-ing participle of lexical verb (e.g., *giving, working*)
130	VVGK	-ing participle catenative (*going* in *be going to*)
131	VVI	infinitive (e.g., *to give... It will work...)*
132	VVN	past participle of lexical verb (e.g., *given, worked*)
133	VVNK	past participle catenative (e.g., *bound* in *be bound to*)
134	VVZ	-s form of lexical verb (e.g., *gives, works*)
135	XX	*not, n't*
136	ZZ1	singular letter of the alphabet (e.g., *A, b*)
137	ZZ2	plural letter of the alphabet (e.g., *s*)

Appendix D

A sample sentence of the BIOCOR part-of-speech-tagged using CLAWS7

Number of the sentence in the text	Number of the word in the sentence (expressed in tens)	Token in the sentence	Part-of-speech UCREL CLAWS7 code
0000015	010	At	II
0000015	020	first	MD
0000015	030	sight	NN1
0000015	040	you	PPY
0000015	050	might	NN1
0000015	060	think	VVI
0000015	070	that	CST
0000015	080	plants	NN2
0000015	090	are	VBR
0000015	100	an	AT1
0000015	110	exception	NN1
0000015	120	to	II
0000015	130	the	AT
0000015	140	rule	NN1

0000015	150	that	CST
0000015	160	all	DB
0000015	170	organisms	NN2
0000015	180	respond	VV0
0000015	190	to	II
0000015	200	stimuli	NN2
0000015	210	.	.

Appendix E

Grammar features describing ESP registers (Biber, 1998)

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 F

Selection of linguistic features from Biber's (1998) framework describing ESP registers which bear relevance to the grammatical component of the POTAI

agentless passive, *by* passives, causative subordination, conditional subordination, *do* as pro-verb, final prepositions, infinitives, necessity modals, past tense verbs, perfect aspect verbs, possibility modals, prediction modals, present participial clauses, present tense verbs, suasive verbs, tense verbs, *that* deletion, *wh*-clauses, *wh*-relative clauses on object position, *wh*-relative clauses on subject position

Appendix G

Interview protocol with English teachers instructing in the 9th grade

Part One:

1. When did you obtain you degree in English teaching?
2. Which university issued your degree in English teaching?
3. How long have you been teaching English?
4. Where did you teach English since obtaining your degree?
5. How long have you been teaching English in the bilingual programme of the school?
6. How long have you been teaching in the bilingual immersion programme of the school?
   (in the 'zero-year' intensive language programme)

Part Two:

1. In your opinion, what kind of skills do 9th graders need to master in order to pass the reading part of their end-term FCE exam?
   (prompts: skimming, scanning, extensive reading, intensive reading, extracting information)
2. Keeping the reading tasks of the FCE exam in mind, which of the following grammar points do you think 9th graders need to be closely familiar with? Why do you think so?
   a, tenses
   b, indirect speech
   c, conditionals
   d, passive voice
   e, relative clauses
   f, infinitives
   g, prepositions at the end of clauses
   h, question tags

I, modal verbs

3. What other grammar points do you think 9th graders should be familiar with in order to pass their end-term FCE exam successfully?

   (prompt: verbs and their prepositions; static vs. dynamic verbs; structures e.g., the very; a most…; the more – the less)

4. What other comments do you have regarding the reading part of the end-term FCE exam for 9th graders?

Appendix H

Interview protocol with biology teachers instructing in the 10th grade

Part One:

1. When did you obtain you degree in English teaching?
2. Which university issued your degree in English teaching?
3. How long have you been teaching biology?
4. Where did you teach biology since obtaining your degree?
5. Did you obtain any other degrees in any other subjects?
6. How long have you been teaching biology in the bilingual programme of the school?

Part Two:

1. In your opinion, what kind of study skills do 10th grade bilingual students need to master by the end of the 9th grade in order to study biology from an English-language biology textbook? (prompts: skimming, scanning, extensive reading, intensive reading, extracting information)
2. In order to perform with outstanding result in biology, what other skills and knowledge do you think 10th grade bilingual students need to master by the end of the 9th grade? (prompts: summary writing, presentation skills, note taking, extensive biology ESP vocabulary, extensive academic English)
3. Keeping the biology textbook used in the 10th grade in mind, which of the following grammar points do you think 10th grade bilingual students need to be closely familiar with? Why do you think so?

   a, tenses

   b, indirect speech

   c, conditionals

   d, passive voice

e, relative clauses

f, infinitives

g, prepositions at the end of clauses

h, question tags

I, modal verbs

(Each category is exemplified by biology sample sentences.)

4. What other grammar points do you think 10th grader bilingual students should be familiar with in order to study biology from an English-language biology textbook?

   (prompt: verbs and their prepositions; static vs. dynamic verbs; structures e.g., the very; a most…; the more – the less)

5. How well do you think 10th grade bilingual students are prepared in English to study biology in English?

Appendix I

List of biology terms and their frequencies in Bands 4-10 in the BIOCOR

(raw frequency and relative frequency expressed in percentages):

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

Appendix J

List of academic English items and their frequencies in Bands 4-10 in the BIOCOR (raw frequency and relative frequency expressed in percentages):

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

Declaration form
for disclosure of the doctoral thesis[5]

I. The data of the doctoral thesis

Name of the author: **Natalia Borza**
MTMT-identifier: 10037847
Title and subtitle of the doctoral thesis: **Register analysis of English for specific purposes discourse. An in-depth exploratory and descriptive theory- and corpus-based study of the case of biology texts in secondary education in Hungary.**
DOI-identifier[6]: 10.15476/ELTE.2015.174
Name of the doctoral school: Doctoral School of Education
Name of the doctoral programme: Programme in Language Pedagogy
Name and scientific degree of the supervisor: Krisztina Károly, PhD, Habil.
Workplace of the supervisor: Eötvös Loránd University, Centre for Teacher Training

II. Declarations

1. As the author of the doctoral thesis,[7]
a) I agree to public disclosure of my doctoral thesis after obtaining a doctoral degree in the storage of ELTE Digital Institutional Repository. I authorize Madar Veronika, the administrator of the Student Affairs and Registrar's Department to upload the thesis and the abstract to ELTE Digital Institutional Repository, and I authorize the administrator to fill all the declarations that are required in this procedure.
b) I request to defer public disclosure to the University Library and the ELTE Digital Institutional Repository until the date of announcement of the patent or protection. For details, see the attached application form;[8]
c) I request in case the doctoral thesis contains qualified data pertaining to national security, to disclose the doctoral thesis publicly to the University Library and the ELTE Digital Institutional Repository ensuing the lapse of the period of the qualification process.;[9]
d) I request to defer public disclosure to the University Library and the ELTE Digital Institutional Repository, in case there is a publishing contract concluded during the doctoral procedure or up until the award of the degree. However, the bibliographical data of the work shall be accessible to the public. If the publication of the doctoral thesis will not be carried out within a year from the award of the degree subject to the publishing contract, I agree to the public disclosure of the doctoral thesis and abstract to the University Library and the ELTE Digital Institutional Repository.[10]

---

[5] Endorsed by Senate Rule CXXXIX/2014. (VI. 30.) on the amendment of the Doctoral Regulations of ELTE. Effective date: 01 July 2014.
[6] Filled by the administrator of the faculty offices.
[7] The relevant part shall be underlined.
[8] Submitting the doctoral thesis to the Disciplinary Doctoral Council, the patent or protection application form and the request for deferment of public disclosure shall also be attached.
[9] Submitting the doctoral thesis, the notarial deed pertaining to the qualified data shall also be attached.
[10] Submitting the doctoral thesis, the publishing contract shall also be attached.

2. As the author of the doctoral thesis, I declare that
a) the doctoral thesis and abstract uploaded to the ELTE Digital Institutional Repository are entirely the result of my own intellectual work and as far as I know, I did not infringe anyone's intellectual property rights.;
b) the printed version of the doctoral thesis and the abstract are identical with the doctoral thesis files (texts and diagrams) submitted on electronic device.

3. As the author of the doctoral thesis, I agree to the inspection of the thesis and the abstract by uploading them to a plagiarism checker software.

Budapest, 26 October, 2015 ……

Signature of thesis author