# HUNGARIAN 8TH GRADERS' WRITING SKILLS IN ENGLISH

# A CRITERION- AND CORPUS-BASED ASSESSMENT PROJECT

**Bors Lídia**

Consultant: Dr Horváth József

Doctoral Programme in Applied Linguistics
University of Pécs, Hungary
2008

**TABLE OF CONTENTS**

**Acknowledgements**

I would first like to thank my consultant, József Horváth, for his advice and scrupulous help on writing my dissertation. I am also indebted to Marianne Nikolov for sharing her expertise and encouraging me in difficult days of doubt.

I am grateful to the authors of the Croatian-Hungarian research project: Jelena Mihaljević-Djigunović, Marianne Nikolov, and István Ottó who made Hungarian students' written products available for investigation.

My special thanks go to raters: Viktória Kusz, Éva Pálfy, Richárd Pércsich, and Andrea Sulyok, for the assessment of a sample of texts and giving feedback on the process. I am grateful to Magdolna Lehmann for sharing her expertise on vocabulary assessment and to Robert Wołosz for designing an XML tool for me to analyze students' accuracy in EFL. I owe special thanks to István Ottó for his help and suggestion on statistical analyses.

I thank Lídia Dobány, Mátyás Dobány and Péter Katymarac who helped me with typing in texts, and to Richárd Pércsich for his support in solving computer problems whenever I was stuck.

I am also indebted to Paul Nation and his colleagues who made their research instruments available to the public on the Internet.

Finally, the dissertation would not have been possible to write without the texts I analyzed. Special thanks go to the young authors.

## *List of abbreviations and acronyms*

| | |
|---|---|
| CA | Contrastive analysis |
| CEFR | Common European Framework of Reference |
| EA | Error analysis |
| EFL | English as a foreign language |
| ESL | English as a second language |
| EXP | Expletive |
| FA | Frequency analysis |
| FC | Frame Curriculum |
| FL | Foreign language |
| GLR | General linguistic range |
| GRA | Grammatical accuracy |
| GSL | General Service List of English Words |
| IL | Interlanguage |

| | |
|---|---|
| LAD | Language acquisition device |
| LD | Lexical density |
| LFP | Lexical Frequency Profile |
| LV | Lexical variation |
| L1 | First language/mother tongue |
| L2 | Second language/ foreign language |
| NCC | National Core Curriculum |
| NEG | Negation |
| OOA | Obligatory occasion analysis |
| OWP | Overall written production |
| PLU | Plural form of nouns |
| PRC | Present continuous tense |
| R | Rater |
| RQ | Research question |
| s | Sentence |
| SES | Socio-economic status |
| SLA | Second Language Acquisition |
| TTR | Type-token ratio |
| TTRL | Type-token ratio of lexical words |
| VOC | Vocabulary range |
| WL | Word list |
| YLLs | Young language learners |

## *List of Tables*

## *List of Figures*

# Introduction

As advances in technology allow people living in different societies and cultures in the world to interact with each other, communication across languages becomes more essential. Although in the 21st century needs and uses of writing have changed dramatically, writing remains an important medium of communication. It is central to people's personal experience and social identities, and often forms a basis for evaluation. In order to achieve their aims, both learners and adults need to know the basics of successful writing. Most school testing requires writing assignments: secondary students are expected to produce high quality texts to graduate from secondary school and enter tertiary education. Adults need to write clear and persuasive curricula vitae and cover letters when applying for a job. All age-groups need to use the Internet to get up-to-date information and keep in touch with friends, to mention just a few areas where writing skills are crucial.

From the beginning of the 90s, together with social changes in Hungary, the need for using foreign languages (FL) and the prestige of being able to communicate in them has grown immensely. In order to communicate in a FL, students need to use not only the spoken, but also the written forms of communication. So as to help students become proficient writers by the end of secondary education, writing pedagogy needs to start as early as possible, that is, the basics of good writing in both the mother tongue (L1) and a second language (L2) are supposed to be learnt in primary education as part of literacy development. When examining students' writing processes and the products, researchers have concentrated mainly on participants of secondary and tertiary education. Primary-school students' writing has rarely been in the focus of researchers' attention (Leki, 2002; Molnár, 2002; Silva & Brice, 2004). The possible reason for this may be the limited linguistic and topical knowledge, and strategic competence of young language learners (YLL). This is why I aim to explore the writing skills of this neglected age group.

The students involved in my research were 13-14 years old at the time of data collection. Their age was already past the age-category of young learners which falls approximately between five and twelve years. In another view (McKey, 2006, p. 1), YLL are those who are learning an L2 or FL during the first six or seven years of formal schooling. The 8th-graders participating in this research can marginally be included in this category, as the years of their L2 learning (from 11 to 14) overlap with those of young learners. This is why I will relate some of the discussed issues to young learners, although I do not intend to examine the research area related to young learners in depths.

In addition to professional interest, I also have personal motivation for writing this study: for 20 years I worked as a teacher of English in primary education and continuously faced the challenge of assessing students' performances as objectively as possible in order to provide feedback on their development. The perplexity of the dubious reliability of my assessment lingered all the time, especially in the case of students' written performances, where qualities of writing may have been quantified if time constraints had permitted. Questions like 'How should I assess students' compositions taking into consideration all the ingredients of good writing? How can I be sure about the reliability of my holistic decision concerning the content, vocabulary, grammar and cohesive devices displayed in students' tests?' emerged and remained unanswered for a long time. Let me share a classroom story that illustrates multiple viewpoints regarding students' performances.

It happened in year seven. Pupils' home assignment was to write a new, 'inside-out' version of the episode of the story we had read in the previous lesson, that is, to interchange main characters and their respective actions. The students responded rather creatively, and we all enjoyed listening to their stories. Besides listening, pupils had a task: they were asked to identify the best story of all. After listening to all compositions, they voted and explained why they liked the selected piece of writing.

Finally, to sum it all up, I also appraised a few outstanding texts. When I finished, one of the students raised her hand and expressed her discontentment regarding my neglecting one of the best compositions. I was astonished, as I did not have a high opinion of that piece of work. Nevertheless, I promised to read it and revise my first impression. So I did, and it proved a lot better in written form.

This event made me stop and think about the fragile reliability of teachers' assessment. This example is not typically representative of the way students' written performances are assessed, as in this case opinions had to be formed promptly after the first listening to students' work. Nevertheless, it shows that the teacher's first holistic impression may misjudge pupils' performances, and, that there exist different viewpoints concerning these products, which are definitely worth considering.

The other area of personal interest concerns 'average' pupils' knowledge in schools of Baranya county. As an advisor for English as a foreign language (EFL) teachers for more than 20 years in county and national competitions, I regularly meet the best-achieving students coming from a lot of schools. On the other hand, I have observed EFL classes both in the city of Pécs and county settlements. The contrast between high-achievers' outstanding performances and the average students' language knowledge observed in EFL classes was

huge. I was interested to find out if these observations could be supported by language test results, and what level of L2 competence average pupils achieve by the end of their primary education. In 1997 and 2000 I participated in research projects exploring the language competence of primary-school students in Pécs (Bors, Nikolov, Pércsich, & Szabó, 1999; Bors, Lugossy, & Nikolov, 2001), but the mapping of pupils' knowledge living in towns and villages of Baranya was still unexplored.

Later, when I was involved in a Croatian-Hungarian research project (Mihaljević-Djigunović, Nikolov, & Ottó, 2008) as an organiser for Baranya and an assessor of writings, I felt that a great opportunity knocked on my door. It was time I set to work to answer both of my questions: the one concerning students' L2 performance in Baranya primary schools, and the other regarding the reliability of teacher assessment. Therefore, in my dissertation I examine Hungarian primary-school students' scripts in EFL. In order to do so I analyze students' English scripts taken as part of a complex language proficiency assessment project that measures their reading, writing and listening skills, as well as the speaking performance of a sub-sample.

All 231 participants studied EFL in the 8$^{th}$ grade of primary schools in villages, towns, and the county seat, Pécs, in 2004 when the research project was carried out. Pupils' task was to compare two pictures illustrating a four-member family in their living room. The compositions are scrutinized with the help of text-linguistics and corpus linguistics.

## *Research questions and an overview of the dissertation*

So as to measure the reliability of the assessment of students' written performances, I will explore both holistic and analytic approaches of decision making, and find out how the results are related to each other. Accordingly, the aims of the dissertation are as follows:

1 to explore 231 Hungarian primary school leavers' writing skills in L2 English
2 to assess students' L2 writing competence by using an analytic assessment scale
3 to analyze language accuracy of L2 scripts regarding four language forms: present continuous tense (PRC), expletives (EXP), plurals (PLU) and negation (NEG) quantitatively
4 to trace developmental patterns of L2 acquisition in students' scripts
5 to assess a sample of students' L2 writing holistically and examine the relationship between holistic and analytic assessment; and analyze interrater reliability.

Table 1: Research questions, data sources, and methods of analysis

| **Research questions (RQ)** | **Data source** | **Method and section in which the RQ is answered** |
|---|---|---|
| 1. How does Baranya students' writing proficiency compare with their other EFL language skills? | Data on four skills gained from an international survey points received for writings | Analytic scale, descriptive statistics; 4.8.1.1 |
| 2. How does Baranya primary-school learners' L2 proficiency compare with the L2 proficiency of the same age-group in a national sample? | Results of a national survey | Descriptive statistics; 4.8.1.1 |
| 3. What level of writing proficiency do Baranya primary-school students reach in English by the end of their primary education? | Students' written performance in English; points received for writings | Descriptive statistics; 4.8.1.2 |
| 4. What is the relationship between students' writing proficiency and the curricular achievement targets? | Points received for writings, National Core Curriculum, CEFR | Content analysis; 4.8.1.2 |
| 5. What is the relationship between students' socio-educational background and their writing proficiency? | Points received for writings socio-educational background of schools | Descriptive statistics; 4.8.1.2 |
| 6. What can we learn about students' texts with the help of a corpus analysis? How accurately do 8th graders use simple target language forms (PRC, EXP, (there/it), NEG, PLU)? What developmental patterns emerge? | Students' written performance in English; Morpheme studies | Corpus analysis; 4.8.2 |
| 7. What is the relationship between criterion-based and corpus-based measurement of accuracy? | Points received for writings, data gained from linguistic analysis | Correlation analysis; 4.8.3, Content analysis; 4.8.5 |
| 8. How are raters' opinions of sample tests related to each other? | Sample tests rank ordered, then assessed by the CEFR criteria for writing | Correlation analysis 4.8.4 4.8.5 |
| 9. How does criterion-based assessment compare with holistic assessment? | Points received for writings, rank-ordered sample tests by assessors, Sample tests assessed by the CEFR criteria for writing | Correlation analysis; 4.8.4 4.8.5 |
| 10. How do raters' interpretations of CEFR scales compare? | Sample tests assessed by the CEFR criteria for writing | Correlation analysis; 4.8.4, Qualitative analysis of rater feedback |
| 11. What are typical performances like? What developmental levels do they reflect? | Five students' written performance in English; Morpheme studies | Descriptive statistics, content analysis; 4.8.5 |
| 12. What range of vocabulary characterizes 8th graders? What is the relationship between criterion-based and corpus-based measurement of accuracy and vocabulary? | Five students' written performance in English | Range and frequency analysis; 4.8.5 |

So as to answer these research questions, I created the learner corpus and error tagged and analyzed 231 texts for specific linguistic features in order to identify the developmental sequences of selected English morphemes through writing. This is the first study in Hungary

to compare students' EFL performances with the curricular requirements, and the levels of the *Common European Framework of Reference* (*CEFR,* Council of Europe, 2001). The reliability of assessment is investigated in a variety of ways: (1) by examining relationships between criterion- and corpus-based assessment, (2) by comparing various types of measurement (holistic and analytic), and (3) by comparing a rater's assessments based on two different analytic scales.

The dissertation is divided into four chapters. The theoretical background to the research studies is outlined in Chapters 1, 2 and 3. The first chapter gives an overview of the nature of writing as connected to reading abilities both in L1 and L2. Next, it presents the main approaches to writing, and summarises the results of studies on the development of learner language. Issues of measuring the richness of vocabulary will be also highlighted in this chapter. Finally, Chapter 1 provides an overview of the research methodology of writing studies. Chapter 2 discusses the basic considerations in assessing writing; overviews the aims and related proficiency levels of the *CEFR* (Council of Europe, 2001), and discusses corpus-linguistic techniques in analysing writing. Chapter 3 describes the context of the study: it presents literacy requirements in L1 and L2 in Hungary, and investigates the relationship between the Hungarian national requirements and the *CEFR*. Then, it summarises recent research findings on Hungarian students' writing skills in both the L1 and EFL.

Chapter 4 presents the main study including the research questions (RQs), the participants, the data collection instruments (tasks, assessment scales), the procedures, the findings, and the discussion of how they are interrelated. General conclusions revisit the findings, discuss the pedagogical implications of the results, and present the limitations of the study.

# Chapter 1: What is involved in writing?

## *1.1 Definition of terms: literacy and bilingual literacy*

Writing is an integral part of literacy; therefore, I start defining terms by the description of the broader category. Literacy has been defined from different viewpoints. According to the *Dictionary of Language Teaching and Applied Linguistics* literacy means 'the ability to read and write in a language', while "functional literacy refers to the ability to use reading and writing skills sufficiently well for the purposes and activities which normally require literacy in adult life or in a person's social position" (1992, p. 216). The first definition reflects an extremely general view, while the latter implies successful performance from a social viewpoint. However, researchers have interpreted this concept in several ways, focusing on different aspects of language use, the most important of which I will overview and summarize in this section.

Linguistic dimensions of literacy involving the mastery of a writing system cannot be separated from cognitive processes needed to achieve this mastery. Literacy, first of all, is related to language and knowledge of how it is used, and secondly, to the writing system. One of the approaches sees literacy as a psycholinguistic process (Barnett, 1989; Gathercole & Baddeley**,** 1993; Oakhill, 1993) connected to dealing with texts. This process includes letter recognition, encoding, decoding, word recognition, as well as higher-order cognitive operations of sentence comprehension, and the comprehension of a text.

Reading, as well as writing, demands the reader's active participation at a cognitive level: when reading a text, readers must rely on their knowledge of written symbols and relate those to the knowledge of language, texts, content areas, and the world, in order to build up the meaning of a text. Both reading and writing can be seen as acts of meaning construction in which individuals make connections between textual elements and existing knowledge to design new knowledge.

According to this cognitive view (Bereiter & Scardamalia, 1987; Gathercole & Baddeley**,** 1993; Hayes & Flower, 1980; Oakhill, 1993; Raimes, 1987; Zamel, 1983), writing, together with reading, is a central aspect of literacy, a learnt ability that facilitates logical thinking and participation in different roles of modern society. This view focuses on cognitive skills and the text itself. It frames literacy as a set of discreet, value-free technical skills of encoding and

decoding meanings, perceiving shape-sound correspondences, mostly acquired in educational institutions.

Students developing literacy in two or more languages can learn the psycholinguistic process in one language, but must learn the specific symbol system, words, grammar, and text structure of each language. Bialystok (2001, p. 5) calls learners bilingual who develop some level of proficiency in two or more languages. The cognitive development of bilingual learners has been studied over many years (Bialystok, 2001; Bialystok & Hakuta, 1999; Cummins, 2001), but, despite this, Bialystok points out that research literature on the development of bilingual children is thin due to the difficulty of doing research in this area. Cummins found evidence of a positive association between bilingualism and both cognitive ability and divergent thinking.

According to Kern (2000, p. 3), text-centric and cognitive-skills-oriented views of literacy share a number of limitations in the context of second language education, as they interpret literacy as an end product of instruction "instead of as a variable set of processes contingent on textual, cognitive, and social factors." Reflecting upon dangers of strict public demands on education for accountability, Kern claims that students may develop limited literacy skills if teachers concentrate on instruction to reach the criterion level of learner skills without trying to integrate them. He also points out the importance of contextual factors stating that the individual range of abilities and knowledge is used for versatile purposes and functions.

Another perspective focuses on literacy in relation to context and process. According to Gee (1989, p. 23), "literacy is control of secondary uses of language (that is, uses of language in secondary discourses)." Primary discourses take place among intimates who share a great deal of knowledge, whereas secondary discourses serve for communication in institutions such as schools, workplaces, or stores. The functions of literacy vary according to institutional context so much (Gee, 1996) that an individual cannot be accepted as a full member of a given group until he or she has mastered the relevant discourse (ways of behaving, interacting, valuing, speaking, often reading, and writing). Kaplan and Palmer (1992) also state that literacy is a complex phenomenon which is difficult to define properly. They see literacy as a set of practices that vary depending on the context, the functional uses of literacy and the demands of larger social structure. August and Hakuta's (1997, p. 54) view on literacy being a social practice that "assumes participation in a community that uses literacy communicatively" echoes Kaplan and Palmer's (1992) definition.

This claim has not changed in the past ten years, as Cummins (2007, cited in Blades, 2007, no page) also pointed out at a recent annual conference of the organization of California

Teachers of Other Languages (CATESOL) that literacy attainment is directly related to literacy engagement. “When students’ identities are affirmed in the classroom, they feel comfortable investing their identities into the literacy activities and practices, and they learn more.” When children are encouraged to share unique personal experiences, when use of their first language is not discouraged, when ‘decoding’ techniques are not the sole and ultimate aim of instruction, when students feel they have a voice in the classroom, they will become engaged with their own education. Cummins (2007, cited in Blades, 2007, no page) accentuated that “new understandings are constructed on a foundation of existing understandings and experiences.” These ideas should also be implemented in Hungarian educational institutions where literacy instruction generally does not rely on students’ active engagement and reading texts discussed in L1 and L2 classes are often not motivating for learners (Nagy, 2004).

This social view of literacy highlights that writing and reading vary depending on context and cannot be simplified to a set of cognitive or technical abilities. As writing is a means of communication in social settings, it cannot be merely an observable exercise of abstract skills. According to this view, there is no single acceptable literacy, but a wide variety of practices appropriate for particular times, places, participants and purposes.

Becoming literate in a bilingual, English as a second language (ESL) or FL context is one of these various practices, which raises several questions concerning the influence of the first language on the second language, or vice versa. In the next section I aim to review Cummins’ (1981) Interdependence Hypothesis, the notion that there is an underlying common proficiency across languages. The following set of principles suggests an interpretation of literacy relevant to the L2 context.

The sociocognitive view, proposed by Kern (2000, p. 16), can be seen as an integration of previously discussed standpoints related to second language education. He outlines this view in seven principles suggesting that literacy involves:

(1) *interpretation practiced by both writers and readers of texts*: writers manifest their view of the world in their texts, whereas readers try to understand the mediated ideas by using their own conceptions of reality.

(2) *collaboration*: writers, having considered readers’ expectations, decide what they need to say and what is common sense, readers in turn must contribute their motivation and knowledge to decode meaning from the text.

(3) *conventions*: they guide readers and writers of different cultures in creating and interpreting texts, but they can be modified for individual purposes; they are not universal phenomena.

(4) *cultural knowledge*: a system of beliefs, customs, and values. L2 readers and writers may misunderstand texts of the target language due to the differences in these cultural systems.

(5) *problem solving*: disentangling embedded meaning from texts, as described earlier in this section, in relation to the psycholinguistic approach.

(6) *reflection on language* and the world and *self-reflection* are also induced by practicing literacy.

(7) *language use*: besides using writing systems, and lexical and grammatical knowledge, is also a constituent of literacy. It involves the knowledge of creating discourse.

Kern's (2000) integrated view on literacy is based on the ideas of researchers in cognitive psychology (Bereiter & Scardamalia, 1987; Gathercole & Baddeley**,** 1993; Hayes & Flower, 1980; Oakhill, 1993; Raimes, 1987; Zamel, 1983), as well as educators in the social-cultural field (August & Hakuta, 1997; Gee, 1996; Kaplan & Palmer, 1992). Halliday and Hasan's (1976) definition of text, Hymes' (1972) term 'communicative competence', Widdowson's (1978) suggestion to focus on language use instead of language usage, and Breen and Candlin's (1980) term 'language as communication' all contributed to the pedagogical shift from structural to communicative frameworks.

These principles of literacy can also be applied to human communication; as Kern (2000, p. 17) puts it: "literacy involves communication". The pedagogical implication of bridging literacy and communication is that practicing literacy in a non-native language should involve learning not only vocabulary and grammar, but discourse as well. Moreover, students need to learn to deal with uncertainties and try to organize their thoughts in new ways instead of simply learning facts about the other culture. As Kern (2000) claims, becoming literate does not mean simply the achievement of a particular level of reading or writing performance. In his view, it also involves individuals using reading and writing as tools for thinking and learning in order to expand their understanding of themselves and the world.

After reviewing different perspectives on the complex term of literacy, in the next section I intend to focus on the productive form of literacy by relating it (1) to the other productive skill, speaking, then (2) to the other visual way of communication, to reading. Then, three basic approaches to the examination of written products will be demonstrated echoing the

main views on literacy described above: decontextualised text analysis; psycholinguistic and sociolinguistic approaches. Finally, the relationship between writing and L2 proficiency will be considered.

## *1.2 The nature of writing ability*

In language teaching it is traditional to categorize instances of language use into four skills: reading, writing, listening, and speaking, which can be paired according to either the channel of communication (aural versus visual) or mode (productive vs. receptive). Instead of trying to come up with a concise definition of writing, first, I will investigate the relationship between writing and the two skills closely linked to it: speaking and reading. Then, I will widen the concept of writing ability through the lenses of the disciplines of (1) linguistics, (2) cognitive psychology, and (3) sociolinguistics contemplating educational implications as well.

### 1.2.1 Speaking and writing

Before the 1960s priority was given to spoken language in education, because writing was defined as an orthographic representation of speech based on Bloomfield's and Fries' concept on language as speech (Kroll, 2003, p. 16). Today it is recognized that written language is not simply speech put on paper. Although their lexical and grammatical resources are the same, writing is often more dense lexically (Halliday, 1989), with a higher proportion of content words than speech, and, at the same time, it is grammatically less complicated with shorter stretches of clauses. Another disparity between them is that they rely on different lexical and syntactic structures. From an emotional perspective, speaker involvement with the listener is much higher in the case of conversations than that of the writer with the reader. The writer's detachment is observable by comparing passive constructions and abstract subjects used in speech and writing. All these and many other disparities between the two productive channels of communication should not, however, form a basis for claiming superiority of either.

After long years of controversy, Grabe and Kaplan (1996) summarized linguists' and educational researchers' contradictory positions regarding the superiority of one or the other skill and claimed that a consensus had been emerging among researchers. Instead of contemplating superiority, this new perspective underlined the dimensions of diversity: oral and written texts do vary across a number of dimensions including textual features, cognitive processes involved in text production, as well as socio-cultural norms and patterns of use. The awareness of the relationship between speaking and writing is important for both language teachers and language testing experts, as test developers need to be aware of the fact that

writing draws on many of the same linguistic resources as speaking, but it also relies on different mental processes.

### 1.2.2 Reading and writing

#### 1.2.2.1 Reading in L1

The other skill closely related to writing is reading, which 'is not simply an act of absorbing information, but a communicative act that involves creating discourse from text (Kern, 2000, p. 107). Kern argues that the sense of context makes face-to-face oral interaction different from reading and writing. Whereas situational context is shared in face-to-face interaction, when reading, it is not readily inferable from the linguistic text. In written interaction, on the other hand, the writer must provide relevant contextual information for the reader to understand intended messages. The reader, in turn, is required to reconstruct the background information based on the writer's cues and on his own experience and knowledge (Nystrand, 1989, p. 75).

Kern (2000, pp. 107-127) explains reading in terms of a design model, where every text a reader encounters is the result of a particular act of design. It encodes particular reader-writer relationships and a particular interpretation of the world. Readers try to decode these relationships by using their 'available designs' (knowledge of language, genres, schemata, etc.) in order to create meaning. These interpretations depend on readers' purposes, the topic, and the physical situation. If readers share the same 'available designs' to a given text, their mental representations will be similar. All the above is influenced by the larger sociocultural context that is related to the functions of reading, the social status of readers, and the discourse community they belong to.

Concerning the instructional steps to be taken to reach available designs, that is, the beginning methods for teaching literacy, in English-speaking countries an intensive controversy has developed between advocates of skill-building and those of the 'whole language' method (Krashen, 2003). While skill-builders emphasize the conscious learning of letter-sound correspondences, the supporters of 'whole language' claim that this knowledge emerges naturally as children learn to read. Krashen reviews several method comparison studies (e.g., Evans & Carr, 1985; Eldridge, 1991 cited in Krashen, 2003, p. 152) and finds strong results supporting the Comprehension Hypothesis claiming that we acquire language by understanding it, but other conditions, such as open attitude, low affective filter, and developmentally appropriate input must also be met for comprehension.

Researchers of students' L1 reading competence in Hungary focus on reading instruction, stressing the role and responsibility of teachers. Let us hear a few recent voices articulating future tasks in L1 reading instruction.

Nagy (2004) attributes Hungarian students' poor reading results in the international measurements to traditional methods applied in reading instruction. He argues that the majority of students are able to become proficient readers if they are trained with appropriate methods. He points out that one of the preconditions of fluent reading is the knowledge of the most frequent 5,000 words, but for novice readers short texts constructed from the most frequent 2,000 words are sufficient. So as to enhance pupils' reading comprehension he suggests the use of motivating, comprehensive texts for beginners, as well as playful methods of teaching basic vocabulary. His view on texts to be applied is in line with Krashen's (2003) Comprehension Hypothesis.

Molnár (2006), relating text interpretation to reading instruction, also emphasizes the responsibility of teachers, as senior members of the institutional discourse community, in developing students' reading comprehension. She perceives teachers as mediators between textbook writers and students, who are the targeted readers of textbooks. Molnár claims that teachers' background knowledge of texts, their mental representations concerning the theme, together with the perception of their role as teachers highly influence how the students will be able to interpret texts.

Csíkos and Steklács (2006) point out that learners' metacognitive strategies need to be developed in primary reading classes, as their flexible use contributes to both reading comprehension and effective learning. This view corresponds with Kern's (2000) concept of literacy, which involves using reading and writing as tools for thinking and learning, in order to enhance one's understanding of oneself and the world.

We have seen views on reading as a process of creating meaning, as well as concerns related to reading instruction in L1. Reading in L1 is a complex process involving cognitive, metacognitive and social constituents; the same skill performed in L2 is interdependent with the first language processes.

#### 1.2.2.2 Reading in L1 and L2: The Interdependence Hypothesis

Second language reading performance is a more complex phenomenon; it is influenced by linguistic, cognitive, and sociocultural factors.

Regarding linguistic background, Grabe (2002, p. 55) argues that in comparison with L1 readers, L2 readers begin to read without the initial language knowledge that can be assumed

to L2 readers. Most L1 readers have a vocabulary of at least 6,000 words and the basic structures of their language when they start receiving reading instruction. L2 readers, in contrast, may have relatively little spoken language knowledge of the L2 when reading instruction begins.

Investigations on the relationship of L1 and L2 reading (Goodman, 1971, 1985) found that the basic process of reading is universal, involving the formation, testing, modification and confirmation of hypotheses based on the text itself as well as the reader's knowledge. Cummins (1981, pp. 23-24), researching bilinguals' literacy, adopted this universalist stance stating that the literacy-related aspects of bilinguals' proficiency in L1 and L2 are interdependent across languages. This Interdependence Hypothesis claims that once one develops an ability to deal with context-reduced uses of language, that ability does not need to be reacquired when learning a new language.

However, Cummins (1976) claimed that bilinguals would have to attain a 'threshold level of L2 competence' to successfully realize linguistic interdependence. He also pointed out that this threshold level should not be interpreted as something static, but rather as a variable feature depending on the time spent using the L2 and the cognitive operations the learner needed to perform. Taillefer (1996) showed that the relative importance of second language proficiency varied with the type of reading task; more difficult tasks relied on a higher language threshold level. Other studies suggested that successful second language reading is not simply the matter of transferring L1 abilities when the threshold level of L2 proficiency is acquired; L2 reading can remain less efficient than L1 reading well past the threshold point of language proficiency and, also, reading in different languages can involve qualitatively different perceptual and processing strategies (Favreau & Segalowitz, 1982; Koda, 1993).

Reading and writing are intrinsically linked complementary processes; "… the thought processes that writers go through in the act of writing promote a sensitivity to language that makes close, analytic reading possible" (Kern, 2000, p. 171). If reading is based on creating discourse from texts, writing involves designing texts for potential readers; it supposes the knowledge of rhetorical and stylistic devices, genres, and formatting conventions. Thus, similarly to reading, we can think of writing as design). The issues of writing in a second language will be attended to later, after reviewing the most influential approaches to writing in an L1. Now, I will show what research tells us on the relationship between reading and writing. The main issue is whether writing develops by writing more or by reading more.

#### 1.2.2.3 Reading and writing in L1

Smith (1983) maintains that the conventions of writing are acquired by reading; to learn how to write in different styles and genres one needs to read them. The hypothesis does not predict a perfect correlation between the amount of pleasure reading accomplished and writing quality, it rather says that there is a minimum amount of reading that every good writer has done. Smith (1981) argues that at least two conditions must be met for a reader to acquire a writer's code: (1) the expectation of success (2) and the feeling of membership of the 'club'. These conditions reflect Gardner and Lambert's (1972) views on integrative motivation. They claim that success in second language acquisition is dependent on certain affective variables.

#### 1.2.2.4 Reading and writing in L2

Krashen's (1981) Affective Filter Hypothesis resembles Smith's (1981) conditions for success in writing. When success is assured and the anxiety level concerning language acquisition is low, the acquisition of writing competence is more probable. Krashen reviews several studies indicating that voluntary pleasure reading contributes to the development of writing ability (Applebee, 1978; Donalson, 1967; Woodward & Phillips, 1967 cited in Krashen, 1984, p. 78). Krashen (1984, 1993) hypothesizes that writing competence comes only from large amounts of self-motivated reading for interest. It is acquired subconsciously while reading. Leki (2002) underlines the impact of Krashen's work on L2 writing teachers' thinking; their move from a narrow focus on language errors toward a more comprehensive understanding of L2 writing.

#### 1.2.2.5 Extensive reading and writing development

Elley (1998) reports findings related to the beneficial effects of 'book floods' on pupils' motivation towards the L2, as well as their performance in it. The 'book flood' projects were conducted mainly on the islands of the South Pacific and in South Africa. The reason for starting reading projects in the third world was the harsh contradiction between the expectations, according to which children should become literate in a L2, and the resources to achieve this aim. Most countries (Fiji, Niue, Sri Lanka) were characterised not only by a lack of resources and well-qualified teachers, but children also had insufficient exposure to the target language. Their motivation to learn a L2 was also low, as they already had a L1 to express their needs.

Reading projects in Singapore, Sri Lanka and South Africa proved that an enriched book reading programme with interesting, well-illustrated books had a substantial effect on pupils' language development. The Spanish Book Flood with Free Reading (Schön, Hopkins &

Davies, 1982 cited in Elley, 1998, p. 12) and the Extensive Reading (ER) by Pakistani Pupils (Hafiz & Tudor, 1989 cited in Elley, 1998, p. 23) had similar gains by independent silent reading. In Sri Lanka, South Africa and Pakistan writing test results showed that the pupils in READ programmes were able to compose more coherent texts than their control group counterparts. These test results are encouraging, as they support the hypothesis that ER develops not only pupils' reading skills, but also their writing skills. Most studies report on additional positive attitudinal changes towards English relative to control group.

In Hungary there had been no large-scale FL reading programmes until the 1990s except for the one conducted by the English Teaching Resource Centre, Pécs, in primary schools of Baranya county between 1995-98 (Bors, 1999). The most important aims of the Baranya Reading Project were to develop both teachers' and pupils' attitudes towards extensive reading and also to improve students' language competence. According to teachers, the benefits gained from participating in the project were numerous. Among them they mentioned students' language improvement in all skills, as well as the gains in their general knowledge and the growth of pupils' self-esteem.

As the aims of the Baranya Reading Project were to develop attitudes towards L2, on the one hand, and enhance language competences on the other hand, I (Bors, 1999) included a 'writing element' in the project: I suggested students compose a reading diary. In my survey, I wanted to find out how pupils reacted to this task, so my main focus was students' motivation to write, not the linguistic accuracy of the products. Consequently, the quality of writing was not assessed. The answers of the participating teachers' (N=16) questionnaires reflected diverse reactions: five teachers said that for some pupils writing was a new motivating element, but other children found it to be extremely time-consuming and were discouraged by it. Two answers claimed that for most of the children diaries meant an interesting challenge. Another two teachers referred to the difficulty perceived by pupils who, nevertheless, accepted that the evaluation would be more objective by the scoring of diaries. One teacher's children rejected the idea of writing; two educators said that some of their students even stopped reading when they heard about diaries.

Let us now see the reasons for participating pupils' (N=80) acceptance or refusal of writing a diary. 31 percent of the children indicated that they preferred reading one more book instead of spending time on writing. In their replies 9 percent of the children said that they wrote diaries, because it was compulsory and they were assessed on doing it, whereas teachers did not mention the compulsory feature of this task. Eleven percent of the pupils wrote diaries for their teacher's suggestion, while two children mentioned that they worked

for a good mark, and one was eager to get more scores in the competition. Five children appreciated the benefits of re-reading and writing, as it resulted in their ability to recall the story better. Two children found that they understood the story better if they compiled reading diaries, while four pupils said they did enjoy writing.

These replies suggest that writing is not a favourite activity with most children. One reason is that they are not taught how to do it effectively, as the main aim of primary FL education is to develop oral skills. Focusing on this aim results in teachers' preference to keep writing periods low in class and assign written exercises for home. These tasks require no creativity, thus children, participating in the BRP, faced a new challenge when they were invited to produce a reading diary. Nevertheless, the ones who agreed to compile diaries were reported to have improved their writing skills.

Summarising what research tells us on the benefits of extensive reading programmes, we must admit that there are limitations to research projects, as there are so many variables involved that it is difficult to draw conclusions from the findings. Most of these programmes obtained confirmation of their hypothesis concerning the positive effects of ER, though Williams (1999, p. 41) warns us about the various concomitant changes in the South African READ project which made its evaluation extremely difficult.

On the issue of the development of writing through reading, research informs us that good writers have done a certain amount of pleasure reading, though, reading alone is not sufficient to make an efficient writer (Kiszely, 2003; Smith, 1983). Projects conducted in the area of ER have reported mostly positive results regarding students' attitudes towards the L2 and the growth in their language proficiency.

In the previous sections I addressed the nature of writing as opposed to speaking and reading; then, I further explored the relationships between L1 and L2 literacy. First, the relationship between reading in L1 and L2 was presented through the Interdependence Hypothesis (Cummins, 1981). Next, the influence of reading in L1 on writing in L1 was investigated, and the links between reading and writing in L2 were introduced by summarizing the results of ER projects. Now, let us investigate how linguists approach writing; what factors they consider to be crucial in this mental process and how they analyze the product of the process. These theories provide a frame for the narrower area of investigation of the present study.

### 1.2.3 Approaches to researching writing

The various purposes of writing, the complexity of its contexts of use and the diverse backgrounds of those wishing to learn it, all necessarily widen the range of possible frameworks of analysis. Hyland (2002, p. 5) identifies three main approaches to writing, though he admits that this classification implies no rigid divisions. These approaches are based on theories concerned with texts, readers or writers respectively, similarly to the views on literacy that I surveyed in the first section. Leki (2002, p. 60) also argues that "L2 writing instruction and research have gradually broadened their perspective by shifting focus from texts, to processes (i.e., composing), to disciplinary and socio-political contexts (i.e., social construction)." These recent developments indicate that researchers' interest has shifted from the product to the process and the writer of the text.

#### 1.2.3.1 Text-oriented approach

The first approach investigates the products of writing by examining texts in various ways, either focusing on linguistic elements or on rhetorical resources available to writers. The dominant model for years considered writing as a textual product, which can be analyzed independently of particular contexts, writers or readers (Shannon & Weaver, 1963 cited in Hyland 2002, p. 7). The ideas behind this view originated from Chomsky's *Transformational Grammar* (1957). According to this paradigm, texts are orderly arrangements of words, clauses and sentences. By following morphological and syntactic rules and correctly arranging these elements writers can transfer ideas mechanically to the reader. It implies that writing is decontextualised, as meanings can be encoded by the writer and the same meanings recovered by anyone with the right decoding skills. This view of writing is still alive, as in many schools writing is mainly conducted to demonstrate the knowledge of decontextualised facts with little awareness of the reader. In these contexts the main criteria of good writing are: factual display and clear exposition. Hyland (2002) claims that from this perspective learners' compositions are seen as a demonstration of their knowledge of forms and rules of creating texts.

Guided composition is the main teaching method corresponding to this belief. As for assessment, indirect measures, typically multiple choice tasks, cloze tests, gap-filling, or error recognition are considered reliable measures of writing skills (Alderson, 2000). This focus on form generated considerable interest in identifying microstructural characteristics of texts, and the immense development in computer technology has supported the analysis of large learner corpora, for example, temporal frequency expressed in writing (Kennedy, 1987 cited in

Hyland, 2002, p. 7). From this viewpoint, writing improvement can be measured by counting gross increases in certain features, such as number of words, clauses, and sentences. This study will also rely on corpus analysis to obtain objective data on learners' L2 performance. Concentrating exclusively on formal features of texts as a measure of writing competence, however, ignores how texts are the writer's response to a particular communicative setting. Social views on the writing process emphasize the latter feature.

#### 1.2.3.2 Social and cultural approaches

Writing is not a product of solely cognitive efforts but is a meaning-making activity which is socially purposeful. As Hayes (1996, p. 5) states, "What we write, how we write, and who we write to is shaped by social convention and by our history of social interaction … The genres in which we write were invented by other writers and the phrases we write often reflect phrases earlier writers have written." The genre of dissertation can easily be related to this claim, as it also has its conventions that researchers need to learn and apply if they want to join the academic discourse community.

In the early 1970s a new field of linguistic research emerged in order to find ways to analyze texts beyond sentence level, that is, to interpret texts as discourse. Discourse refers to language as use, and to the purposes and functions linguistic forms serve in texts. According to this view, writers have certain goals and intentions, certain relationship to their readers and the linguistic forms of the text are means to accomplish their social aims. An important early contribution to this model was the work of the Prague School (e.g., Firbas, 1986), examining clause structures, that were signing the writer's assumptions on what is known and what is new to the reader. This functional sentence perspective was further developed by Halliday (1994). The rhetorical functions of particular discourse units were investigated earlier by Winter (1977) and Hoey (1983). They show that readers, relying on their knowledge of text patterns, can infer the semantic connections between clauses, sentences or paragraphs.

These approaches reflect the idea that the coherence of writing is created in the readers' interpretive processes. Scripts or schemata (Schank & Abelson, 1977), cognitively organized conventionalized knowledge, help interpreting discourse by analogy with earlier experiences, and impose a coherent frame on a message.

A more pragmatic than cognitive approach (Grice, 1975) proposes that writers try to create optimally relevant texts; thus, successful communication is based on interactants' mutual assumptions of rationality and cooperation. Kramsch (1997), drawing on current

theories of discourse analysis, argues that the construction of meaning from texts is a rhetorical process as well as a cognitive one.

Cultural aspects of writing have seen numerous debates from the 70s through the 90s. Kaplan's (1988) method of contrastive rhetoric, in which he analyzed learners' written products and called attention to distinctive variation in the written discourse of students from different cultures, has been subject to a number of criticisms (Brown, 1994; Leki, 1992). Nevertheless, recently it has regained respect, as it has become evident for researchers that many aspects of writing are rooted in culture. Investigation into contrastive rhetoric has shown that cultural expectations can have a consequence for the coherence of texts, as coherence is not an inherent feature of the text, it rather results from the writer's right or wrong judgment of what the reader will be able to infer from the text. As readers bring their own expectations and background knowledge to reading (Hoey, 2001), misreadings are possible. The role of reader expectations has important implications for the scoring of writing tasks and will be dealt with later in the study.

In sum, social and cultural paradigms place the reader into the focus of attention and examine how writers engage with an audience in producing coherent texts. The main proposition of the social approach suggests that the social setting plays an important part in the construction and interpretation of texts.

After surveying text-oriented and social perspectives on writing I turn to overview the cognitive view of the writing process.

#### 1.2.3.3 Cognitive approach

This paradigm focuses on the writer and the writing process by describing writing in terms of the psycholinguistic processes used to create texts. The main positions in the process writing movement focus on (1) the personal creativity of the individual writer, (2) the cognitive processes of writing, and (3) the writer's immediate context (Hyland, 2002).

In order to capture the differences between novice and expert writers and to describe the constituents of the writing process several models of writing have been developed. Though models of complex cognitive activities can never be completely accurate or proven, they call our attention to various factors that influence the process. Many of these models originally focused on L1 writing, and they are not specifically concerned with the development of language proficiency as expressed through writing, but assume a stable language system. They focus on the development of cognitive and metacognitive strategies that are involved in generating a coherent text for specific audiences and purposes. Nevertheless, research has

shown that L1 and L2 writing share a lot of common features (Krapels, 1990; Silva, 1993), so it is worth getting acquainted with them. I will briefly review the most influential models here to exemplify the richness and multidimensional character of this research area, though this study will restrict itself to a small section of the possible areas of investigation.

A pioneering model considering cognitive factors was developed by Hayes and Flower (1980). Hayes and Flower (1980) depict the writing process as a problem-solving activity in terms of the task environment, that is, the writing assignment itself and the text produced by the student. It looks at students' background knowledge and long-term memory, including knowledge of topic and audience, and a number of cognitive processes, such as planning, translating thought into text, and revising. Flower and Hayes (1981) note that good writers have a great deal of tacit knowledge of formal features of reader-based prose, thus they can rely on more options to organize their thoughts and express their ideas. Novice writers, on the other hand, spend less time thinking about the reader, they focus on the topic. One important insight of the model is that writing is not a linear, but a recursive process, whereby writers discover and reformulate their ideas in order to approximate meaning (Zamel, 1983), which has implications for the teaching of writing.

Sixteen years after the birth of the Hayes and Flower (1980) writing theory Hayes (1996) designed a new model, which investigates the task environment, but it places more emphasis on the individual. He thinks that individual aspects of writing involve interactions among four components: motivation and affect, working memory, cognitive processes, and long-term memory. Hayes conceptualizes working memory as being composed of three parts (1) phonological memory, which stores auditory information, (2) the visual-spatial sketchpad embraces visually coded information, and (3) semantic memory stores conceptual input. Cognitive processes in this model involve text interpretation, reflection, and text production. Hayes also emphasizes the importance of reading in the writing process, which includes reflective reading of own text and reading source texts and instructions. With respect to second language writing assessment, a shortcoming of the model seems to be the lack of attention paid to L2 competence.

Grabe and Kaplan (1996) cater for this unattended aspect in their model developed in the same year. Their view of language knowledge, building on the works of Hymes (1972), Canale and Swain (1980) and Bachman (1990), divides language knowledge into three areas: linguistic, discourse and sociolinguistic. Linguistic knowledge comprises the knowledge of basic structural elements of the language, while discourse knowledge supports the

construction of cohesive texts, and sociolinguistic knowledge helps in using language appropriately in a variety of social settings.

Bereiter and Scardamalia (1987) proposed an influential two-model description of writing which I will summarize without going into fine details. The two versions of the model, named knowledge-telling and knowledge-transforming parts, relate to novice and advanced writers, respectively. Here I will focus on the knowledge-telling part of the model describing novice writers' mental processes. As interactive elements related to speech are missing from the individual writing process, the writer, while generating content in the absence of a partner, needs to rely on three sources: (1) the topic (or task), and (2) his own discourse schema, that is, the knowledge of the elements to be included and their arrangement. The third source is (3) the writer's own text written so far, which can be used for generating further ideas. Thus, in Bereiter and Scardamalia's (1987) view, knowledge telling follows the straight-ahead form of speech production and does not require much planning or goal setting. In my view, relying on the third source, the writer's own text, involves reflection and possible reorganization of ideas, which makes it different from speech production.

The other part of this model, knowledge transformation, involves more effort and skill, as it expects the writer to create new knowledge by completing problem analysis and problem solving as well. The model proposed by Bereiter and Scardamalia (1987) calls attention to differences between skilled and unskilled writers, pointing out the substantial disparity between the writing strategies they are able to use. It also offers valuable insights both for writing pedagogy and assessment in relation to task difficulty and genre familiarity. Writing tasks that are familiar to students and can be accomplished through a knowledge-telling process may be attainable for beginners, but may not distinguish between better and poorer writers. On the other hand, if the task requires students to indulge in an unknown genre, otherwise good achievers may not perform well. This underperformance is due to the extra cognitive effort made to meet challenges concerning content and rhetorical issues.

We have seen that theoreticians constructed several models aiming to describe the products and complex processes of writing, while placing emphasis on different aspects of this multidimensional activity. Although these theories emerged in relation to first language literacy, many of their claims can be linked with second language learning. In the next section I will survey the relationship between writing in L1 and L2, and address specific difficulties met by second language writers.

### 1.2.4 Writing in L1 and L2 and specific issues of writing in a second language

When examining L2 writings, researchers' first dilemma is to what extent overall second language proficiency influences the quality of writing in the second language? Another question concerns transfer of writing proficiency from first to second languages, that is, whether the same cognitive processes are responsible for constructing L1 and L2 texts.

Regarding the transfer of writing ability from L1 to L2, researchers (Cummins, 1981; Cumming, 1989; Krapels, 1990; Kroll, 1990) agree that, given a certain level of language proficiency, expertise in writing can be transferred from the L1 to the L2. The reason for this is that L2 students rely on similar cognitive processes to those of L1 writers. Nevertheless, L2 learners usually face language difficulties while engaged in writing, they have a restricted syntactic and stylistic repertoire, as well as limited range of accessible lexis that can be used in writing. In sum, Silva (1993) points out that writing in an L2 is more constrained and less effective than writing in the L1. He found that L2 learners plan less, revise for content less, and write less fluently and accurately than they do in their L1.

The origin of these deficiencies lies in poorer language competence: while L2 learners need to focus on language issues trying to keep words and structures in their working memory, their higher-order cognitive processes responsible for generating content and organizing ideas cannot be involved. To put it in another way, learners cannot rely on meta-cognition as long as they struggle with basic cognitive operations, as the capacity of their working memory is limited (Gathercole & Baddeley, 1993).

A Japanese study (Kobayashi & Rinnert, 2002) inquiring into high-school students' L1 writing instruction aimed to help university L2 English writing teachers to understand their students' needs. The results call into question the common assumption that Japanese high-school students receive little training in L1 writing. The authors emphasize that a more in-depth analysis of how students' L1 literacy background, especially essay-writing experience affects L2 writing in terms of text construction and composing processes would provide useful information for L2 writing specialists at universities. By drawing on students' strengths in terms of their literacy background, teachers can help them make connections between their L1 and L2 writing.

According to Magnuczné (2003), Hungarian university students rely on knowledge telling rather than knowledge transforming (Bereiter & Scardamalia, 1987) in their English essays. She compared 30 Hungarian and 30 USA students' argumentative writings and found that Hungarian students were not aware of the rhetorical requirements of argumentation. The knowledge-telling approach the Hungarian students relied on reflected their previous essay-

writing experiences both in their mother tongue and L2 essays during secondary education. In Hungarian secondary schools the main aim of writing is the demonstration of learnt knowledge as opposed to using argumentative strategies. Magnuczné's findings illustrate that writing in an L2 is not only a question of language competence, but strategies for analyzing and organizing content also need to be taught, as well as genre characteristics.

Szalai (2003) underlines the importance of considering students' personalities in EFL writing classes. She describes the phases of process writing, experimental, and cooperative writing, as main methods to make writing the instrument of thinking, creativity and self-reflection.

There are few inquiries into the language proficiency levels of the same students in their L1 and L2. Kiszely (2003) conducted an innovative research investigating 30 Hungarian secondary school students' L1 and L2 argumentative essays by applying topical and discourse structure analysis. The research study was complemented with a standardized essay assessment procedure and it also scrutinized students' L1 and L2 writing processes and literacy backgrounds. Kiszely found that L1 writings were of better quality, students used more arguments to prove their points in their mother tongue than in the L2. Relying on discourse structure analysis and topical structure analysis, he claims that discourse structures are transferable, but coherence patterns are not. He supposes that different threshold levels exist for the transfer of separate textual features: for discourse structures a lower level of L2 proficiency is sufficient, whereas for coherence patterns a higher level is necessary. Kiszely also points out that the results of the process approach to writing have not become an integral part of secondary school teaching practice in Hungary.

In tertiary education, however, a process writing syllabus was initiated and exploited by Horváth (2001), who demonstrates how the development of writing skills is achieved at university level in Hungary. He describes the process syllabus designed at the English Department of Janus Pannonius University (JPU) aimed at enhancing students' essay writing skills. The classroom tasks and techniques, as well as the out-of-class activities were designed to raise students' awareness of constituents of successful writing. Both teacher and peer feedback contributed to the improvement of students' texts collected in their personal portfolios.

These examples from Hungarian research results prove that limitations in the meta-cognitive processes while composing in a L2 and the lack of sufficient training leading to successful writing are challenges in L2 writing. The application of a process syllabus, on the other hand, can contribute to writing development.

Besides the lack of training in good writing and problems occurring in metacognitive processes, another area of difficulty for L2 learners may be text interpretation and text generation. These challenges are described among cognitive processes in Hayes's (1996) model. Poor language proficiency on the perception side will reduce performance in writing; thus, in reading tasks the misunderstanding of the source text will lead to erroneous writing performance, and evaluating one's own writing is also problematic if reading skills are underdeveloped. Krashen (1984), Elley (1998), and Bors (1999) report the beneficial effects of ER on pupils' motivation towards the L2, as well as their language performance, including writing skills.

In addition to limited language competence, second language writers are also handicapped by not being aware of appropriate ways of expressing functions in writing. Also, expectations of readers in a different culture may not be understood by L2 learners.

The success in L2 writing, as well as in L2 learning in general, is determined by the learners' motivation to acquire the new language and the extent of their desire to integrate into the new culture (Gardner & Lambert, 1972; Peirce, 1995; Shen, 1988; Schumann, 1978). However, for FL learners, studying a L2 in a classroom setting, the integration into the culture of the target language may not be a motivating force. In institutional learning environments grades, impressing the teacher or peers, and future jobs seem to be more realistic goals for L2 acquisition.

An additional issue in L2 writing is the amount of time allocated. Time constraints pose a relatively important problem for L2 writers, as searching for linguistic devices in an L2 is more time consuming than in the mother tongue.

This study mainly focuses on the linguistic and discourse features (vocabulary, accuracy, and text cohesion) of EFL texts produced by primary-school students and explores the writing strategies of a sub-sample both in linguistic and metacognitive domains. The investigation of the process of writing is not one of the focal points of this study; nevertheless, the review of psycholinguistic perspectives will support the qualitative analysis of students' writing.

In this section I have demonstrated that, on the one hand, the basic language skills are interrelated, as they rely on similar linguistic resources, but on the other hand, they also differ in triggering distinct mental processes. Concerning the influence of L1 literacy on that of L2, research informs us that the basic underlying mental processes for comprehending texts are universal. The Interdependence Hypothesis (Cummins, 1981) claims that literacy skills acquired in the mother tongue can be transferred to a second language in case the learner has attained a threshold level of L2 competence.

In the next section I will overview how researchers tried to gain information on L2 students' language competence and the phases of its development. Although most of these inquiries rely on oral data collection, their findings are relevant to this study. The results of studies on learner language have important pedagogical implications; by better understanding the second language acquisition process we can apply the findings in language teaching: syllabus design, materials development, task design, and testing.

## *1.3 Studies on the development of learner language*

### 1.3.1 Investigating learner language

For many researchers the goal of the study of SLA is the description and explanation of L2 learners' competence and its development. Definitions of competence vary, but according to all linguists it comprises the underlying systems of linguistic knowledge (Canale & Swain, 1980). The question arises how we can get information about this mostly hidden entity. Linguistic knowledge implies both implicit and explicit elements, implicit ones being mostly formulaic chunks and the unconscious knowledge of abstract patterns and rules. Explicit knowledge is conscious and declarative; it implies awareness of the schemas existing in implicit knowledge and the metalanguage that can verbalize this analyzed knowledge (Bialystok, 1982; Ellis, 2004; Krashen, 1981). Although many linguists agree that linguistic knowledge comprising competence is basically implicit, and the main goal of SLA is to account for this implicit knowledge, they also realise that the investigation of this area cannot be direct. Thus, examining learners' performance remains the main source of information to make assumptions about their L2 competence and how language acquisition takes place.

Consequently, the study of the characteristics of learner language - language performance - is one of the focal points in researchers' investigations. Four aspects of learner language have been scrutinised: (1) errors, (2) developmental sequences and acquisition orders, (3) variability, and (4) pragmatic features.

Besides language performance, the other area of research investigates external factors of language acquisition related to the context of learning, language input and interaction with peers which the learner experiences. The third area concerns internal mental mechanism in the learner, which are used in communication. They relate to the transfer of knowledge from the learner's mother tongue; the universal processes involved in converting input into intake and restructuring existing L2 knowledge systems; the use of linguistic universals; and the

processes for using L2 knowledge in performance. The study of external and internal factors aims to explain how L2 language acquisition takes place.

After the short enumeration of universal constituents contributing to L2 learning the question of individual differences and the reasons that cause them also need to be identified. Learners differ regarding their motivation and aptitude and also in their use of various learning strategies. The study of these factors helps us to understand why some students learn more rapidly than others and why they reach higher level of proficiency. The four areas of investigation (1) language performance, (2) external factors in language acquisition, (3) internal mechanisms used in language acquisition, and (4) individual differences of language learners interrelate; therefore, a lot of research papers have inquired into more than one of these areas. These focal points of interest reflect researchers' basic views on learning, that is, whether they are concerned about the social influence on language learning or the psycholinguistic processes of the learner, or they are ready to acknowledge both.

In the next section I will select from the abundance of methods in which learner language has been investigated and summarize the approaches and techniques of those relevant to my research. I will discuss the main methods focusing on language performance: contrastive analysis (CA), error analysis (EA), and obligatory occasion analysis (OOA).

### 1.3.2 Types of linguistic analyses of learner language

#### 1.3.2.1 Contrastive analysis and error analysis

Error analysis (EA) is the earliest method that was created to analyze learners' errors; its roots date back to the 18$^{th}$ century. It consists of a set of procedures for identifying, describing and explaining errors that learners make in their speech or writing. Corder (1967) claims that learner errors serve three important purposes: pedagogic, research and learning purposes. The pedagogic purpose shows teachers what learners already know and what they have not acquired. Research purpose is aided by the evidence learner errors provide on the process of learning, and the learning purpose is supported by feedback on errors by which they can discover rules of the target language.

EA as a research tool was promoted in the 1960s as an alternative approach to contrastive analysis (CA) (Brooks, 1960) for understanding language learning. CA, built on the behaviourist view of learning, suggested that L2 learning took place in the same way as any other kind of learning, that is, by habit formation. Habits are formed when learners respond to stimuli in the environment and subsequently have their responses reinforced (Skinner, 1957). Habit formation calls for over-learning, which ensures that learner responses are automatic.

So, already learned habits interfere with the learning of new ones by inhibiting the process of acquisition. To provide a cultural example: when you are offered something in Hungarian, and you want to refuse politely, you say: Thank you, no. It is difficult to change this word order into the English one: No, thank you. Thus, CA supposes that L2 learners face a hard task of overcoming the interference of L1.

Whereas CA investigated only the learner's native language and the target language, EA provided a methodology for scrutinising learner language. Based on their research results, error analysists refuted CA assumptions. Error analysists emphasised that the study of learner errors indicated that errors were only partly caused by transferring L1 'habits'. Other errors reflected learners' creative contribution to the process of language learning. EA also revealed that learners went through stages of acquisition, as errors reflected certain levels of development (Bailey, Madden & Krashen, 1974; Dulay & Burt, 1974; Krashen, 1981; Krashen, Butler, Birnbaum & Robertson, 1978; Larsen-Freeman, 1976).

EA became associated with nativist views of language learning and the emergence of interlanguage (IL) theory (Selinker, 1972). Whereas behaviourism assumed the influence of environmental stimuli, nativist theories emphasize the mental processes that take place in the learner's mind. According to this paradigm, linguistic input is computed internally by an innate cognitive faculty (Chomsky, 1965), called the language acquisition device (LAD), resulting in a knowledge system that is reflected in the learner's actual performance. The cognitive mechanisms are responsible for both 'noticing' language features and processing them as L2 knowledge, that is, the learner's IL.

Selinker (1972) used the term 'interlanguage' first to refer to the mental grammar that a learner constructs at a specific stage of the language acquisition process. He originally defined it as "a separate linguistic system based on the observable output which results from a learner's attempted projection of a TL (target language) norm" (Selinker, 1972, p. 214). IL theory has evolved considerably since its emergence, but the main premises on L2 acquisition have remained intact. Relying on Ellis's (1997, p. 33) summary, the following assertions characterize a learner's IL:

- It consists primarily of implicit linguistic knowledge.
- It constitutes a system in the same sense as the native speakers' grammar is a system. This system of rules is viewed as a 'mental grammar'.
- It is permeable and transitional, that is, it is open to change. New linguistic forms are constructed either externally from input or internally through processes such as over-

generalization, omission or transfer. This ever-restructured language forms an IL continuum reflecting stages of development.

- It is variable; at any stage of development the learner will employ different forms for the same grammatical structure. This variability may be partly random (free variation), but it is largely systematic, suggesting that learners are likely to have competing rules at any stage of development. However, other researchers claim (Gregg, 1990) that IL systems are homogeneous, and variability in learner language reflects mistakes, so it is related to performance only, language competence is not variable at any stage.
- It is the product of general learning strategies, such as L1 transfer and intralingual ones, like simplification.
- It may be supplemented by means of communication strategies to compensate for gaps in L2 knowledge.
- It may fossilize (it fails to achieve full native speaker grammar).

The concept of IL offers a general account of L2 acquisition implying elements from both mentalist theories of linguistics (language acquisition device) and cognitive psychology (learning strategies). Besides this prevailing psycholinguistic perspective on IL there are others which are concerned with social factors of language acquisition. One of these theories (Tarone, 1988) states that IL comprises different styles to be used in different conditions of language use, while another social view (Peirce, 1995) considers learners' social identities that are negotiated in their interactions.

Returning to the psycholinguistic approach described above in detail, CA is credited for compiling IL premises which were supported by the results of error analysis. An example for variability from Ellis' study (1984) illustrates how J, a Portuguese boy, varied in his use of pre-verbal negation, applying *me no cut, not finished, she don't* in the same speech sample. This example shows that variation involves the usage of not only one deviant target-language form, but several different deviant forms.

The starting point for EA is the clear definition of an error. This in not an easy task, as a decision needs to be made if grammaticality or acceptability would be a criterion. If grammaticality is selected, an error can be defined as a 'breach of the rule of the code' (Corder, 1971). Acceptability depends on the context of language use, so it is more dependent on subjective stylistic considerations. Steps in conducting error analysis involve: (1)

collection of sample of learner language, (2) identification of errors, (3) description of errors, (4) explanation of errors, and (5) error evaluation.

Explanation of errors is a delicate area of investigation, as 'many errors are likely to be explicable in terms of multiple rather than single sources. Thus, it is, perhaps not surprising that researchers have produced different estimations of the percentage of errors that can be traced to interlingual and intralingual sources' (Ellis & Barkhuizen, 2005, p. 66). The limitation of EA is that it focuses on the errors that learners make without considering learners' correct language usage. To get a fuller picture of learner performance, obligatory occasion analysis investigates both correct and incorrect forms of learner language.

#### 1.3.2.2 Obligatory occasion analysis

##### 1.3.2.2.1 Morpheme studies

Performance analysis, the analysis of learners' L2 production examines samples of learner language in their totality. Brown (1973) describes obligatory occasion analysis (OOA) as a method that inquires about the accuracy of specific linguistic features performed by language learners. Similarly to EA, it compares forms used by language learners with target language norms (standard native speaker variety). OOA was first elaborated to investigate L1 acquisition by developing a means of specifying to what extent learners have acquired a certain linguistic feature. It also established a method for comparing the extent to which different features have been acquired. Brown (1973), in a longitudinal study, scrutinized the accuracy of a number of English grammatical morphemes, such as present progressive *–ing,* plural *–s*, and regular past *–ed,* performed by three children. The procedure was then adopted by second language acquisition researchers both in longitudinal (Hakuta, 1974) and cross-sectional studies (Dulay & Burt, 1973; Larsen-Freeman, 1976). These inquiries, known as the morpheme studies, aimed to learn whether there was a universal order of morpheme acquisition for all learners irrespective of variables as their mother tongue, age or setting. They also wanted to find out if this order was the same or different from the acquisitional order of L1.

##### 1.3.2.2.2 Developmental stages and acquisition sequences found in morpheme studies

The morpheme studies were important in deciding whether behaviourist or nativist views of language learning would win the theoretical battle. If it could be demonstrated that L2 learners followed a universal order of acquisition, then nativist thinking would be proved right claiming that learning was internally driven, irrespective of the external differences. If, in

addition, evidence was found for the same acquisitional order for L1 and L2 acquisition, it would prove the existence of the Language Acquisition Device (Chomsky, 1965), which governs the route of development for all humans. Conversely, if it was evidenced that learners followed different orders of acquisition, it would suggest that learning was environmentally directed as proposed by behaviourist theories. However, if differences were found between L1 and L2 acquisition orders, interpretation would be difficult, as L2 learners meet the second language in different environments and at a different age from L1 learners, thus it is hard to establish which variable is responsible for the disparity.

The main assumption in the morpheme studies is that the accuracy of a grammatical morpheme indicates the extent to which it has been acquired. In the cross-sectional studies, the order of accuracy is equated with the order of acquisition. The order of acquisition is established by ranking morphemes that have been investigated regarding their accuracy. One of the problems with rank ordering is that it fails to acknowledge differences in degrees of accuracy, but it can be solved by grouping morphemes with accuracy scores that are close (Dulay, Burt, & Krashen, 1982). Another important question is at what accuracy level a particular morpheme can be regarded acquired. Traditionally (Brown, 1973), a morpheme is considered 'acquired' if a learner achieves the accuracy score of 90 percent or higher. In other studies (McDaniel, McKee & Cairns, 1998) the cut-off point is 75 percent correct use in obligatory contexts.

However, the level of acquisition of a morpheme can be miscalculated, as some morphemes go through a U-shaped development: first, they are produced correctly, then, due to overgeneralization, they are replaced by incorrect forms. Finally, the correct form is acquired (e.g., came, comed, came). To avoid misjudgments on learners' acquisition of morphemes, researchers need to ensure that learners are grouped according to general L2 proficiency, and accuracy orders defined separately for each group.

Cross-sectional and longitudinal studies of L2 learners carried out in the 1970s produced conflicting results. Whereas cross-sectional ones (Bailey, Madden & Krashen, 1974; Larsen-Freeman, 1976) indicated that there was a regular order of accuracy for English grammatical morphemes, longitudinal ones produced more mixed results. In cross-sectional investigations acquisition hierarchy was not influenced by learners' mother tongue, age or setting. On the other hand, it was affected by the way of data collection. For example written and oral samples in separate research studies did not reinforce each other (Krashen, Butler, Birnbaum & Robertson., 1978; Larsen-Freeman, 1976).

The morpheme studies found both similarities and differences between L1 and L2 acquisition orders. For example, Brown (1973) and Dulay and Burt (1973) both reported that progressive *-ing* was the most accurately used morpheme and that 3rd person –s of present tense was one of the least accurately used ones. On the other hand, these studies also found differences in the use of articles, the copula and the auxiliary 'be', which were performed more accurately by L2 learners. The results of the morpheme studies were interpreted as lending support to a nativist view of L2 acquisition. According to Dulay, Burt, and Krashen (1982), Brown's (1973, p. 105) conclusion for L1 acquisition can be extended to L2 acquisition: children work out rules based on the speech they hear, passing from levels of lesser to greater complexity. Dulay, Burt and Krashen (1982) claim that this rule also applies to adults.

The results of the morpheme studies contributed to the retreat of behaviourist theories, as the same acquisition order of morphemes irrespective of age, L1 or learning context could not be explained by habit formation. The existence of universal order indicated that learners relied on internal mechanism in language acquisition.

Although the methodology utilized in the early morpheme studies received harsh criticism (Long & Sato, 1984), its basic findings proved right. A limitation of OOA is that it fails to inform us whether learners can use the acquired morphemes correctly in the context, that is, if they also know the function of the morpheme (Long & Sato, 1984). It is possible to address this issue by means of target-like use analysis, which takes account of the functional overuse of a morpheme (Lightbown, 1983; Pica, 1984). Larsen-Freeman and Long (1991), summarising results of numerous studies, noted that despite admitted limitations in some areas, the morpheme studies provide strong evidence that ILs exhibit common accuracy/acquisition orders.

#### 1.3.2.2.3 Order of morpheme acquisition revisited

While the morpheme studies phased out in the 1980s, interest in the natural order did not. Researchers turned their attention to the explanation of the order of acquisition, and they wanted to find factors accounting for the accuracy order. Since the time of the first morpheme studies in the 1970s, more than fifty L2 morpheme studies have been reported, using data from a variety of L1 backgrounds and analysis procedures (Ellis, 1994; Larsen-Freeman & Long, 1991).

An interesting research study conducted in Hungary (Nikolov & Krashen, 1997) compared children who studied EFL with a content-based approach with similar children who studied

English with a form-based traditional approach. The pupils of the experimental group were slightly more accurate in their production of grammatical morphemes in an oral interview, and were more fluent, confirming that communication-based approaches do not sacrifice accuracy for fluency.

Two decades after the boom of morpheme studies Goldschneider and DeKeyser (2001) re-examined the accuracy results of oral production data gained in twelve cross-sectional studies. They investigated the predictive power of five variables (frequency, phonological salience, semantic complexity, syntactic category and morphophonological regularity) for accuracy. Goldschneider and DeKeyser (2001) suggest that the salience of a morpheme plays the most important role in its acquisition. Thus, no single variable can account for accuracy, but rather a number of variables contribute to the differential salience of individual morphemes. Goldschneider and DeKeyser's (2001) meta-analysis suggests that L2 acquisition is the product of both input and the learner's internal mechanisms.

Their view is more compatible with connectionist models that posit a general and minimalist cognitive structure tuned to notice and process input features (Ellis, 2002) than with Chomsky's nativist view of language learning, which states the existence of an elaborated LAD that directs learners to seek evidence from input.

Lee and Huang (2004) investigated a specific area of IL of young Chinese ESL learners. They examined 270 Hong Kong primary-school children's variable use of 'be' in a story-writing task for developmental patterns of English language acquisition versus L1 transfer phenomena. The results did not suggest a strong role for L1 transfer, but tended towards developmental aspects of omission and overgeneralisation. Lee and Huang found that the sequential order in the acquisition of the copula 'be' and auxiliary 'be' by Chinese ESL beginners was similar to the order established in earlier morpheme studies (Brown, 1973; Hakuta, 1974), that is, the copula is acquired earlier. A further search into the children's course books revealed that the learning environment and conditions for learning (textbook input and transfer of training) could play an equally significant role in the IL 'be' system of Chinese ESL pupils.

The role of input was also found an important factor in the morpheme acquisition order in Muñoz's study (2006, p. 114). She carried out a research in FL setting seeing that "the morpheme studies have been taken as evidence that there is not a qualitative difference in the way in which L2 learning progresses after a certain maturational point" in acquisition-rich environment, regardless of instruction. Muñoz explored whether different-aged learners in an FL situation also proceed in similar ways in morphological development, and whether there

are age-related differences in the rate of morpheme acquisition, as reflected in the accuracy with which those morphemes are used. Muñoz took oral interviews of six groups of students participating in the Barcelona Age Factor (BAF) project.

She found that the learning context does not affect accuracy orders if learners have had a certain amount of exposure to the target language and have progressed beyond the very elementary levels of proficiency. As performance on morphemes improves with increased exposure to the target language, a commonality of ordering emerges, though it is not invariant. The study highlights the role of input, as one of the determinant factors of morpheme ordering was the frequency of their occurrence (Goldschneider & DeKeyser, 2001) Muñoz's study (2006) also provides evidence of older learners' superior rate of learning to use morphological features accurately, especially in initial stages of language acquisition.

The role of input, that is exposure to language, is also raised by Demuth (in McDaniel, McKee & Cairns, 1998, p. 19), while summarising the methodology of collecting spontaneous speech production data from children. Whereas these spontaneous data can provide positive evidence for the presence of a grammatical construction, they offer limited use in determining whether the absence of a particular language form is due to lack of linguistic ability, lack of exposure to the construction, or lack of appropriate discourse contexts in the sample. Although Demuth's (in McDaniel, McKee & Cairns, 1998) observation is related to speech production, the first two concerns can be raised in connection with writing as well.

A relatively new area of investigating machine-readable learner language samples is corpus linguistics (Leech, 1997). McEnery, Xiaio and Tono (2006) aimed to test the results of the early morpheme studies by exploiting a commercial learner corpus called the Longman Learners' Corpus. They examined how Japanese learners acquire grammatical morphemes such as the ones investigated by Dulay and Burt (1973), and other researchers in the 1970s. McEnery, Xiaio and Tono's (2006) study considerably verified early studies' claims by using large learner-corpus data.

In this section I have overviewed the types of linguistic analysis of learner language that have relevance to this study. Then, I summarized the findings of the morpheme studies carried out in the 80s (Dulay, Burt, & Krashen, 1982) and the results of the re-examined investigation of the original data 20 years later (Goldschneider & DeKeyser, 2001).

In the next section I will address the research methodology concerning L2 acquisition and performance placing emphasis on the written products of second language students.

## *1.4 Research methodology of writing studies*

### 1.4.1 Overview of research methodology: a mixed approach

The study of second language acquisition (SLA) draws on insights of research from a number of disciplines, comprising linguistics, psychology, sociology, psycholinguistics, sociolinguistics and education. This multidisciplinary aspect of SLA is reflected in the numerous surveys published on the field (Ellis 1985 and 1994; Gass & Selinker, 1994; Towell and Hawkins, 1994). Researchers have enquired into diverse aspects of language acquisition: they have examined learners' L2 performance by collecting and analysing samples of learner language, reports of learners' introspections or their intuitions about appropriate L2 behaviour.

My empirical study relies on both quantitative and qualitative research methods. The use of multiple research techniques contributes to the credibility of the investigation (Mackey & Gass, 2005, p. 164) "One reason for the persistence of the distinction between quantitative and qualitative research is that the two approaches represent different ways of thinking about and understanding the world around us" (Nunan, 1995, p. 10). While quantitative research is based on the positivistic notion that the main function of research is to uncover facts and truths that are independent of the researcher, qualitative researchers do not believe in the notion of objective reality.

The truth seems to be unfolding by seeing the world from multiple viewpoints. Researchers (S. D. Sieber, 1973; Jick, 1979 cited in Creswell, 2003, p. 15) recognized that all methods have limitations, and tried to neutralise the biases of one single method by applying different ones. This triangulation of data seeks convergence across quantitative and qualitative approaches and provides information from multiple perspectives. From the original concept of triangulation emerged additional reasons for mixing different types of data, e.g., (1) the results of one method can inform the other method, or (2) one method can be placed within another method to provide insight into different levels of analysis (Creswell, 2003). Within mixed methods of inquiry researchers can shape the procedures to best suit their research aims. I have chosen concurrent procedures to combine quantitative and qualitative data in order to provide a comprehensive analysis of my research results.

### 1.4.2 Research on second language writing

During the past thirty years the amount of empirical research on L2 writing has been increasing, and the approaches to doing it are extremely varied due to the range of backgrounds and the interdisciplinary nature of the field (Matsuda, 1998). The objects of inquiry include writers' texts, writing processes, attitudes, and the social context among others (Polio, 2003). Hedgcock (2005, p. 610), summarising research and pedagogy on L2 writing, claims that "no unitary theory or model of L2 writing has yet emerged as a foundation on which to build a coherent disciplinary identity". He argues that an intensified reflexivity is needed between research and pedagogy in the 21st century, and L2 writing professionals will be required to pursue multiple types of inquiry. They will need to call on the new findings in allied disciplines, for example, in second language acquisition, computational linguistics, cognitive and educational psychology.

Leki (2002, p. 61) argues that while for most users outside academic circles L2 writing may be limited to functions such as writing short notes, the majority of published research on L2 writing has dealt with extended writing in professional settings. The researchers aimed at finding out how best to teach L2 writing, but she claims that "This research question, however, is premature since, before teaching L2 writing, it would seem necessary to understand and characterize good writing, even to specify what it means to be a good writer."

Jarvis, Grant, Bikowski and Ferris (2003) agree that a satisfactory description of L2 writing is problematic. They argue that even when variables such as proficiency, language background, topic, and audience have been controlled, straightforward predictive relationships between linguistic variables and quality ratings have remained elusive, and perhaps they always will. Jarvis et al. explore multiple profiles of highly rated timed compositions and describe how they compare concerning their lexical, grammatical, and discourse features. They investigated the use of 21 linguistic features; among them text length, mean word length, diversity of vocabulary (type/token ratio), conjuncts, and hedges. Within their two data sets, the profiles of highly rated texts differed significantly. Some profiles exhibited above-average levels for several linguistic features, whereas others showed below-average levels.

Jarvis et al. (2003) suggest that complementarity is a characteristic feature of texts; a number of linguistic features that contribute to the overall quality of a written text may bring about low levels of other features. Compensation is the second dimension of the multiple-profiles phenomenon. It is a strategy by which successful writers may be able to compensate for potential deficiencies in their writing by relying on a few of their strengths. The results of

the study suggest that the quality of a written text may depend less on the use of individual linguistic features than on how these features are used in tandem.

The few investigations on children's ESL writing report on controversial findings. Whereas Hudelson (1988) found that texts produced by young ESL writers were similar to those of young native speakers, she admitted that L1 culture had an influence on the functions and purposes of writing. Maguire and Graves (2001, p. 588), on the other hand, claim that school-age L2 learners use different genres, modalities, semantic and syntactic structures.

Other researchers who studied the writing development of young learners state that it takes years of persistent an attentive teaching for L2 pupils to achieve sufficient language control to produce effective academic writing similar to that of L1 children of the same age (Valdes & Sanders, 1999).

As this study is concerned with primary-school students' texts, I will focus on this aspect of inquiry. Most of the experimental studies in this area investigate students at a higher proficiency level (Leki, 2002) than that of the pupils in this research; thus, few of these inquiries can be related to my study. What makes them interesting, though, is the variety of techniques and measures used to study either the overall quality of the text (Engber, 1995) or a specific feature within the writing. Examples for areas of investigations are the studies on linguistic accuracy (Hamp-Lyons & Henning, 1991; Lee & Huang, 2004), the lexicon (Laufer & Nation, 1995; Lehmann, 2006; Nation, 2005; Schmitt, 2000), content (Hedgcock & Lefkowitz, 1992), and coherence (Kiszely, 2003; Magnuczné, 2003; Reynolds, 1995).

As regards measures, I refer back to the previous section on qualitative and quantitative research methods stating that the choice of the best measure is not straightforward. Researchers who want to quantify the quality of a text may use holistic measures, which assign one overall score to a script (Engber, 1995). Analytic scales help with assessing various aspects of texts (Wolfe-Quintero, Inagaki, & Kim, 1998 cited in Polio, 2003, p. 42). Assessment issues will be discussed in more detail in Chapter 2.

In the next part I aim to introduce a few studies investigating FL learners' written productions that can be related to the present study in either their focus, or the age group, or the research methodology applied.

Recently, the debate on the age factor in acquiring a L2 has enlivened, some researchers inquired into the influence of starting age on the development of language fluency and accuracy. As EFL situations had rarely been included in earlier research projects in this area, the following studies set out to explore language development in FL acquisition settings.

Lasagabaster and Doiz (2003) investigated 62 Spanish-Basque bilingual students of three different age groups: 11-12-year-olds (N=31), 15-16-year-olds (N=18) and 17-18-year-olds (N=13) with a similar amount of time exposure to English as a FL. They applied several approaches of evaluation: (1) raters evaluated texts on general impression, then, (2) quantitative analysis of writings was carried out concerning fluency, complexity and accuracy measures, finally, (3) students' errors were described qualitatively.

Lasagabaster and Doiz (2003) argue that the nature of the assignment (letter writing) entailed the absence of specific types of errors that may have been frequent in other tasks. For example, there were hardly any negative or interrogative sentences in the students' letters, but this was the consequence of the task rather than the lack of students' familiarity with these structures.

Raters found that the older the students, the more developed communicative ability is displayed in their texts in all language areas: content, organisation, vocabulary, use of language and mechanics. The quantitative analysis led the authors to the same conclusion, more competent students produced longer, more complex, and more accurate texts. The qualitative examination of errors showed different trends (1) younger students made more basic errors (spelling mistakes, omissions of verb), (2) the oldest age group committed more omissions of the infinitive particle 'to' and also semantic malformations, while (3) the middle group had most mistakes concerning missing articles and verb tenses.

Lasagabaster and Doiz (2003) explain these different trends of errors by the existence or lack of linguistic competence: the youngest group can not make mistakes in areas they do not know yet, whereas the intermediate group has poorer linguistic competence than the oldest group, so they are liable to commit errors in areas in which older students are more competent and do not make mistakes.

Trying to find reasons for the results, they argue that little attention is paid to the development of the writing skills until the beginning of secondary education (from the age of 13-14 years). They point out that the higher writing competence of older students may be due to the longer exposure to formal education. The authors argue that the age factor cannot be isolated from other factors that interact with it, such as the level of competence achieved in the L1 and L2, and several affective factors.

Torras, Naves, Celaya and Pérez-Dival (2006) point out that the issue of age and IL development in instructed learners, measured through writing, has not yet received enough attention. Thus, they investigated 495 EFL learners from the Barcelona Age Factor (BAF, Muñoz, 2006) project distributed in six groups according to age of onset and hours of

instruction over a wide time frame, from 8 to 17 years of age. The same amount of instruction hours was completed over different periods of time; early starters received 726 hours of instruction in nine academic years, while late starters had the same amount of lessons in seven school years. Students' written production was measured on a composition task, and it was analyzed in four areas of competence: fluency, lexical complexity, grammatical complexity, and accuracy.

Torras et al. (2006) concluded that an early start at the age of 8 does not result in a higher level of attainment at the age of 16, after 726 hours of instruction. They see the reason of this in the reduced number of language lessons after the age of 12 for the early starters. They found differences in the distribution of gains that affect the rate of acquisition of language variables in the different age spans. Some fluency variables (number of words/clauses/sentences) and two lexical variables (noun and verb types) increase steadily from the beginning of instruction in both groups. By contrast, the age 12 appeared a turning point for some variables within grammatical complexity (e.g., clauses per sentence, number of subordinate clauses).

Another important conclusion of this study is that neither the areas of language nor the variables included in them develop in tandem, whereas their rate of development seems to be affected by age. Torras et al. (2006) also conclude that benefits of an early start in some fluency aspects and lexis may have a delayed effect in the long run, after a more extensive period of instruction than the one in this study.

One interesting point in the conclusion includes the misleading nature of some of the measures used in the study. The authors found that the number of sentences is misleading evidence of the development of writing proficiency, since differences do lie in the internal complexity of sentences, such as coordination and subordination. The measure of words per sentence also reflected complexity rather than fluency. Torres et al. (2006) suggest the phrase/clause ratio which would allow investigation of their constituents.

The pedagogical implications of the study are that the early start provides gains in fluency, whereas starting at about the age of 12 will result in greater development in complexity of linguistic devices. This age represented a turning point in the acquisition of written competence, due to the simultaneous operation of several factors, such as age, instructional time, or proficiency (e.g., Li, 2000; Torras and Celaya, 2001; Wolfe-Quintero et al., 1998).

The findings from the BAF project are consonant with those obtained in similar research conducted in the Basque Country on the acquisition of EFL by bilingual (Spanish and Basque) children (e.g., Doiz and Lasagabaster, 2004; Sagasta, 2003).

One of the RQs of this study aims to investigate the vocabulary features of students' writings, thus the next section will concentrate on vocabulary measurement issues.

### 1.4.3 Measuring vocabulary features of students' writings

In order to take part in successful communication language learners rely on linguistic, cognitive and social resources. One of the important constituents of linguistic knowledge is vocabulary, defined by McCarthy (1990, p. 1) as the "words of a certain language, where words are seen as freestanding items of a language that have meaning".

Different measures have been used to assess learners' lexical knowledge: c-tests, cloze-tests, vocabulary-scales, among the most frequent methods. The main question of interest concerning vocabulary for L2 learners and teachers is usually what and how extensive vocabulary a language learner needs. To facilitate the decision on vocabulary selection for FLs various word lists have been produced.

One of the oldest ones is *A General Service List of English Words* (GSL) compiled by West (1953). This list contains 2,000 head-words based on frequency patterns drawn from a written corpus of five million words. Classroom implications suggest that for fast results in L2 English learning the 2,000 GSL words should be among the words acquired first. Nevertheless, we will see from the results of this study that even primary-school students know words that belong to base lists 2-4, while they may not know all of the most frequent 2,000.

Linguists have been researching ways of characterising learners' vocabulary size and other features as well. Laufer and Nation (1995) proposed a new measure of lexical richness in writing, called the Lexical Frequency Profile, which was based on the GSL and the University Word List. This objective tool can be useful in measuring vocabulary size, as it shows the relative proportion of words from different frequency levels in any sample of writing. This measure, the type-token ratio (TTR, Nation, 1995), shows the proportion of different words (types) and all words (tokens) used in the text. A similar ratio can be calculated for lexical or content words excluding all function words (articles, auxiliaries, prepositions, pronouns) from counting. This measure is called the type-token ratio of lexical words (TTRL).

According to Read (2000, p. 200), lexical richness has four main components, by which a reader may assess the lexical features of a script:

1) the type-token ratio (lexical variation),
2) lexical sophistication, which shows the ratio of relatively low-frequency words and high-frequency words,
3) lexical density: the proportion of content words and all words,
4) the number of errors in the text, which may be caused by wrong word choice or incorrect form.

A more detailed description of lexical measures will be presented in section 4.7.3.4, and an analysis of sample texts in 4.8.5.

A shortcoming of TTR is its sensitivity to text length (Read, 2000; Richards, 1987; Vermeer, 2004), basically because the rate at which new word-types appear in a text decreases as the text size increases. As TTR decreases in longer texts, the comparison of data becomes less reliable (Read, 2000). It means that TTR can be used reliably for comparing texts of the same or of very similar length. However, unfortunately, students' writings are rarely exactly the same length.

There have been attempts to overcome this problem by constructing variations of TTR, but they were not feasible solutions (Miralpeix, 2006). Fixing the length of all samples to be analyzed could be another solution to this problem, but this would result in data loss, due to cutting the texts.

To compensate for the shortcomings of TTR, a new measure, D, was proposed by Malvern and Richards (1997, 2002). It is claimed to be 'more informative than TTR, because as opposed to the single value of TTR, it represents how TTR varies over a range of token sizes for each speaker or writer' (Miralpeix, 2006). This measure uses all words in the text, there is no need to standardise text length. Its other advantage is that it can be used with lower levels of language proficiency, as the calculation is applied successfully even with a short text containing as few as 50 tokens.

Although it is difficult to predict if the D formula will become a standard measure for lexical richness, it seems to be one of the most reliable indices of lexical diversity at present.

Miralpeix compared TTR and D in her research within the BAF project. She examined the oral and written vocabulary of Catalan/Spanish bilingual early starters and late starters by various methods. She found moderate correlations between D and TTR without keeping length constant, but after standardising text length these correlations became stronger.

Lexical density (LD) has also been used to assess lexical richness of scripts (Ure, 1971). It may be a misleading measure at low levels of proficiency, though, as students at this level often use telegraphic style; they tend to omit function words. In these cases the ratio of content words becomes high suggesting high lexical richness, although, it actually reflects the writer's inability to construct a coherent text (Hyltenstam, 1988).

However promising the D might be, it is not available for the author, so this study will rely on the Range software (Nation, 1995) to examine the lexical features of some of the students' texts and investigate how text length influences TTR.

### *1.5 Summary*

In this chapter I investigated the connection between reading and writing skills in L1 and L2; then, I examined several models on different aspects of writing. Next, I surveyed the relationship between writing in L1 and L2 focusing on specific difficulties they pose for second language writers.

Then, studies investigating learner language were introduced, as well as the methodology of exploring this transitional and variable phenomenon, so as to provide the theoretical background of my research project. Reviewing the research methodology of writing studies illustrated that mixed methods proved to be most efficient, as by using them results of one method can inform the other, and researchers can choose the procedures that best suit their aims.

Research on second language writing of young learners was reviewed to see what questions interest researchers in the international arena of EFL writing that can be related to this study. The review of different instruments measuring vocabulary features has helped me choose the best research methods to analyze students' texts.

The next chapter will discuss assessment issues of writing.

# Chapter 2: Assessing writing

## *2.1 Basic considerations in assessing writing*

In the previous chapter I have overviewed the nature of writing ability as opposed to other language abilities, explored the main approaches to writing, and introduced studies on the development of writing skills focusing on learner language. In this chapter I will turn my attention to the assessment of writing and look at writing tests as a specific type of language test. First, test purposes are considered, then, the relationship between language performance and abilities are discussed. Next, types of scoring procedures are introduced and concerns about the reliability of rating are shared with the reader.

### 2.1.1 Test purpose

As we cannot observe language ability directly, we use performance data to infer the underlying abilities. When designing a writing test, the first question to be considered is what we are going to use the test for. Bachman and Palmer (1996) emphasise two main purposes: (1) to make inferences on learners' language ability and (2) to make decisions based on these inferences.

The assessment of writing can focus either on the product of learners' writing activity or the complex process of creating this product. We have seen several models aiming to describe the complex psycholinguistic processes of writing (Bereiter & Scardamalia, 1987; Grabe & Kaplan, 1996; Hayes & Flower, 1980), in section 1.2.3.3. As this study concentrates on the products of writing, I will present only a few examples for process-focused methods of teaching and assessment.

The process-oriented literature describes several useful ways of developing learners' writing proficiency. Portfolios, that is, collections of students' work prepared over a period of time, have become a strong component of assessment in elementary education in some countries (Puckett & Black, 2000; Weigle, 2002). Project work or contract of work (Puckett & Black, 2000) is a negotiated task between the student and the teacher. Self- and peer-assessment are also effective strategies to encourage children to be active participants in the assessment process (Scarino, Vale, McKay, & Clark, 1988).

Assessment should encourage and motivate learners. Swain (1985) describes tasks and assessment procedures that 'bias for best'; tasks are appropriate and motivating and give some indication of success. Large-scale external tests for young learners can motivate by using two

or three shield rewards rather than pass-fail results (McKay, 2006). A good example for this is the Cambridge Young Learner English Test (Taylor & Saville, 2002).

### 2.1.2 Language ability test performance

First of all, we need to clarify what we mean by language ability; what the constituents of this complex phenomenon are in real-world language use and in a language test. Language ability is viewed as a construct, and test designers have to define which parts of the construct are essential to what they want to measure, and which are not. Thus, for each testing situation, a definition of ability, or construct, of interest must be developed. This specific construct will consider test takers, the purpose of the test, and the target language use situation.

According to applied linguists (Bachman & Palmer, 1996; Douglas, 2000; McNamara, 1996), communicative language ability consists of interactions between aspects of language knowledge, on the one hand, and strategic competence, on the other. As opposed to Grabe and Kaplan's (1996) model of writing, a more general taxonomy of components of language ability (Bachman & Palmer, 1996; Douglas, 2000) comprises

(1) grammatical knowledge, or the knowledge of the building blocks of language,

(2) textual knowledge, or the knowledge of constructing coherent texts from these blocks,

(3) functional knowledge, or the knowledge of how to get through messages by applying communicative functions, and

(4) sociolinguistic knowledge, or the ability to use appropriate language in various social ontexts.

Strategic competence is a set of metacognitive components, or strategies, which provide cognitive management function in language use. They can be viewed as higher order executive processes (Bachman & Palmer, 1996, p.70). Strategic competence provides the link between one's language knowledge and the external situation, and also between the language knowledge and other individual characteristics. Bachman and Palmer (1996) see strategic competence as having three components: goal setting, assessment, and planning. Thus, it is not a language-specific ability, but rather a problem-solving one. On the other hand, when we think of strategic competence regarding writing, we will find strategies specifically characteristic of this area, e.g., the process of reflection (Hayes, 1996), and rhetorical decision-making (Bereiter & Scardamalia, 1987).

In addition to language knowledge and strategic competence, other factors also influence learners' language performance. Topical knowledge caters for the content of the message, as without it even the most excellent language knowledge may remain unveiled. Personality and

emotional factors may also enhance or hinder performance. Topical knowledge may not be part of the test construct, depending on the purpose of the test.

This study is mostly concerned with linguistic knowledge, which corresponds to grammatical knowledge and textual knowledge in Bachman and Palmer's (1996) taxonomy. Lexis must be part of the grammatical knowledge category, as it can be viewed as a building block of language (Lewis, 1993).

### 2.1.3 Test usefulness

Issues on test usefulness will be dealt with in short, as the main focus of this paper is not test construction, but the analysis of test results and the comparison of different assessing methods. Test usefulness is defined by six qualities: reliability, construct validity, authenticity, interactiveness, impact, and practicality. As all these qualities cannot be maximized, the test developer needs to determine an appropriate balance among them.

Reliability is defined as consistency of measurement across different characteristics of the testing situation, such as different prompts and different raters (Weigle, 2002). A test is reliable if learners receive the same score from one prompt or rater to the next. Also, if a group of test takers is rank ordered in the same way on different versions of the test, or by different raters, the test is said to be reliable (Weigle, 2002). *Cronbach's alpha* is a commonly known and applied statistic for interval-level data that responds to the consistency of observers when numerical judgments are rendered to a set of units. It is called a reliability coefficient, but according to Hayes and Krippendorff (2007), it does not measure concrete agreement. Instead, it quantifies the consistency by which assessors judge units on an interval scale without being sensitive to how much the observers actually agree in their judgments. Hayes and Krippendorff (2007) pinpoint that without defining scale points with valid reliability interpretations it is unsuitable to assess reliability of judgments. They state that Cronbach's alpha "… is appropriate as a measure of the reliability of an aggregate measure across observers, such as the arithmetic mean judgment, but it does not directly index the extent to which observers actually agree in their judgments" (Hayes & Krippendorff, 2007, p. 5).

After describing the criteria for a good measure of reliability, they propose Krippendorff's alpha as the standard reliability measure. They claim that it can be used regardless of the number of observers, levels of measurement, sample sizes, and presence or absence of missing data. In this study both Cronbach's and Krippendorff's alpha will be used.

Construct validity will be discussed in connection with research methodology (see 4.4.2). It is related to "the meaningfulness and appropriateness of the interpretations that we make on the basis of the test scores" (Bachman & Palmer, 1996, p. 21). Construct validation is the process of determining whether a test is actually measuring what it is intended to measure.

## *2.2 Scoring procedures*

A score in a writing assessment is the outcome of an interaction between the test taker, the task, the written text, the rater and the rating scale (Hamp-Lyons, 1990; McNamara, 1996). Two of the aforementioned categories represent central considerations in scoring: compiling the rating scale and ensuring that raters use the scale appropriately and consistently. According to McNamara (1996), the scale that is used in assessing performance tasks such as writing tests represents the test developer's notion of what skills or abilities are being measured by the test. For this reason the development of a scale and the descriptors for each level are of crucial importance for the validity of the assessment.

### 2.2.1 Types of rating scales

When determining a system for scoring, the first decision should be made about the type of the rating scale, that is, should a single score be given to each student writing, or should several features of the text be scored. There are four types of rating scales used in writing assessment (see Hamp-Lyons, 1990), but I will describe only the two most characteristic ones, holistic and analytic scales, as I aim to rely on those in this study.

#### 2.2.1.1 Holistic scoring

Holistic scoring assigns a single score to a script based on the rater's overall impression of the writing. Raters work from a scoring rubric that outlines the scoring criteria. They are trained to adhere to the rubric when assessing scripts. The predecessor of holistic rating was called general impression marking, which did not rely on explicitly stated criteria. A well-known example of a holistic scoring rubric in ESL is the scale for the Test of English as a foreign language (TOEFL) writing test.

One of the advantages of holistic scoring is that it is fast, raters do not need to reread scripts several times for different aspects of writing. It also reflects the reader's authentic, personal reaction to a text, thus, in White's view (1984, p. 409), it is more valid than analytic

scoring. One drawback of holistic scoring, however, is that one single score does not allow raters to distinguish between various aspects of writing such as content, richness of vocabulary, control of syntax, and organization. It often conceals a lack of consensus on writing quality (Barkaoui, 2007). Thus, holistic rating does not provide diagnostic information about the person's writing ability. It is not easy to interpret either, as the details are blurred under the holistic judgment.

#### 2.2.1.2 Analytic scoring

Analytic scoring schemes are preferred by many writing specialists, as they provide more detailed information on a test taker's performance in different aspects of writing (Alderson, 1991; Hamp-Lyons, 1990; Weir, 1990). Hamp-Lyons (1990) describes the difficulties of establishing the "true" score related to a paper when assessors highly disagree. She admits that raters can not consistently agree with each other on the score awarded to the same script, moreover, sometimes we make different judgments about the same writing on different occasions. I will support this view in the empirical section of the dissertation.

The main advantage of analytic rating over holistic schemes is that it provides more useful and detailed diagnostic information about students' writing abilities. It is especially useful in L2 education where students tend to show an uneven profile across different aspects of writing. For example, a script may be well organized, but display poor vocabulary used inappropriately, or it excels in richness of vocabulary, and has numerous grammatical errors, but the message is clear. Hamp-Lyons (1991) suggests that an analytic scoring scheme tends to improve reliability, as multiple scores are given to each script.

The main disadvantage of analytic scoring is that it is time-consuming, since raters are required to make more than one decision concerning each piece of writing; thus, they are to read scripts several times. Some analytic scales have a large number of categories in order to profile achievement. Such scales have been argued to be less appropriate for assessment, as raters tend to find it difficult to cope with more than four-five categories.

### 2.2.2 The rating worry

The process of rating written language performance is still not well understood, despite a body of work exploring it over the last 20 years (e.g., Bukta, 2007; Cumming, 1990; Milanovic *et al*., 1996; Vaughan, 1991). In recent years linguistically more informed research into rating and rater behaviour has emerged, particularly in second language writing assessment. For example, Cumming, Kantor, and Powers (2002) looked closely at the

decision-making process that experienced raters of essays use, including asking them to enumerate elements that characterize effective writing. Cumming et al. also used think-aloud protocol to shed light on raters' internal criteria. They found that in addition to a focus on content and rhetoric (e.g., assessing task completion), and a language focus (e.g., considering error frequency, lexis, syntax), raters also self-monitored, for example by rereading the essay, or comparing it with other essays.

Lumley (2002) investigated the process by which raters of ESL learners' texts make their scoring decisions using an analytic rating scale designed for multiple test forms. Four trained and experienced raters provided think-aloud protocols describing their rating process. Data show that although raters follow a similar process, the relationship between scale contents and text quality remains obscure. Lumley observed that raters first formed uniquely complex impressions of texts, independently of the scale wordings. Then, they tried to reconcile their impression of the script, the specific features of the text, and the wordings of the rating scale, thereby producing a set of scores. While they tried to ease the tension between the rules and the intuitive impression, they went through an indeterminate process.

> Because this judgement is so complex, so multi-faceted, we can never really be sure which of the multitude of influences raters have relied on in making their judgements, or how they arbitrated between conflicting components of the scale. Likewise, we have no basis for evaluating the judgement that would have been made if a different scale were used. (Lumley, 2002, p. 24)

Lumley claims that rating can succeed in yielding consistent scores if raters are trained, and supported with guidelines to deal with problems.

When educators need to decide what type of rating scale they would use, their decisions depend on the aims of assessment. There has been little empirical research, however, on how different rating scales affect L2 rating processes, and how raters perceive them. Barkaoui (2007) employed a mixed-method approach to investigate the effects of two different rating scales on EFL essay scores, rating processes, and raters' perceptions. Four EFL teachers in Tunisia assessed a set of EFL essays silently and two subsets of four essays while thinking aloud using a holistic scale and then a multiple-trait rating scale.

The holistic scale resulted in higher inter-rater agreement, although raters employed similar processes with both rating scales. They used essay-comparison and self-generated criteria relatively more frequently while assessing texts holistically. These "supplementary criteria," however, seem to have led to a higher inter-rater agreement. These findings contradict Hamp-Lyons's (1991) claim that an analytic scoring scheme tends to improve reliability.

In Barkaoui's (2007, p. 103) study raters were the major source of variability in terms of the scores assigned and the frequencies of the decision-making strategies used, particularly for multiple-trait rating. This variability is surprising, particularly for the three raters who had been teaching the same writing course and rating essays together for a few years, and, thus, were expected to have developed a relatively common set of criteria that they could have used with the new scales. As the raters differ in many respects (e.g., teaching and rating experience, academic background), it is difficult to determine which rater variables affected the scores they assigned. Barkaoui suggests that the interaction between raters and rating scales is another area for further research.

Hamp-Lyons (2007) summarizes concerns about the reliability of rating and states that two solutions have most often been proposed and implemented: the combination of direct testing of writing with standardized tests; and improved rater training. She raises the issue of automated scoring as a possible solution for consistent assessment. She wonders whether automation (Ericsson & Haswell, 2006; Whithaus, 2005 cited in Hamp-Lyons, 2007, p. 6) is removing a 'problem' with human scoring, or washing out the richness of multiple interpretations. She claims in accordance with Conference on College Composition and Communication that direct assessment in the classroom should provide response that serves formative purposes, helping students develop and shape ideas, as well as organize and edit texts.

> Automated assessment programs do *not* respond as human readers. While they may promise consistency, they distort the very nature of writing as a complex and context-rich interaction between people. They simplify writing in ways that can mislead writers to focus more on structure and grammar than on what they are saying by using a given structure and style.

(Position Statement on Writing Assessment at the Conference on College Composition and Communication, the College section of the (U.S.) National Council of Teachers of English, 2006, cited in Hamp-Lyons, 2007, p. 9)

Bachman and Palmer (1996) note that the choice of test procedures should suit the testing situation. For research purposes practicality and impact may be of lesser significance, but reliability and construct validity will be central concerns. In my study both analytic an holistic scoring are applied, moreover, raters faced an additional challenge of having to interpret the descriptors related to writing ability in the scales of the *Common European Framework of Reference* (*CEFR*, Council of Europe, 2001). Thus, the next section will overview the aims and the main features of the *CEFR,* and discusses concerns about its comprehensiveness, transparency and coherence.

## *2.3 Writing assessment in the Common European Framework of Reference*

### 2.3.1 Aims and measurement issues

The *Common European Framework of Reference* is a guiding document concerned with language proficiency levels elaborated for teachers and learners of modern languages. It has been developed by members of the teaching profession and "it describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively" (Council of Europe, 2001, p. 1). It is intended to be a comprehensive, transparent and coherent frame of reference for language learning, teaching and assessment, based on a very general view of language use.

The description of language proficiency levels is arranged in a scale divided into six bands. Each band outlines parameters of communicative activities and communicative competence required at the respective level of language proficiency. In considering the vertical dimension of the *CEFR*, we need to keep in mind that any attempt to establish levels of proficiency is to some extent arbitrary, as the case is in any area of knowledge. However, for practical purposes it is desirable to create a scale of defined language levels to segment the learning process.

The scale needs to be comprehensive, that is, the description of proficiency levels has to be clear and relatively simple so as to be accessible to practitioners. On the other hand, if the users of the scale proceed along the scale vertically, they consider only one dimension of knowledge and competences. Nevertheless, progress is not merely moving up a vertical scale, but it also involves horizontally widening knowledge. Taking into account the subskills of communicative competence, it is easy to accept that learners may develop more rapidly in one of the areas, but lag behind in another one. This places them in different bands along the vertical dimension.

I will now deal with measurement issues that were crucial in designing this international document. The main criteria for creating the scale were:

- The points on the scale should be objectively determined based on a theory of measurement (Council of Europe, 2001, p. 21).
- The number of levels on the scale should be adequate to illustrate progression in different language abilities, but it should conform to the capability of people of making reasonable distinctions.

These criteria can only be satisfied by a combination of intuitive, quantitative and qualitative methods. The development of the *CEFR* exploited systematic combination of these three methods. The main weakness of reliance merely on intuition is that the wording of competences at particular levels tends to be subjective; it may be valid in one special context, but fail to measure language proficiency in others. The *CEFR* has gone through a long process of validation relying on quantitative analyses, but its details are beyond the scope of this study. The authors claim that "there is a growing body of evidence to suggest that the criteria … [of measurement] are at least partially fulfilled." (Council of Europe, 2001, p. 22) Nevertheless, leading specialists on testing have pointed out several limitations of the framework (Jones, 2002; Alderson, Figueras, Kuijper, Nold, Takala, & Tardieu, 2004). According to Weir (2005, p. 281),

> though … containing much valuable information on language proficiency and advice for practitioners, in its present form the *CEFR* is not sufficiently comprehensive, coherent or transparent for uncritical use in language testing. First, the descriptor scales take insufficient account of how variation in terms of contextual parameters may affect performances by raising or lowering the actual difficulty level of carrying out the target 'Can-do' statement.

Jones (2002) found that different people tend to understand 'can-do' somewhat differently, which suggests the need to search for greater precision and explicitness in test specification than is currently provided by the document. I will relate this lack of precision and explicitness to the assessors' concerns during the rating process of students' writings in this research study. Weir (2005) also mentions the lack of theory-based validity parameters – a function of the processing involved in carrying out these 'can-do' statements. Failure to comprehensively define the construct to be tested prevents current attempts to use the *CEFR* as the basis for developing comparable test forms within and across languages and levels (Alderson *et al.*, 2004; Huhta *et al.*, 2002; Jones, 2002).

As McKay (2006) points out, the Council of Europe has provided a conceptual framework and guidelines for assessment for all language learners, but there is no direct reference to young language learners (YLLs) in it. Work is being carried out in Europe translating these frameworks into syllabuses and specific assessment frameworks for young learners; Hasselgreen (2005) takes account of adaptation procedures.

By reference to the *CEFR* and the European Language Portfolio (ELP, a self-assessment instrument), Hasselgreen (2005) inquires into how far the special needs of YLLs are being catered for by assessment practices in European schools. She provides examples of the

application of these assessing instruments in Europe with a focus on recent developments in Norway.

The account of one of their projects, the Bergen 'Can-do' project, has demonstrated one way in which the levels and 'Can-do' statements in the *CEFR/ELP* can be adapted for the assessment of YLLs so as to preserve the integrity of the *CEFR* levels, and yet consider the particular characteristics of children and young teenagers.

> To the question of whether descriptors of the type used in *CEFR/ELP* are sufficient in themselves for all description of the ability of YLLs, the answer was 'no'. Teachers have expressed a clear need for material that captures the everyday 'here and now' classroom performance, in terms of language quality, in addition to the more function-focused 'Can-do's, which show long-term rather than immediate progress. (Hasselgreen, 2005, p. 352)

Hasselgreen also concludes that teachers lack training in language assessment, and it is desirable that institutions responsible for initial and in-service training provide courses on language testing and assessment, including learner self-assessment.

Another area of ambiguity concerning *CEFR* levels is the connection between task difficulty and learner performance. An interesting email discussion of *EALTA*-members (European Association for Language Testing and Assessment) sheds light on experts' dilemmas:

> Since the *CEFR* describes behaviour of language users, could we possibly talk about task level? Or should we use the *CEFR* descriptors to describe the language produced by learners when attempting to respond to the input? In other words, is it correct to describe a task as A2, or should we describe the performance because of this task? To make things even more difficult when dealing with Writing, the overall written production scale (*CEFR,* p. 61) was not empirically calibrated and the descriptors are relatively short. In my opinion, describing written performance in such a short sentence is not adequate. (Papageorgiou, 12/10 2005)

Sigott (*EALTA*-members' discussion, 12/10 2005) responds that, actually, Papageorgiou has raised two questions: "(1) Can an A2 prompt elicit performance that is above A2? (2) If so, should we be prepared to rate such performance to be level B1, B2, or higher?"

Clearly, if the answer is yes to question 2, then a new dilemma occurs: is it appropriate, or desirable, to describe the difficulty of tasks related to the *CEFR*? The document itself suggests making distinctions between criterion behaviour (what a task at this level requires in terms of skills and abilities) on the one hand, and how well a learner can handle these skills and abilities, on the other. But the question remains: Should assessors be prepared to rate the response to an A2 task higher than A2? Sigott answers as follows: "If no, then rating somebody's writing performance as A2 means that it is *at least* A2, but possibly beyond. If

yes, then we have to question the wisdom of talking about tasks/prompts for the productive skills in terms of their 'difficulty' with regard to the *CEFR*."

In this study, I will demonstrate that a task aiming to measure language proficiency at A1-A2 level can, in fact, elicit higher level learner performances. In order to feel more comfortable with identifying elusive *CEFR* levels, the next section introduces them, and discusses the descriptors of global orientation of the lowest three proficiency levels, A1-B1, relating them to the issue of comprehensibility.

### 2.3.2 The *Common European Framework of Reference* levels

The classic division of language proficiency into basic, intermediate and advanced levels has been kept in the *CEFR,* renamed as levels of basic, independent and proficient users. These main competence indicators were further divided into two sub-categories at each level, as shown in Table 2.

Table 2: The levels of the *CEFR* (Council of Europe, 2001)

| Levels | | Names of the proficiency levels |
|---|---|---|
| Proficient user | C2 | Mastery |
| | C1 | Effective operational proficiency |
| Independent user | B2 | Vantage |
| | B1 | Threshold |
| Basic user | A2 | Waystage |
| | A1 | Breakthrough |

The authors expect that the wording of the descriptors will develop over time as the experience of institutions with related expertise is incorporated into the description. The common reference points can be presented in different ways for different purposes. There is a global one-page form of the *CEFR* which provides orientation points for teachers and curriculum planners. Most FL teachers in Hungary have become familiar with it by now. This point will be revisited later when I discuss the relationship between the Hungarian *National Core Curriculum* (1995, 2003), the *Frame Curriculum* (2000), and the *CEFR* (Council of Europe, 2001).

#### 2.3.2.1 General description of levels A1-B1

I intend to focus on the three lowest proficiency levels only, levels A1-B1, as the upper levels are beyond the competence of the participants in this study. These levels will be of specific relevance when examining raters' decisions on a sample of students' writings. As McKay (2006, p. 307) puts it, "the first three levels are suitable for use with young FL learners,

though they are skeletal, designed to describe development of beginning learners of all ages." In Table 2 it is shown that level A1 is the lowest level of generative language use. At this point the learners

> can interact in a simple way, ask and answer simple questions about themselves, where they live, people they know, and things they have, initiate and respond to simple statements in areas of immediate need or on very familiar topics, rather than relying purely on a very finite rehearsed, lexically organized repertoire of situation-specific phrases. (Council of Europe, 2001, p. 33)

The descriptors "finite rehearsed, situation-specific phrases" refer to a proficiency level, which is lower than A1, and at this stage real communication cannot take place due to the lack of the ability to use language that would express the learners' own thoughts and feelings.

Level A2 includes the majority of descriptors regarding social functions, such as simple everyday polite forms of greeting and address; handling short social exchanges; talking about work and free time activities; making arrangements; and making and accepting offers. The next level, B1, expects learners to be able to maintain interaction and get across what they want to, in a range of contexts. The other characteristic feature of this level is the ability to cope flexibly with problems of everyday life.

Although the global orientation on proficiency levels serves as a good starting point, both teachers and students need a more detailed overview of the main skills and competences necessary at respective proficiency levels. Such overviews presented in the form of a grid show major categories of language use (e.g., phonological control, grammatical accuracy, and vocabulary range) at each of the six levels. This study is mainly concerned with writing, especially linguistic features reflected in it. Consequently, the detailed descriptors of the following areas are of specific interest: general linguistic range, overall written production, vocabulary, and grammatical accuracy.

#### 2.3.2.2 General linguistic range at levels A1-B1

Regarding general linguistic knowledge, level A1 expects the learner to use a very basic range of simple expressions about their personal needs. Learners at the lower end of A2 are to use basic sentence patterns and communicate with memorised phrases; groups of a few words and formulae about themselves and other people; and they are expected to have a small repertoire of short memorised phrases. The upper end of A2 requires learners to deal with everyday situations with predictable content, but allows for searching for words. At the lower stage of level B1 learners are expected to have enough language to get by, with sufficient vocabulary

to express themselves with some circumlocutions on topics such as family, hobbies and interest, work, travel, and current events. According to the specification, at this stage lexical limitations may cause repetitions and even problems with formulation at time. Speakers belonging to the upper band of B1 are required to describe unpredictable situations; explain the main points in an idea or problem with reasonable precision and express their thoughts on abstract or cultural topics (Council of Europe, 2001, p. 110).

In summary of the previous descriptors of levels A1-B1: the growing expectations regarding language proficiency are related to sentence structures, contexts of actions, topics, and ease of expression of thoughts. At A1 level only simple expressions regarding concrete present needs are expected, while at A2 sentence structures need to be more complicated and clarity and fluency of utterances should also be at a higher level. Though grammatical or sentence structures are not mentioned directly in the description of linguistic knowledge, the expressions concerning 'unpredictable situations' and 'abstract topics' may require using complex structures. The contexts of interactions in which learners can communicate should also grow in number with the increase of abstract topics they can handle.

#### 2.3.2.3 Overall written production at levels A1-B1

The descriptors on this illustrative scale have not been empirically calibrated with the measurement model. They were created by recombining elements of descriptors from other scales (Council of Europe, p. 61). At level A1 learners are required to write simple isolated phrases and sentences. At A2 they are to write a series of simple phrases and sentences linked with simple connectors like 'and', 'but', and 'because'. At B1 they need to be able to write straightforward connected texts on a range of familiar subjects within their field of interest, by linking a series of shorter discrete elements into a linear sequence. This scale focuses on the use of cohesive devices, that is, how learners can produce connected texts.

Kaftandjieva and Takala (2002, p. 113) argue for making the *CEFR* more comprehensive when they state that the construct of language proficiency in writing is not as well defined as in other areas. They claim that 'can-do' statements are not subtly varied between levels A2 and C2 (p. 126). Hawkey and Barker (2004) note that it was necessary to draw upon corpus analysis to identify features of texts that helped discriminate between candidates along the *CEFR* scale, when compiling a common scale for Cambridge ESOL, as they were unable to use functional competence to maintain these distinctions.

Alderson (2002) points out that the *CEFR* contains little information as to the content of any given level other than what is contained in the numerous scales. For example, it is not

easy to determine what sort of written and spoken texts might be appropriate for each level. In relation to response format Alderson, Figueras, Kuijper, Nold, Takala, and Tardieu (2004, p. 10) note that *CEFR* does not contain any information about response format, even though the document is supposed to be a reference point for assessment. In writing, the choice of format will determine whether knowledge telling or knowledge transformation occurs in task completion; these are two very different processing experiences.

#### 2.3.2.4 Vocabulary range at A1-B1 levels

At level A1 learners need a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations. At the lower end of A2 they are expected to have a sufficient vocabulary for (a) the expression of basic communicative needs, and (b) coping with simple survival needs. In the upper band of A2 a sufficient vocabulary is necessary to conduct routine, everyday transactions involving familiar situations and topics. To fulfil the requirements of B1 learners have to show a sufficient vocabulary to express themselves with some circumlocutions on most topics pertinent to their everyday life such as family, hobbies and interests, work, travel, and current events (Council of Europe, p. 112).

After studying the description of vocabulary range, it can be claimed that the *CEFR* provides little assistance in identifying the breadth and depth of productive or receptive lexis that might be needed to operate at the various levels. This general guidance on the learners' lexical resources for productive language use, as Huhta, Luoma, Oscarson, Sajavaara, Takala, and Teasdale (2002, p. 131) point out, does not provide examples of typical vocabulary or structures that would help test compilers or assessors.

#### 2.3.2.5 Grammatical accuracy at A1-B1 levels

Learners at A1, according to the requirements, show limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire. Learners having reached level A2 use simple structures correctly, but still make basic mistakes systematically, for example, they tend to mix up tenses and forget to mark agreement; nevertheless, it is usually clear what they are trying to say. The lower band of B1 expects learners to apply a repertoire of frequently used 'routines' and patterns associated with more predictable situations reasonably accurately. Learners in the upper band of B1 are required to communicate with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. They may commit errors, as long as it is clear what they are trying

to express. Although the descriptors for accuracy provide a few examples to help decision making regarding proficiency levels, they are still not sufficiently concrete.

According to Alderson, Figueras, Kuijper, Nold, Takala, and Tardieu (2004, p. 13), it is important that the *CEFR* is not seen as a prescriptive device but rather a heuristic one, which can be refined by language testers to better meet their needs. In its current form, it exhibits a number of serious limitations such that comparisons based on the scales alone might prove to be misleading, due to the insufficient attention paid in these scales to issues of validity. The *CEFR* as presently constituted does not enable test developers to design comparable tests. In relation to test assessors, the same limitations may cause similar uncertainty regarding ranking learners' proficiency levels.

In response to the need of guidance in relating exams to *CEFR* in a credible manner, a manual (Takala, 2004) was developed to support experts using it. The aim of this manual was to encourage the development of both formal and informal national and international networks of institutions and experts. Thus, it was meant to be not just a technical guidance tool but a rich resource for thoughtful action. The Manual presents four inter-related sets of procedures (familiarisation, specification, standardisation, empirical validation) that users are advised to follow in order to design a linking scheme in terms of self-contained, manageable activities. The activities carried out in all four sets of procedures contribute to the validation process. Hopefully, they contribute to finding the answer to Alderson's (2002) question "How do I know that my Level B1 is your Level B1?"

In this section, besides reviewing issues concerning the limitations of the *CEFR,* I have overviewed a few possible adaptation procedures. I examined the three lower proficiency levels of the referential scales that can be related to the present study, as most learners' written performances fall in the A1-A2 band of the *CEFR*. In the next part I will explore a more objective method for analysing learner language, namely, corpus linguistics.

### *2.4 Corpus-linguistic techniques in analyzing writing*

In this section I aim to describe why and how electronic corpora are useful for linguistic inquiries, then, I will define the term 'corpus', and introduce large-scale projects of learner language based on written samples. Last, I intend to shortly summarize the main methods to aid processing information accessible in a corpus.

Due to the advancement in computer technology that permitted analyses of large amounts of printed or typed text, corpus-based studies have altered the methodologies for analyzing

syntactic, lexical, and collocation features in written English (e.g., Leech, Rayson, & Wilson, 2001; Stubbs, 2001). One important reason for using corpora in linguistic research is to extract linguistic information present in them. Electronic corpora have obvious advantages over their paper-based equivalents. The speed and accuracy of processing large quantities of data is one of them, the avoidance of human bias in analyzing data is the other, which make results more reliable. The scope of application corpora has widened in recent years, and it has contributed to developments in education: syllabus design, teacher education, materials development (Horváth, 2001; Hughes, 1997; Wilson, 1997), and test validation (Alderson, 1996).

In modern linguistics, a corpus can be defined as a body of naturally occurring language, though computer corpora are generally designed with particular purposes in mind, and are often assembled to be representative of some language or text type (Leech, 1992). There is increasing consensus that a corpus is a collection of machine-readable, authentic texts which is sampled to be representative of a particular language or language variety (McEnery, Xiaio, & Tono, 2006, p. 5).

Let me now examine the largest available written learner corpora from different L1 backgrounds representing a range of language proficiency levels. The International Learner Corpus of English (ICLE, Granger, 1998), the Longman Learners' Corpus (LLC), and the Cambridge Learner Corpus (CLC) rely on international data. The ICLE is a corpus of argumentative essays written by advanced learners with different L1 backgrounds. The LLC contains ten million words of students' texts at a range of proficiency levels from 20 different mother tongue backgrounds. It is useful for textbook writers and it is publicly available for commercial purposes. The CLC is an ever-expanding collection sampled from students' Cambridge exam tests from 150 countries. It contains 20 million words, nearly half of which have been coded for errors. Unfortunately, it is not available for public use.

The spectrum is wide concerning learner corpora assembled from students coming from a single L1 background. One of the very first ones was contributed by Hungary. The JPU Learner Corpus contains more than 400,000 words developed throughout 1992-1998 in advanced writing courses at the English Department (Horváth, 2001).

The Chinese Learner English Corpus embraces one million words collected from writings of different age groups at different proficiency levels, the writers ranging from middle school students to English majors (Gui & Yang, 2001). The Japanese EFL Learner Corpus was collected from secondary school students, and was used primarily to study IL errors (Tono, 1999). The Swedish corpus can boast of one million words taken from advanced level

students, while The Polish Learner English Corpus containing 500,000 words is sampled from beginning to highly advanced learners (Lewandowska-Tomaszczyk, 2003).

After reviewing large international and national corpora, I will turn my attention to the ways of processing these huge amounts of linguistic data. When representative data have been collected, they constitute a raw corpus which needs to be processed before it can be used. Thus, written texts need to be rendered machine-readable by keyboarding or scanning, whereas spoken data need to be transcribed. Then, in case the corpus is meant to be used later, corpus mark-up and annotation are necessary.

Corpus mark-up is a system of codes inserted into an electronic document to provide information about the text itself and, also, to govern formatting, printing and other processing (McEnery, Xiaio, & Tono, 2006). This procedure is needed, as the examples of linguistic usage stored in the corpus are taken out of their original context, so their contextual information is lost. This lost knowledge can be regained by using the mark-up scheme. Corpus mark-up provides objectively verifiable information concerning the components of a corpus and the textual structure of each text.

Besides mark-up, annotation is the other essential process of encoding data for further exploitation; it supplies additional, interpretative, linguistic information to the electronic corpus. Corpus annotation, as well as mark-up, adds value to the corpus in that it considerably extends the range of RQs that a corpus can address (McEnery, Xiaio & Tono, 2006, p. 29). Corpus annotation can be achieved fully automatically, by a semi-automatic interaction between a human being and the computer, or manually by analysts.

Corpus annotation can be undertaken at different levels, for example, at the morphological level corpora can be annotated in terms of prefixes, suffixes and stems; at the lexical level they can be annotated for parts of speech (POS tagging), lemmas (lemmatization), and semantic fields (semantic annotation); at the syntactic level corpora can be annotated with syntactic analysis, and so on.

Error tagging is especially relevant to this study, as it is a special type of annotation which is associated with learner corpora and aims to inform language pedagogy. In error tagging, the types of errors occurring in learner corpora receive special codes. Corpora annotated for learner errors can reveal the relative frequency of error types committed by learners of different L1 backgrounds and proficiency levels (Dagneaux, Dennes & Granger, 1998).

As seen above, a number of error-tagged learner corpora are available; ICLE, LLC, and CLC. They are assembled from a variety of L1 backgrounds and are partly tagged for learner

errors. Other learner corpora cover only one L1 background, e.g., Chinese, Japanese, Swedish, and Hungarian.

Error-tagging systems vary from one corpus to the other regarding the types of error codes. Error tagging is a time-consuming task, as it is difficult to develop a rule-based programme to identify errors in learner corpora due to lack of information concerning error patterns and their frequencies (Tono, 2003, p. 804). Nevertheless, a number of attempts have been made to automate this laborious process (Dagneaux, Dennes & Granger, 1998).

In L2 writing assessment, the size and scope of analyses evolved from exclusive reliance on impressionistic ratings of scripts by trained assessors, to computations of frequency rates with which particular syntactic, lexical, and discourse features are applied in learners' scripts (Hinkel, 2005; Lazaraton, 2005; Weigle, 2002).

In summary, electronic learner corpora lend themselves to various linguistic inquiries, which, in turn, help experts to gain exact data on L2 acquisition, thus contributing to developments in education.

In relation to the present study, none of the above described existing error tagging schemes could be utilized, as this study was specially focused on young learners at a relatively low language proficiency level, whereas the well-known learner corpora analyze mostly advanced learners' written performances. On the other hand, some of the corpora were not available for me at the time of tagging my data; consequently, I could not rely on ready-made systems. Although the corpus was relatively large to handle, manual tagging was feasible, the procedures of which will be described in Chapter 4.

The next chapter aims to examine the socio-educational context of Hungarian EFL learners; it overviews literacy requirements in their mother tongue and L2, as well as research results on meeting these achievement targets.

# Chapter 3: Contextualizing the study: The socio-educational context for Hungarian learners' writing in L1 and L2

## *3.1 Literacy requirements in L1 Hungarian: National Core Curriculum and Frame Curriculum*

Educational achievement targets, reflecting radical political and social changes in the early 90s, were modified to answer the challenges of a more open and democratic lifestyle. These changes are also reflected in the *National Core Curriculum* (NCC, 2003) and *Frame Curriculum (*FC*,* 2000).

The Hungarian *National Core Curriculum* (NCC, Nemzeti alaptanterv, 2003) identifies ten different cultural domains, among them one basically concerned with teaching literacy skills is Hungarian Language and Literature (HLL). As the focus of this paper is students' writing, here I will concentrate on this main literacy skill. The requirements of this cultural domain in relation to writing skills prescribe that students in the final year of primary education, that is, 14-15-year-olds should be able to

- present their thoughts in legible, nice handwriting;
- create texts of different genres (descriptive text, narrative from different viewpoints, review, argumentative essay);
- express their thoughts, feelings and opinions from alternative viewpoints;
- prepare reports with the teacher's assistance and individually;
- use their knowledge of grammatical structures and spelling consciously;
- be aware of the necessary register in creative writing (pp. 26-27).

These skills represent an abundance of proficiency areas, separated mostly according to traditional foci of mother tongue teaching: layout of texts, text types, and self-expression in writing, grammar, spelling, and style. The categories used for describing the curricular requirements, though, may seem slightly confusing for the reader, as there are overlapping descriptors of the required competencies. For example, argumentation is mentioned in two places, once among required genres in which students need to produce texts, and then among descriptors of self-expression.

The *Frame Curriculum* (2000), which is supposed to be in harmony with the principles of the *National Core Curriculum* (1995; 2003), describes the requirements of the latter in more detail, so that teachers can translate them into their everyday practices. Though the *NCC* has

been overviewed since the publishing of the *FC* in 2000, the main requirements are still in harmony with the new *NCC* (2003). As an overall description of requirements, the *FC* (2000, p. 140) expects students to acquire reading and writing skills in *HLL* classes that enable them to be successful in their further studies and future job. According to this document, the improvement of mother tongue literacy skills is of crucial importance, as they serve not only the comprehension and production of literary texts, but also the studying of other school subjects and, in parallel, the preparation for lifelong learning. For 7th-graders the *FC* prescribes the following class activities in order to practise and develop written literacy skills:

- forming pupils' own view about issues related to everyday life;
- drawing up a draft individually, using quotations/references;
- creative writing: writing a fictitious report, compiling a newspaper page using appropriate register.

These activities, as seen above, are in accordance with the targets of the *NCC*. For grade 8 the *FC* does not define important additional skills.

## *3.2 Literacy requirements in L2 English*

### 3.2.1 *National Core Curricula* (1995, 2003) and *Frame Curriculum* (2000)

As during the decades prior to the change of regime in 1989 all students were supposed to follow prescribed curricula represented by centrally published teaching materials, discussions concerning new achievement targets were not timely. Such professional debates started only in the 1980s when the idea of a new national curriculum emerged as part of the liberalization of education.

Medgyes and Nikolov (2002, p. 203) see modern languages as "a primary conveyor of innovation in the *NCC"*, keeping the 1995 version of the document in mind. They refer to Enyedi and Medgyes (1998, cited in Medgyes & Nikolov, 2002, p. 203), and Nikolov (1999a, cited ibid.), claiming that the social and political changes after 1989 are most spectacularly represented in the curriculum of modern languages, as Russian ceased to be the main FL accessible in public education, thus, other FLs appeared on offer on a large scale. Medgyes and Nikolov (2002, p. 203) find that Hungary's anticipated accession to the European Union increased the need to speak FLs and adopt European norms. All the documents relating to foreign-language education since the change of regime have been designed accordingly; they have adopted the functional-notional syllabus and have preferred humanistic and communicative principles of education.

While appreciating the new approach to language teaching presented in the *NCC* (1995), we must admit that it also had some drawbacks. One was the lack of requirements for crucial turning points in public education; for graduates from both primary ($8^{th}$-graders) and secondary levels ($12^{th}$-graders). The language area where concrete requirements are communicated is vocabulary, 1,200 words of productive lexicon plus 400 words of receptively known vocabulary are prescribed for $8^{th}$-graders.

As for the *Frame Curriculum* (2000), which was published with the aim of clarifying the requirements of the *NCC* (1995) and offering assistance for schools in producing their local curricula, it assigns achievement targets for each grade from 4-8 of primary education in terms of the four language skills and enumerates the required items of expressing communicative intentions. Regarding writing targets, the *FC* expects students to write simple, structured messages, letters to friends, or texts of about 50 words length conveying factual information. A 50-70-word-long description or account of an event is also called for. The description of language levels A1-C2 of the *CEFR* (Council of Europe, 1996, in Az alapfokú nevelés-oktatás kerettantervei, 2000, p. 199) are demonstrated in the *FC* (2000), but no connection is established between the requirements of the Hungarian document and the Council of Europe's publication. Thus, at that time, most teachers trying to interpret the *FC* (2000) were not able to relate the two messages to each other. The other reason for the difficulty of interpretation was due to the complete novelty of the European document for educators. Since that time, initiated by the Ministry of Education, a 'new', competence-based teaching approach has been gaining ground in all cultural domains, and more and more teachers have been acquainted with effective teaching approaches and classroom techniques such as cooperative learning and methods for catering for individual differences. Foreign language teachers have attended refreshing training courses to study these techniques and get to know European requirements concerning modern languages. As a result, by 2007 most L2 instructors will have heard of the proficiency levels A1-C2 of the *CEFR* (Council of Europe, 2001).

### 3.2.2 Relationship between the Hungarian documents and the *Common European Framework of Reference* (2001)

After the publication of *CEFR,* another version of *NCC* was published in 2003, with a shorter and more reader-friendly description of the requirements for modern languages, identifying proficiency targets for 6th, 8th, 10th, and 12th-graders.

As the requirements of the *NCC* are to be reviewed every three years, a new version of the curriculum was published in 2007 (*Nemzeti alaptanterv, 2007)*, with a slight delay. This document has no significant changes regarding the teaching of FLs, as the main overarching aim of the document, the development of key competences, is in line with the requirements for FLs of the previous *NCC* (2003).

Regarding L2 literacy, the *NCC* (2003) prescribes level A1 of the *CEFR* (Council of Europe, 2001) for students graduating from primary education. In terms of writing skills level A1 requires students to be able to write a short greeting or fill in a simple form. However, a discrepancy can be identified between the writing requirements of the latest version of the *NCC* (2003) and the *FC* (2000) as the *National Core Curriculum* requires level A1 for primary school-leavers and the actual wording of the requirements in the *FC* represent level A2: students should be able to write a private letter, a short message, or a greeting of about 50 words.

This discrepancy may partly be due to the continuous work on these documents over years, which involved the reconciliation of several opposing political and specialist views. The other reason for this disparity may be the desire to accommodate Hungarian requirements to European standards. From the publication of the *NCC* in 1995 eight years passed until its latest version in 2003, during which period radical changes took place in European language policy. The publication of *CEFR* (Council of Europe, 2001) meant a new signpost in standardised teaching and assessment targets for modern languages, and Hungarian language policy was eagerly attuned to these new proposals. The limitations of writing requirements (Kaftandjieva and Takala, 2002) published in the *CEFR* were elaborated on earlier in this study.

Speaking of *CEFR* (Council of Europe, 2001, p. 99) requirements, *writing a private letter or a short message* covers a higher level (A2) target, whereas *expressing opinion in an argumentative essay* belongs to proficiency level B2 according to this document. These proficiency levels will be of interest for us when the details of the current research, for example, students' tasks in English are described. Also, the difficulties in interpretation of descriptors by expert teachers will be analyzed.

After reviewing curricular achievement targets in the official documents (*NCC,* 1995, 2003; *FCC,* 2000) let us now examine what research tells us about what Hungarian students can do.

## *3.3 Meeting the achievement targets*

### 3.3.1 Research on Hungarian students' writing skills in L1

In this section I will overview research conducted in the past four decades on primary schools students' Hungarian language competence. As most of the studies to be overviewed examine both primary and secondary education, attention will be focused on primary school students' data and, where necessary, mention will be made of language proficiency of students in secondary education.

Orosz (1972), in a national representative survey, examined essay-techniques of four age-groups (11, 14, 16 and 18) using the following text types: description, narrative, and characterization. In order to study Hungarian students' composing skills objectively he categorized the operations of essay-writing techniques into 18 groups and developed measuring instruments for these subskills. Unfortunately, the abundance of areas (18) of language competence to be assessed according to his suggestion prevented teachers from using this method in their everyday practice. The other drawback of his measuring instrument seems to be the negative scoring of skills. Students' errors, following the error analysis tradition, were counted in each subskill and then compared to the average length of compositions using a relatively difficult system. His findings revealed significant improvement from 11- to 18-year-olds only in three areas of essay writing: material organisation, sentence structuring, and referencing. Orosz concluded that, in general, the results of the survey on students' quality of writing were discouraging - reflecting a negative view of language development.

The first attempts to compare Hungarian students' levels of achievement to those of other countries date back to the early 1980s. Kádárné (1990) describes the Hungarian results of an international study on students' essay writing skills, which was conducted by the International Association for the Evaluation of Educational Achievement (IEA) in 14 countries on three continents. IEA had studied students' achievements in several school subjects from as early as 1962. Essay writing got into the focus of interest of IEA as an essay was hypothesized to be the product of complex cognitive and linguistic processes, representing both cognitive and communicative skills. Thus, to research these processes, an investigation integrating various disciplines had to be designed. The survey was conducted individually by each participating

country, whereas the instruments and logistic instructions were prepared in cooperation. Fourteen different writing tasks designed for the survey covered dominant communicative functions (emotive, referential, conative, phatic, metalingual) and cognitive processing operations (reproducing, organising, inventing).

The three examined age groups (10-, 14-, 18-year-olds) involved 8th-graders, who were in the final year of their primary education in 1983, when the study was carried out. Each student was expected to write three compositions out of the eight writing tasks devised for this age group: different letter-writing tasks, an argumentative essay, a reflective essay and a personal narrative. Each essay was rated by two judges: first holistically, then along an analytic scale containing six criteria: content, organisation, style, accuracy, spelling and layout. Two judges had to reach consensus in rating, so they continuously met to compare their scores. So as to ensure high reliability, the raters were provided with several sample essays previously judged by an international jury. The sample compositions represented three different levels of writing along a five-point scale. Due to conscientious test design the reliability of scoring was very high, exceeding the internationally accepted 0.75 Cronbach's alpha with all raters, reaching an average of 0.84 for the holistic impression scores.

As for the results, the representative study found that the differences between the average scores of students living in the capital (3.52), big towns (2.92) or villages (2.52) were small. However, if the average data are transformed into percentages, we see that students in the capital reached 70 percent, those living in big towns achieved 58 percent, while village-dwellers performed 50 percent. In my view, these differences cannot be called small. The differences were greater between schools within the capital city and the same applied to big towns.

As the Hungarian study was part of an international survey, Kádárné (1990) compared Hungarian students' achievements with those of English and Dutch participants'. She found that English students' results were slightly higher, while their Dutch counterparts' scores were lower, the results being diverse in different types of schools in both countries.

The results of the IEA study concern us in two areas: the letter-writing task and the argumentative essay, as these tasks were similar to the tasks assigned for the participants of this research in their mother tongue. Kádárné reported that the lowest scores were awarded for the argumentative writing task for both 14- and 18-year-olds, while letter writing proved to be an easier task. Most 8th-graders performed best on one of the letter-types, the letter of excuse.

The next study focusing on students' development in writing from primary schools to secondary education was part of a large-scale research on a representative sample, carried out

in Szeged in 1999. Molnár (2002) conducted a survey on composition to explore the improvement of composing skills between the ages of 13 and 17 (7th- and 11th-graders in schools).

The participants, 427 seventh-grade and 371 eleventh-grade students, completed one writing task: a letter about their opinion on advertisements to a commercial television channel. The evaluation scale was adapted from the IEA scale (Kádárné, 1990), and partly from the Pedagogy Department of Szeged University (Vidákovich, 1990), and it comprised six criteria: holistic impression, content, organisation, style, accuracy, and layout. As Molnár (2002) focused mainly on students' text processing activities, she ignored spelling in the assessment of students' letters. Similarly to the IEA study, each criterion was judged along a five-point scale by two independent assessors. The interrater reliability was moderately strong: 0.52 for the holistic impression (Molnár, 2002, p. 200).

The research results showed moderate development in students' writing proficiency from age 13 to 17. The differences between students' composing competence were higher within both respective age groups than the difference between the results of the two cohorts. This finding echoes the results of the international project introduced by Kádárné (1990). An interesting finding of Molnár's (2002) study revealed that students of different social backgrounds showed parallel development in writing proficiency, but children of less favourable socio-economic status (SES) reached only the level of proficiency in writing by the 11th grade, from which level the pupils of the other group started in 7th grade. Molnár (2002) was interested in the relationship between students' school grades and their performance on the test. She did not find a significant correlation between students' literature grades and their quality of writing, whereas she found a relationship between students' grammar grades and their quality of writing performance. Molnár concluded that analyzing literary texts, as it is common practice in Hungarian Literature classes, is not a sufficient tool to develop composing skills. She could not explain the latter relationship adequately, but supposed that students' higher awareness of language as a system might underlie both good grammar grades and successful text production. Thus, students' cognitive skills may be responsible for both grammar grades and the success of writing products.

The next survey intended to explore students' mother tongue performance at the end of their secondary education. Horváth's (1998) study was conducted in 15 secondary schools in 1995. The focus of her interest was also the level of consensus between judges, and she examined it in the context of the standardised school leaving examination. The students had to solve two reading and two writing tasks. Their texts were judged by two assessors, one of

them being the teacher of the respective student, the other an independent rater. The rating system was concerned with three main aspects of writing besides a holistic opinion of raters: content, structure and language use. The correlation between the two raters was weak concerning several criteria like reasoning (0.19, 0.2) and spelling (0.3). As had been anticipated (Horváth, 1998), the students' teachers assessed not only the actual performance of their students, but, perhaps unconsciously, they also considered other factors while rating tests. Horváth's survey is relevant to my study, as interrater reliability will be one of the focal points of interest in it.

This section summarised the scarce empirical data gained during the past few decades on Hungarian students' mother tongue writing competence. In the next part I will look at the studies concerned with students' L2 language competence, focusing on L2 English results in primary education.

### 3.3.2 Research on Hungarian students' writing skills in L2

Before introducing facts from several projects that examined students' FL proficiency, let us see what people claim about their language competence in Hungary. First, data gained by questionnaires on the Hungarian population's self-reported FL competence will be overviewed. Then, I will look at studies concerning students' measured FL proficiency.

Three sources have published data on self-reported competence: (1) Terestyéni's (1996) study on a representative sample, (2) data from population censuses, and (3) results of a study conducted by the Ministry of Education in 2003.

The first large-scale survey implemented in 1994 (Terestyéni, 1996) demonstrated that the FL competence of the Hungarian population was on a very low level. The research was carried out on a representative sample of people over the age of 14 to inquire about their self-reported FL competence. Although 32 percent of the sample claimed that they spoke a FL at a certain level, only 11.8 percent of them thought that they could use the respective FL for communicative purposes. Only 3.6 percent of the population claimed that they could use two FLs effectively, whereas 0.8 percent said they managed in three FLs. Among the FLs mastered most people identified German (17.3%), English came second (11.5%), while Russian was ranked third (8.8%) in the middle of the 90s.

The population censuses taken in 1990 and 2001 showed that the number of people speaking English developed more dynamically (2.2%-9.8%) than the number of speakers of German (4.4%-10.2%), the latter still outnumbering all other FLs spoken in Hungary at that point. According to the European survey *(Eurobarometer 2005,* I Vágó & Vass*,* 2006) the

number of speakers of English and German measured up, reaching a respective 16 percent of the responding population. In secondary schools these two languages have been the most popular with a tendency of English getting ahead of German (Halász & Lannert, 1998; Halász & Lannert, 2000). By 2005 this tendency strengthened, and more secondary students studied English (53%) than German (37%) (Vágó & Vass, 2006, p. 247).

It is worth comparing these data with the ones concerning the FL competence of the population of the European Union. The latest survey was conducted in 25 countries of the European Union (EU) and the additional four candidate countries in November-December 2006, and included Hungarian respondents as well. It gained data from more than 28 thousand people aged over 15 years (*Special Eurobarometer survey 64.3 'Europeans and Languages',* 2006*)*. Though results differ from country to country, the summarised data demonstrate that more than half (56%) of the EU population claims that they speak at least one FL, and 28 percent of the respondents state that they speak two FLs well enough to have a conversation. Still, almost half of the respondents, 44 percent, admit not knowing any other language than their mother tongue. In six Member States, however, the majority of citizens belong to this group, the countries being Ireland (66%), the United Kingdom (62%), Italy (59%), Hungary (58%), Portugal (58%) and Spain (56%).

According to the 2006 survey, the European rate of speakers of one FL (56%) is significantly higher than the one in Hungary (42%), but it has to be added that, similarly to the Hungarian research data, the proficiency level of the allegedly spoken languages cannot be identified from the EU study, either.

After this short overview of the self-reported language competence of the population let us examine research data on the FL competence of students in Hungarian primary education. The studies to be reviewed in the first section were mostly small-scale surveys carried out in individual counties or cities on a relatively small sample of students. The reason for this was that they were rather the realisation of local authorities' wishes to know more about students' FL competence than part of a national longitudinal survey.

The past decade saw three independent surveys (1996, 1997, and 2000) conducted on students' FL competence in Baranya county, each of them with different aims. A different survey was done in Csongrád county in 1999 and two in Budapest in 1998 and 2000.

Bölcskei (1996) examined 8$^{th}$-graders' English language reading competence using traditional, de-contextualised, discreet-point tests. The instrument measured one element of students' communicative competence exclusively, the use of English. The 341 students' results could be placed along a wide scale. The reliability of the test was weakened by the

high probability of students' producing right answers by pure guessing, as most of the tasks offered 25-50 percent chance to do so. Therefore, the outcomes of this project can hardly be compared to other assessments.

Two other studies explored the language competence of primary-school leavers in Pécs. In the first research conducted in 1997 (Bors, Nikolov, Pércsich, & Szabó, 1999) all, nearly 1,000 students in town completed the test in both their first and second FL (if they studied a second language besides the first one). Three of the language skills, listening and reading comprehension, and writing were measured on meaning-focused tasks. The writing task was a picture description, where students had to write as much as possible about a funny drawing depicting lots of actions. The assessors rated the compositions along a scale based on three criteria: content/vocabulary, accuracy, and cohesion. Each of the criteria comprised three levels of language proficiency: basic, intermediate, and high. One of the drawbacks of the analytic scale was that it did not describe clearly the language competence meant by these levels, though raters received training to standardise criteria. The overall results ranged from very weak to the exceptionally good language competence; as for the three skills, the lowest results were achieved on the writing task, as anticipated.

One of the important findings of this study was that it did not prove an unambiguous relationship between the weekly number of FL classes and the students' performance on the tests, as a good number of students studying the respective FLs in 2-3 classes a week achieved higher scores than their peers receiving 5-6 hours of FL instruction a week. The analysis of these results showed that these high-achievers studied the respective FL as a second foreign language. They had had previous FL learning experience and they studied their first FL in 4-5 lessons per week. Another significant outcome of the study was that it found a strong correlation between students' socio-cultural backgrounds and their test results on the one hand, and it revealed a similarly strong relationship between the number of weekly FL instruction and students' test achievements, on the other. Thus, it was not possible to claim which of these factors influenced students' FL achievement more. It is probable that parents with higher SES enrol their children in schools where more intensive FL instruction is available (Vágó, 2007), and this factor interacts with all the other variables, including the socio-economic factor. Unfortunately, no further analyses are available on this population.

Similar results were gained from another study carried out in eight primary schools of Pécs in 2000 (Bors, Lugossy, & Nikolov, 2001). Besides collecting data from 6$^{th}$- and 8$^{th}$-graders on three language skills (reading, listening, and writing), this study provided insights into classroom processes, as well as into students' and teachers' thinking by way of using

classroom observation and questionnaires. The achievements of 485 students were distributed along a wide scale. It was proved that students studying in traditionally prestigious schools can consistently achieve high scores on FL tests. The study found that these talented students, based on their high achievements in basic literacy domains, had been streamed into intensive FL classes (four to six classes a week) early in their primary education, while low-achievers were instructed in two-three classes weekly. Thus, the differences between high- and low-achievers were destined to increase. This education policy must have contributed to the diverse achievements of students, and accounted for the fact that some of the children did not even start doing the tasks, while others could solve them without mistakes.

A study conducted in Csongrád county, which was part of a comprehensive educational assessment project on students' literacy skills in grades 7 and 11 (Csapó, 2002; Bukta & Nikolov, 2002) intended to explore students' FL skills. Participants were the same population whose mother tongue achievements were scrutinised by Molnár (2002) discussed in the previous section. Skills of 11$^{th}$-graders were examined in two types of secondary schools: vocational and grammar schools. The aim of the study was to compare students' achievements on two English tests: one of which consisted of tasks measuring mostly grammatical competence, whereas the other contained communicative tasks (Bukta & Nikolov, 2002). Findings showed that students' performance was spread along a wide scale not only across schools, but between groups within institutions, as well. Secondary-school students learning in vocational schools had similar achievements to those of 7$^{th}$-graders. The authors attribute these results to the circumstance that vocational schools attract less able students in general, but they suggest that further research is needed to clarify the reasons for the low performance of vocational school students. These outcomes are worth comparing to those of Molnár's (2002) findings concerning mother tongue writing skills of the same cohort from a socio-economic viewpoint: students of different social backgrounds showed parallel development in writing proficiency, but learners of less favourable SES reached only the level of proficiency in writing by the 11$^{th}$ grade, from which level the pupils of the other group started in 7$^{th}$ grade. As learners' SES strongly correlates with their school performances, these results may be due to the fact that more children of less qualified parents and a less favourable SES, who are low achievers in primary school, attend vocational schools, and in such institutions their development is slower than at other secondary schools.

The study also revealed that students performed generally higher on traditional grammar-focused tasks than on meaning-focused ones, this finding being true for all secondary school students. As for primary school students, there are several examples showing the opposite of

this trend. One interesting finding was that two of the primary school groups performed nearly as high on the communicative tests as secondary school groups (35%, 32%) but their results on the grammar-focused tasks were much lower (25%, 16%) than those of secondary students (52%, 78%) (Bukta & Nikolov, 2002, p. 190). This result indicates that firm knowledge of grammar is not a prerequisite but a component of communicative language use. On the other hand, the study revealed that school instruction tends to develop grammatical competence rather than communicative competence, as all secondary school students achieved higher scores on traditional tests. One of the discouraging findings of this research showed that English language competence of students participating in the survey was at a relatively low level and the estimated development in the four years between 13 and 17 was also minimal.

Besides the investigations in Baranya and Csongrád, assessments have been conducted into Budapest students' performances as well. In the capital, a Dutch-Hungarian joint research project inquired into students' performance in several cultural domains, including the L1 and FLs. The project involved 1,031 8$^{th}$-graders from 59 schools in 1999, and 428 students studying in 30 primary schools of Budapest in 2000. Regarding FLs, the two cohorts' English and German reading, listening and writing proficiency was tested and compared on the same tasks in the consecutive years.

In this section I will focus on students' achievements in English, but comparison will be made to results in German where relevant. The average test result measured on three skills was 51.9 percent in 1999, while in 2000 it reached 56.18 percent (Várnai, 2000, p.79.) The lowest achievements characterised students' writing proficiency in both English and German, English test results being: 20.5 percent and 43.2 percent in the two consecutive years. The significant improvement in results in one year is remarkable and the same tendency was demonstrated in the German tests taken in 1998 and 2000. Analysing these data Sturman (2001, p. 72.) hypothesised that improvements must have been caused by a higher ratio of able and high-achieving students in the second sample and, additionally, by the possible influence of the implementation of the new methodology suggested by the *NCC* (1995). It is more probable, however, that teachers had been acquainted with the letter-writing task and got their students practise it before the test. This supposition is strengthened by the fact that students' results on other skills (reading and listening comprehension) did not improve to a similar extent as their writing skills. Furthermore, classroom observation projects (Nikolov, 1999; Nikolov, 2003) and a nation-wide survey into the frequency of typical classroom activities in primary and secondary education revealed that teachers most often apply techniques of the audio-lingual and grammar-translation method both in English and German

classes (Nikolov & Csapó, 2002; Nikolov, 2003, 2004) and the writing skills tend to be the least developed among the four skills (Bors, Nikolov, Pércsich, & Szabó, 1999; Csapó, 2001). As teaching beliefs and practice generally do not change in a short time (Freeman, 1996; Gebhard et al. 2003; Kennedy & Kennedy, 1996), Sturman's (2001) suggestion seems to be both naive and unfounded.

One of the important findings of this study was that, even considering the more than twofold improvement (from 20.5% and 43.2%) in consecutive years, achievements on the FL letter-writing task were extremely low in both years and in both languages. It is useful to compare these data with students' mother tongue achievements in the same project: the average result in Hungarian letter-writing was 58 percent in 1999, and 54.5 percent in the following year (Hegedűs, 2001, p. 23). The slightly lower scores in the second survey do not reinforce the above mentioned assumption (Sturman, 2001) on the tangible influence of the new *NCC* requirements. The reason why the *NCC* could not be a point of reference is that its requirements concerned only students starting their year 1 and those starting year 7 in 1995. Thus, students who started their primary education in 1995 did not reach grade 8 till 2002, so they were not involved in this survey.

In order to analyze and compare reliably, more precise description of the project would have been necessary. In the project report it is not clarified how the respective samples of population solving tasks in different cultural domains compared with one another in the two cross-sectional projects. Furthermore, the relationship between the difficulty levels of the tasks in the three languages is not stated in the summary of the study. What is clear from the published data is that the average achievement of $8^{th}$-graders in Budapest ranged between 50 percent and 62 percent in all three languages (Hungarian, English, and German) on tasks compiled on the basis of the *NCC*. The actual writing tasks or the assessment criteria are not available in the publication.

The final analysis of the relationship between students' achievements and the weekly number of FL classes showed that in both languages students provided with more intensive instruction outperformed their peers studying in fewer weekly lessons. Várnai (2001, p. 87), summarising English as a FL results, identified the instruction in four classes per week to be the dividing line between low and high achievements. Students learning in 2-3 lessons a week produced very low results, those learning in four lessons achieved significantly higher scores, while their peers having five classes of English instruction a week performed best on the tests. The analysis of the survey did not specify if the students achieving excellent results in the less intensive programme learnt that language as a first or second FL. Parents' SES and its relation

to students' performances was not analyzed, thus, lacking these, the data collected in this project cannot be compared with those of the Pécs or Csongrád studies (Bors, Nikolov, Pércsich, & Szabó, 1999; Bors, Lugossy, & Nikolov, 2001; Bukta & Nikolov, 2002).

Another comprehensive study involved volunteering 9th graders in Budapest in 2000. This survey is relevant for the present study, as 9th grade is the first year of secondary education, after streaming primary school students to vocational and grammar schools. The survey aimed at tapping into students' knowledge in eight cultural domains, among them English and German languages on both general and high levels of proficiency (Kiss, 2001). The summary of the survey claims that students' achievements were better in English on both examined levels (Kiss, 2001, p. 6.), but the studies, analysing results separately in the two FLs, did not specify how the subskills measured by different tasks in the two languages compared, and there were no data on calibration, either (Lindner, 2001; Nagyné, 2001; Tagányiné, 2001 a & b). Although the tests included anchor tasks in both FLs for students studying on different levels, the results of students in general and advanced language proficiency classes were not compared to one another. According to the description of the tests, the construct was not clearly determined, separate tasks did not cover separate skills to be measured. In both languages tasks were strongly focused on students' grammatical awareness, translation skills and background knowledge. The lowest results were found on the writing tasks and students' grammatical knowledge was also very weak. Similarly to the Hungarian-Dutch project (Noijon & Várnai, 2001), the tasks were not matched either between the mother tongue and FLs, or between the two FLs. Having summarised small-scale research results, let us gain insight into large-scale national surveys of FL levels in Hungarian public education.

Hungarian students' FL skills were assessed to monitor the levels and efficiency of FL education in state schools in 2000 and 2002 (Csapó & Nikolov, 2002; Nikolov & Csapó, 2002) and in 2003 (Józsa, 2003; Nikolov, 2003; Nikolov & Józsa, 2006). Some of these measurements were parts of major Hungarian educational research projects in other school subjects (Csapó, 1998; 2002). Besides language skills tests, data were collected on the participants' social and language learning background and plans, attitudes, motivation, and classroom activities. To estimate their general cognitive abilities, a standardized inductive reasoning skill test was used (Csapó, 1998).

Representative samples of students participated in three projects inquiring into their English and German language competence in 2000, 2002 (Nikolov, 2003) and 2003 (Nikolov & Józsa, 2006). The sample for data collection involved school classes of grades 6, 8, 10 and 12 from approximately 330 state schools, including nearly 50,000 FL learners. The results of

the 2002 national survey (Csapó & Nikolov, in press) will be related to the data of the present study, as tasks and assessment methods were basically identical in the two projects. In the national survey the poorest achievements characterized students' writing proficiency, the national mean (2002) of 3,653 8th-graders being as low as 33.7 percent. In the regional breakdown of results the average of the Southern Trans-Danubian region, which Baranya county belongs to, surpassed the national performance reaching 38.8 percent. Some other findings of these large-scale projects are also relevant for the present study: (1) students' language performances showed a lot of variation in all cohorts: learners in some groups achieved top scores, while other groups performed on extremely low levels; (2) moderate to strong correlation emerged between the size of settlement, learners' SES, cognitive skills, and weekly language classes. These results confirm that in Hungary, where access to the examined target languages is mostly limited to the classroom, students' SES and developmental level of cognitive skills strongly influence their achievements in a FL. Similar findings characterise achievements in other school subjects (Andor, 2000; Csapó, 1998, 2002).

An international comparative research study (Mihaljević-Djigunović, Nikolov, & Ottó, 2006, 2008) is worth mentioning as my study examines one of the skills of participating Hungarian students. This project provided insights into Croatian and Hungarian 8th graders' performances in their mother tongue and in EFL. More than 700 fourteen-year-old students participated in the study in two neighbouring regions of Croatia and Hungary. The number of Hungarian students was 247. The study examined how Croatian and Hungarian 8th graders' performances compare on EFL tests; how they compare by groups, within groups, by length of instruction, weekly classes, and size of group; and the relationship between Croatian and Hungarian students' achievements on tests in their L1 and EFL. The findings showed that Croatian students performed significantly better on the EFL proficiency tests than their Hungarian counterparts; larger differences characterised performances between groups in Hungary than in Croatia, whereas no significant differences were revealed within groups. Learners who started EFL earlier tended to achieve higher scores than later starters, whereas the findings on group size and weekly classes were more controversial.

The number of weekly English classes correlated with students' EFL performance in the Hungarian cohort. The results of these students coincide with findings in a nation-wide survey involving learners of English in years 6 and 10 where higher weekly number of classes resulted in higher test achievements (Józsa, 2003; Nikolov, 2003).

As seen from the results overviewed in this chapter, it is extremely difficult to draw generalisable conclusions concerning the level of FL competence accessible in public education based on achievement tests piloted and used so far. What seems transparent is that there are significant differences within age groups and types of institutions, as well as between classes and also between individual students.

Vágó (2007), investigating routes of language learning in Hungary, points out that the beneficiaries of an early start in FL learning, that is of a longer language learning route, are the offsprings of highly qualified parents. The other predictor of the availability for an early start is the type and size of settlement the school is located in. In large towns two thirds, in Budapest 62 percent of the pupils receive an additional year to the three years of language instruction, whereas in small towns this extra FL tuition is available for 50 percent of the children. In villages 43 percent of youngsters can start studying an FL earlier than the mandatory year ($4^{th}$ grade). These claims strengthen the results of other studies concerning the strong influence of children's SES on their access to FL learning, and, consequently, their language proficiency levels measured in these investigations (Andor, 2000; Bors, Lugossy, & Nikolov, 2001; Bors, Nikolov, Pércsich, & Szabó, 1999; Bukta & Nikolov, 2002; Csapó, 1998, 2002).

### *3.4 Summary*

The assessment projects examined in this section were not related to one another, as they were carried out independently. They revealed great differences in a lot of areas: the selection of participants, the theoretical foundations, tasks and text-types used in the measuring instruments, the management and implementation, ways of data processing and analysis. Due to these facts, the validity and reliability of independent surveys seem to be weakened, and, furthermore, the results cannot be compared to each other. Nevertheless, a discouraging conclusion can be drawn from the previously discussed studies: students' FL writing competence proved to be the least developed skill of the three language skills measured in most surveys (Bors, Nikolov, Pércsich, & Szabó, 1999; Bukta & Nikolov, 2002; Nikolov, 2003; Sturman, 2001; Tagányiné, 2001 a, b). This is also true for mother tongue competence, according to assessment surveys reviewed in the previous section (Kádárné, 1990; Molnár, 2002; Orosz, 1972). Although the *National Core Curriculum* (2003) stresses the importance of developing oral skills in primary education, writing is still an important tool of

communication, which is essential in students' future life, and is definitely worth using efficiently.

In the first three chapters of my dissertation I investigated theoretical and practical considerations connected to L2 writing; I summarized approaches to analyzing writing and explored international research findings on students' texts in L1 and L2. Then, I described types of linguistic analysis of learner language and showed the results of the morpheme studies. I introduced the research methodology of writing studies and dealt with the key concepts of the assessment of writing. Then, I examined the socio-educational context of Hungarian learners' writing in L1 and L2 by studying both curriculum requirements and research findings concerning the level of meeting these requirements.

In the next chapter, the empirical study on 231 Hungarian primary school leavers' English writing proficiency will be presented. It investigates students' scripts by using mixed methods; quantitative data on employing four selected language forms will be gained by applying corpus linguistics, whereas qualitative description of learners' writing will be accomplished on the basis of text linguistics. Students' writing proficiency will be related to curriculum requirements. So as to measure the reliability of assessment, the relationship between different types of assessment will be scrutinised. The problems of the interpretation of assessment scales and possible reasons for it will also be explored.

# Chapter 4: An empirical study on 8th graders writing skills in English as a foreign language

## *4.1 Background to study*

My study is partly based on data gained from an international comparative project (Mihaljević-Djigunović, Nikolov, & Ottó, 2006, 2008), which tapped into four language skills of Croatian and Hungarian eighth-graders and investigated relationships between students' performances and different background factors. I participated in this project as the coordinator of research in Hungarian (Baranya county) schools, and as an assessor of these students' scripts. My study focuses on Baranya students' writing skills in L2 English.

## *4.2 Rationale*

L2 writing of primary-school learners has rarely been analyzed (e.g., Lasagabaster & Doiz, 2003; Lee & Huang, 2004; Silva & Brice, 2004; Torras, Naves, Celaya, & Pérez-Dival, 2006); this is also a white spot on the Hungarian map of applied linguistic research. Apart from local studies (Bors, Nikolov, Pércsich, & Szabó, 1999; Bors, Lugossy, & Nikolov, 2001; Bukta & Nikolov, 2002; Várnai, 2000), only a few representative national investigations (e.g., Nikolov, 2003; Nikolov & Józsa, 2006) involved research on young EFL learners' writing proficiency. The reason for this scarce interest might lie in the lower language proficiency level of these pupils, which narrows possible research foci. By analysing Baranya students' writing I can answer questions concerning typical pupils' EFL competence in the county, and then, these results can be compared with the national L2 standard of the same age group, as well as with the requirements of the *NCC* and A1 level of the *CEFR* (Council of Europe, 2001).

In addition to the limited number of studies providing quantitative data on young learners' writing, there has been no published qualitative analysis of students' EFL written performances in this age group in Hungary, so, I am convinced this area is worth exploring. By the close investigation of learners' scripts I can gain data on one of my research questions (RQs) inquiring into the development of learner language. This area has been studied mostly by gathering oral production data of L2 learners (Bailey, Madden, & Krashen, 1974; Dulay & Burt, 1974; Ionin & Wexler, 2002; Pica, 1984). I intend to scrutinise the features of learner language through written performances.

The other focus of this study concerns the reliability of assessment, which will be explored by three methods: (1) by comparing raters' holistic and analytic judgments, and (2) by comparing one assessor's decisions (Rater 5) made on the same scripts holistically, then, based on two different scales. (3) Rater 5's judgments will also be compared with the quantitative data on language accuracy gained from corpus-tagging.

### *4.3 Research questions*

I will aim to find answers to the following RQs:

1. How does Baranya students' writing proficiency compare with their other EFL language skills?
2. How does Baranya primary-school learners' L2 proficiency compare with the L2 proficiency of the same age-group in a national sample?
3. What level of writing proficiency do Baranya primary school students reach in English by the end of their primary education?
4. What is the relationship between students' writing proficiency and the curricular achievement targets?
5. What is the relationship between students' socio-educational background and their writing proficiency?
6. What can we learn about students' texts from a corpus analysis? How accurately do 8th-graders use simple target language forms (PRC, EXP (there/it), NEG, & PLU)? What developmental patterns emerge?
7. What is the relationship between criterion-based and corpus-based measurement of accuracy?
8. How are raters' opinions of sample tests related to each other?
9. How does criterion-based assessment compare with holistic assessment?
10. How do raters' interpretations of *CEFR* scales compare?
11. What are typical performances like? What developmental levels do they reflect?
12. What range of vocabulary characterizes 8$^{th}$ graders? What is the relationship between criterion-based and corpus-based measurement of accuracy and vocabulary?

## *4.4 Overview of research methodology: A mixed approach*

### 4.4.1 Triangulation of data

My empirical study relies on both quantitative and qualitative data collection methods, as the use of multiple research techniques enhances the credibility of the investigation (Mackey & Gass, 2005, p. 164). Whereas quantitative research aims to unveil facts that are independent of the researcher, qualitative researchers want to interpret the background behind the data. The line dividing qualitative and quantitative research types is not clear; according to Grotjahn (1987) the distinction between them is an oversimplification. His analysis relies on the (1) data collection method (experimental or non-experimental), (2) the type of data that resulted (qualitative or quantitative), and (3) the type of analysis conducted on the data (statistical or interpretive). Grotjahn (1987) points out that by the combination of these three factors a wide variety of research types exists. Survey research provides some common ground between qualitative and quantitative approaches (Brown & Rodgers, 2002). Within mixed methods of inquiry researchers can shape the procedures to best suit their research aims. I have chosen concurrent procedures to integrate quantitative and qualitative data (Dörnyei, 2007) in order to provide a comprehensive analysis of my research results

### 4.4.2 Data processing and validation procedures

In order to acquire quantitative data on students' language use, first, I needed to quantify qualitative data (Creswell, 2003), in other words, code learners' texts. This involved multiple steps:

(1) reading and assessing 231 students' writings with the help of analytic scales, thus getting scores. The analytic scales had been designed and validated in an earlier national research study in 2002 (Csapó & Nikolov, in press); they described L2 proficiency along four criteria: task achievement, vocabulary, accuracy, and text cohesion;

(2) analyzing scores quantitatively, which was done by SPSS software;

(3) typing in students' writings;

(4) creating codes for selected target language forms used in students' writings, and identifying the number of times these forms occurred in the text. Codes were named by shortening the name of the investigated language form; counting was accomplished by ticking the occurrences of correct and various erroneous forms of the respective language features in an excel table.

(5) counting the range and frequency of vocabulary items with the help of a software.

This quantification procedure enabled me to compare quantitative results with the qualitative data.

Based on the assumption that statistical data illuminate only general features of students' L2 competence, I also investigated a sample of writings qualitatively providing detailed analysis of their task achievement, vocabulary, accuracy, and text cohesion. As part of triangulating data, I asked five experts to rate 15 selected compositions, first by rank ordering them, then, by placing them in the appropriate band of the *CEFR* (Council of Europe, 2001). This procedure had three aims: (1) it served to compare holistic assessment with criterion-oriented assessment, and (2) it allowed an investigation of correlation between raters, plus (3) it inquired about correlation between the first rating and the second assessment done by the new raters. Thus, reliability of the first rating was checked, and, as the first rater also participated in the second phase of assessment, intra-rater reliability was explored.

## *4.5 Participants*

### 4.5.1 Students

A total of 247 primary-school students from small villages, small towns and from the biggest town of Baranya County participated in the project, out of which 231 took the writing test. These types of school were selected, as they were considered relevant from the point of view of learners' socio-educational context. All the students studied in 8$^{th}$ grade, the final year of their primary education. The sample is close to representative considering the geographical distribution of students by location, though city-dwellers are a bit over-represented by contributing nearly 60 percent of participants from five schools, town-dwellers of three schools are slightly under-represented by giving 23 percent of the cohort. The discrepancy is only 5-6 percent compared to other national studies, for example, Csapó (2001), Csizér (2007), Nikolov and Józsa (2006), so we can state that representativity is sufficient. Villagers are well represented by supplying 18 percent of participants studying in three different institutions (see Table 3).

Students' socio-educational background is one of the focal points of my investigation. Their L2 proficiency is mostly examined by school in this study, as it is not the individual socio-cultural background that is of crucial interest to me, but the educational setting. The map (Figure 1) illustrates the geographical distribution of institutions participating in the research project.

Table 3 Participants by location and schools

| Location | School code / Number/location | Number of students' written performances | % of total number |
|---|---|---|---|
| Large town | 1 | 55 | 24 |
| | 2 | 44 | 19 |
| | 3 | 10 | 4 |
| | 4 | 26 | 11 |
| | 11 | 3 | 1 |
| **Large town total** | **5** | **138** | **59** |
| Small town | 5 | 13 | 6 |
| | 6 | 18 | 8 |
| | 7 | 20 | 9 |
| **Small town total** | **3** | **51** | **23** |
| Village | 8 | 11 | 4.5 |
| | 9 | 11 | 4.5 |
| | 10 | 20 | 9 |
| **Village total** | **3** | **42** | **18** |
| **Total** | **11** | **231** | **100** |

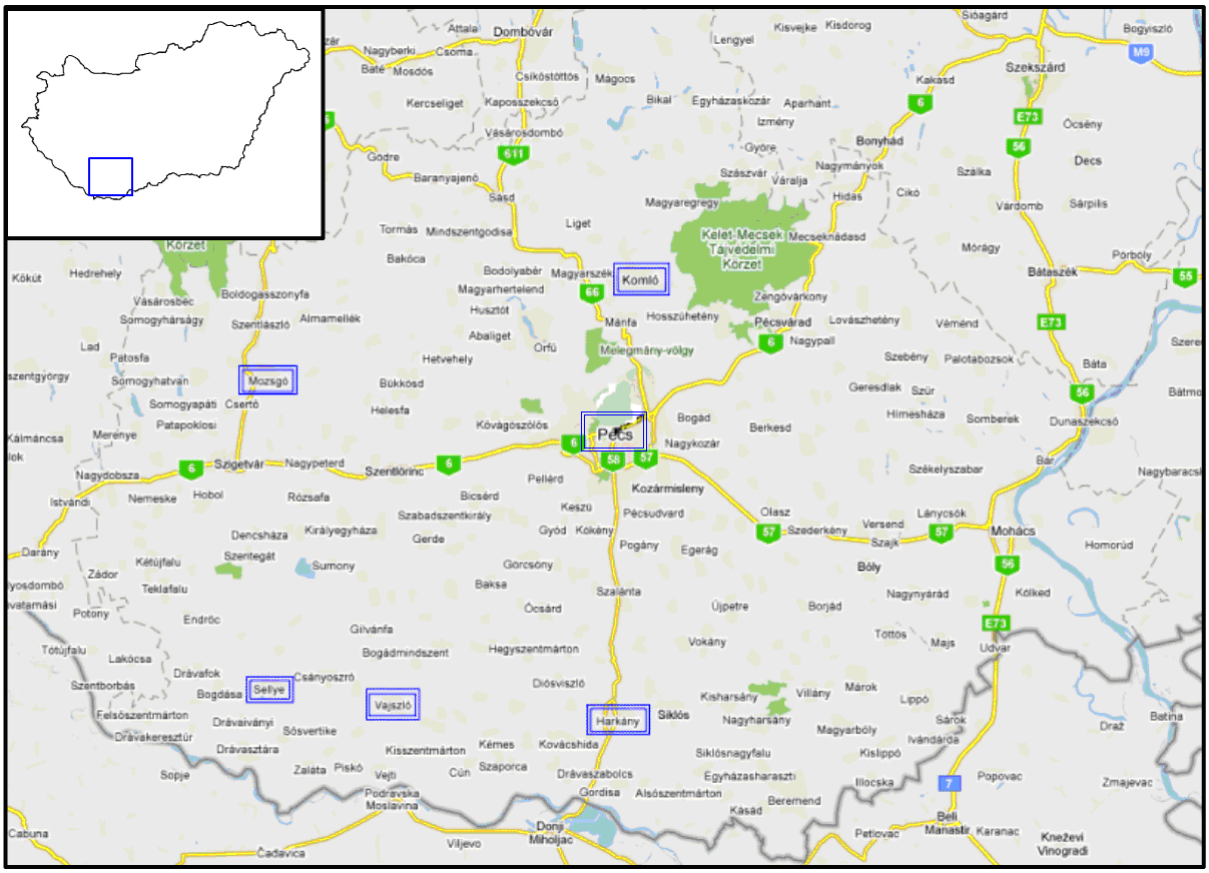



Figure 1: Map of Baranya county with villages, towns of participating schools labelled

### 4.5.2 Raters

The second group of participants comprises six teachers of English who undertook the job of assessing students' written performances. Two assessors were responsible for the standardization of scores based on an analytical scale, then, one of them (the author) evaluated all tests. In the next stage of my inquiry, five teachers of English, including myself, reassessed sample tests along new criteria. These experts had either had long years of experience in teaching and testing students' classroom L2 performance, or had been well trained in the theory of testing. Two of them (Raters 3 and 4) worked as item writers in The Exam Reform Project and received continuous training for four years between 1998-2002. This project was funded in May 1998 by the British Council, and was officially launched with the aim of developing a standardised, multi-level, valid and reliable school-leaving examination in English for the Hungarian School-leaving Examination Reform. Item writers took part in several one-week workshops a year where they were trained in item writing and task design, pilot design, statistical analyses, and data interpretation. Two of our assessors (Raters 2 and 5) received training from experts previously trained in the Exam Reform Project. One of the assessors (Rater 1) took part in trainings organised by a major examination board in Hungary (European Consortium for the Certificate of Attainment in Modern Languages). All of them were acquainted with the basic approach to assessing language proficiency demonstrated by *CEFR* (Council of Europe, 2001).

## *4.6 Instruments*

### 4.6.1 L2 English tests

#### 4.6.1.1 Reading and listening comprehension tests and pragmatics in L2

For EFL a battery of tests consisting of two test booklets (one on reading comprehension and pragmatics, and one on listening comprehension and writing) and a speaking test were used. The tests in the two booklets had been designed and validated first in Hungary in 2002 and later in Croatia (Mihaljević-Djigunović, Nikolov, & Ottó, 2006). All the tests were meaning-focused and were based on the *CEFR* (Council of Europe, 2001) levels the EFL learners were expected to reach at end of year 8 (A1 and lower band of A2 level).

The topics as well as task types were familiar to the participants: they were not borrowed from any particular teaching material but they were highly similar to the tasks in the course books they used in school. The vocabulary and structures of the texts were expected to be on or a bit beyond the level of the target population. The estimated level was hypothesised to cover a relatively wide range (A1 and lower band of A2) as EFL learners in year 8 include

learners who started learning English at the mandatory start in year 4 and those who had started earlier in one to three classes per week (Mihaljević-Djigunović, Nikolov, & Ottó, 2006).

The length of the texts in the items ranged from a word, an expression, or a sentence, to a short passage. All tasks focused on meaning, not form, and reflected the achievement targets defined in the national curricula for these age groups in both countries. The texts were authentic, except for the listening tasks, where scripted materials were used. The reading booklet was compiled in two versions: the actual tasks being identical with an altered sequence. The rubrics were given in English and all listening and reading tasks started with an example to ensure that even if the instructions were slightly beyond test-takers they could understand what to do.

#### 4.6.1.2 L2 writing test

The writing task expected pupils to write a short text based on two picture prompts that looked the same but included ten small differences (Figure 2). The two nearly identical pictures placed side by side illustrated a four-member family at home. Students were to describe differences between the two pictures. The rubrics guided students' attention by providing the ten items to compare. These vocabulary items were expected to be known to all learners, as they were all related to topics prescribed in the *National Core Curriculum* (1995). The reason for listing these items was to avoid the influence of additional cognitive demands such as the necessity of perception of differences. By supplying these words, the prompt offered ten vocabulary items for the students to use, thus not testing them on productive vocabulary knowledge of these words, and making the task easier. Pupils were expected to recognize the items and then, to use them in their writings.

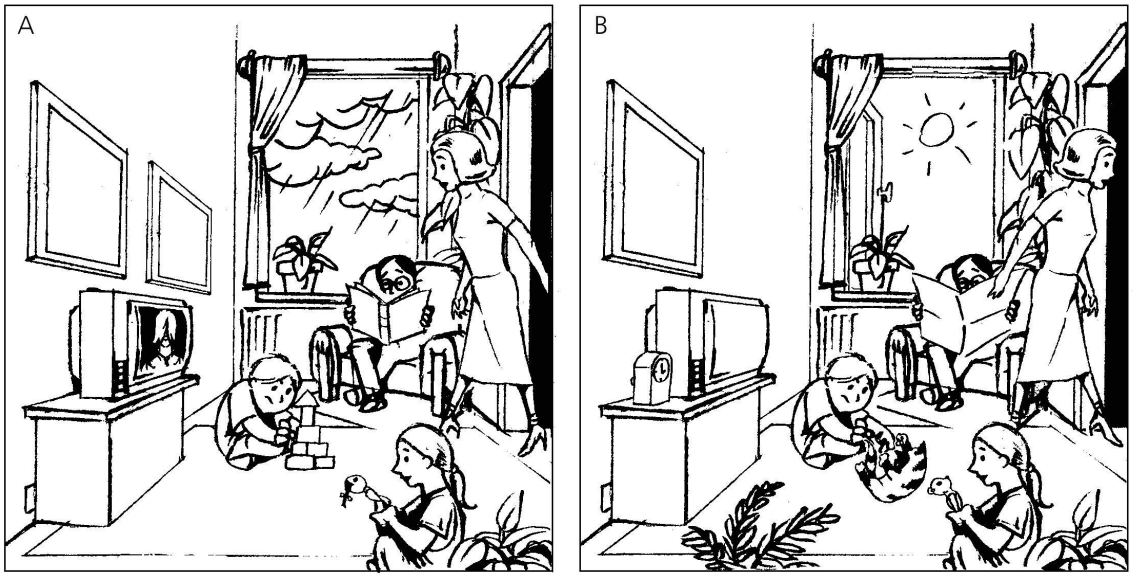



Figure 2: L2 writing task
Look at pictures A and B. Write about the differences between them. Write about the boy and the girl, the man, the woman, the TV, the clock, the wall, the plants, the window and the weather outside.

A model version of the completed task was written by the author to see what solutions can be expected.

In picture A the *boy* is playing with a toy but in picture B he is playing with a cat.
In picture A the *girl* is playing with a doll, but in picture B she is playing with a teddy bear.
In picture A the *man* is reading a book, but in B he is reading a newspaper.
In the first picture the *woman* is coming into the room, but in B she is going out.
In the first picture there are two *pictures* on the wall, but in B there's only one.
In picture A there isn't a *clock* by the TV, in B there is one.
In picture A there are three *plants* but in picture B there are four.
In picture A the *TV* is on, in B it is switched off.
In picture A the *window* is closed, but in picture B it's open.
In the first picture it's raining, but in the second one the sun is shining.

The length of the model version is 161 words, in which nine out of the ten vocabulary items offered by the rubrics have been supplied, *weather* is not used. If the often repeated unnecessary expressions *in picture A* and *in picture B* are deducted from the total sum of words, an approximately 100-word-long written performance should meet the task requirement.

In terms of vocabulary, simple everyday words were expected, verbs like *play, read, come, go, be on/off, rain,* and *shine,* but some others were also possible: *sit, build, switch, watch,* and *leave.*

The expected nouns included *toy, cat, doll, (teddy) bear, book, newspaper, room*, and *picture,* whereas adverbs to be used were related to the window: *open* and *closed*.

In relation to grammar the following structures were targeted for picture description: *present continuous tense* for specifying the four people's actions, the *expletive there* for stating the presence or absence of the clock, the number of pictures on the wall, and the number of plants in the room. The other *expletive it* was required when reporting on the weather outside. The use of the verb 'be' as a copula was expected in connection with the TV and the window. The *negative* construction was required to report about the absence of the clock. NEG was also available as an avoidance strategy when pupils did not know a word, for example, *he is not reading a book.* In these cases NEG was anticipated for expressing the absence of an action or the object of a certain action.

As for cohesion, simple cohesive devices were required in order to connect parts of sentences and clauses, such as the conjuncts *and* or *but*, or anaphoric references, such as the personal pronoun in the second clause: *First, the man is reading a book, in picture B he is reading a newspaper.*

Students' written performances will be referred to as *writing*, *text*, or *script* interchangeably in the study.

### 4.6.2 Mother tongue tests

#### 4.6.2.1 Reading and listening comprehension tests and pragmatics

I will summarise mother tongue tests in short, as they are not the main focus of this study. L1 test booklets included tasks in reading comprehension, listening comprehension, writing, and pragmatics. These were new tests designed and piloted on a similar population in both countries specifically for the purpose of this project. The reading and listening comprehension tests comprised authentic input texts. The task types included multiple choice items and required short answers from participants.

The listening booklet comprised three tasks which were recordings of authentic national radio programs that 14-year-olds would be expected to listen to and find intrinsically motivating. The task types used in the tests investigated close listening as well as skimming.

The reading booklet included five tasks. In terms of content the tasks comprised tables, popular science texts, news from daily papers, texts from encyclopaedias and literary texts. The reading subskills tapped were skimming, scanning as well as intensive reading.

The pragmatics test included one task with six multiple choice items related to situations such as asking politely for a bill in a restaurant, and reacting to friends being late for the cinema.

#### 4.6.2.2 L1 writing booklet

The mother tongue writing tests included two task types. In Writing task 1 the students were asked to write a letter to a travel agent concerning a place where the participant and three friends would spend a few days in their holidays. Writing task 2 required participants to express their opinion on allowing students to draw graffiti on the wall of their institution. This task also provided clues concerning people whose opinions should be contemplated. This included parents and other schools.

These task types are not well-known to students, as in their mother tongue lessons most written tasks are concerned with their compulsory readings. Literary essays related to poems or short stories are better-known and more thoroughly practised genres than personal letters or argumentative essays. Thus, the writing test meant a real challenge for many of the participants.

### 4.6.3 Questionnaire

A questionnaire was administered in the participants' mother tongue. Besides detailed background data, participants were asked about the time they started studying English and how many classes they had in a week, as well as about their English marks.

### 4.6.4 Assessment scales

#### 4.6.4.1 L2 assessment scale

For the L2 writing task an assessment scale (Appendix A) was constructed along four criteria: task achievement, vocabulary, grammar/ accuracy and text cohesion. Each criterion was assessed according to the descriptors divided in five bands. The lowest band served to identify completely insufficient knowledge that was worth zero points. In each of the following four bands one or two scores were provided for participants' relevant performance, so judges were free to decide if the students' writing fell in the lower or upper end of the language proficiency level explained by band descriptors. Thus, eight scores were awarded for the best achievements in each performance category, and altogether 32 points could be gathered.

Let us get acquainted with the band descriptors worth one to eight scores one by one. Students received 0 score for task achievement if they wrote only a few words or their text

was irrelevant to the prompt. In case of 0 task achievement the other criteria were not even considered by raters.

Performances fell in the first band if they were characterised by a very limited scale and inappropriate vocabulary; the text was unstructured and incomprehensible because of grammatical or spelling mistakes. The second band characterised language performances by the following descriptors respectively: two or three things relevant to the pictures; a limited scale and choice of vocabulary; many mistakes in a partly comprehensible text, where the same sentence type is repeated. The middle band (Band three) was meant to draw the line between insufficient and acceptable performances in relation to curricular achievement targets. Concerning task achievement the pass-band says: "Text is on five or six things relevant to pictures, or on seven or eight partly relevant things. Text is on both A and B." Students meeting these criteria were provided three or four scores depending on the measure of relevance to the prompt.

As for vocabulary, the pass-band required a "good scale or choice of vocabulary, mostly appropriate to task", which was appreciated by three or four scores depending on the level of appropriacy. The accuracy requirement of the middle-band allowed students to make several mistakes but required most of the text to be comprehensible. To obtain half of the maximum score for cohesion pupils' texts were to consist of a sequence of sentences varying one or two sentence types.

Adding up the scores of the four criteria in the medium band (Band three) we can see that achieving 12-16 scores would mean an acceptable performance of the task. Judges participating in the standardisation process agreed that performances falling in the medium band would approximately cover level A1 of the *Common European Framework of Reference* (Council of Europe, 2001). In the section analyzing the relationship of raters' assessments (4.8.4.4) and in the qualitative analysis of five learners' scripts (4.8.5) we will see to what extent this hypothesis proved to be true.

As I described in the introduction of the second band, scores below twelve indicate a low level of language proficiency in one or more areas, for example, a limited scale and choice of vocabulary or often inappropriately used words. At this proficiency level many mistakes occur, and only part of the text is comprehensible, thus the student's language proficiency is below A1.

On the other hand, results between 20-24 points (Band four) should represent good, reliable knowledge of basic vocabulary and structures, while student performances exceeding 27 points (Band five) indicate above average proficiency using a rich scale and good choice of

vocabulary with few grammar or spelling mistakes which do not interfere with comprehension. In this top band students' texts are well structured: parts on different things are separated and sentences are logically linked. According to descriptors, there are some complex sentences and more than three sentence types vary. This top band describes L2 performance that is beyond level A1, the targeted language proficiency for all primary school-leavers. Let us see one of the best scripts produced by a student from School 1, who received 32 points for this composition:

> In picture A the little boy is playing with some toys, but in picture B he's playing with a cat.
> The little girl next to him is playing with a doll in picture A, but in picture B she is playing with a bear.
> The man behind them is reading a book in picture A, but in picture B he's reading a newspaper.
> In picture A a woman is coming into the room, in picture B she's going out.
> In picture A there're two pictures on the wall, but in picture B there's only one.
> In picture B there is a clock next to the TV, but there isn't in picture A.
> There is a plant next to the TV in picture B, but in picture A there isn't.
> In picture A the TV has switched on, but in picture B it has switched off.
> The window has been opened in picture B, because the weather is beautiful, the sun is shining, but in picture A the window has been closed, and the weather is bad, it is cloudy and it is raining.

This pupil displayed a good choice and wide scale of vocabulary, as well as complex verb forms (i.e., has switched on; has been closed) that are not required by the task. Although he used these complex forms inappropriately, he attained maximum points for accuracy, as the mistakes did not interfere with comprehension. I will provide detailed analyses of five students' scripts under 4.8.5.

Student performances often reflect uneven language proficiency levels concerning the four criteria, as their subskills may develop more rapidly in some of the areas. Thus, there are students whose results fell in different bands concerning task achievement, vocabulary, accuracy or cohesion. These performances attained a variety of scores between 8-12, 16-20, or 25-28 points.

#### 4.6.4.2 Raters' task sheet

The five raters had two main tasks; their first task was to rank order a sample of student scripts according to their level of English language proficiency using holistic assessment (Appendix B). Then, they were asked to place writings in the appropriate box of a grid based on the language features of the text, described by *CEFR* (Council of Europe, 2001) descriptors for writing (Appendix C). Their last task was to write how well they knew the

*CEFR* and summarize their comments on the process of assessment. I will discuss raters' tasks in detail under procedures 4.7.2.

## *4.7 Procedures*

### 4.7.1 Test administration and first assessment

The paper and pencil tests were administered to students in classroom-size groups in May, 2004. Participants used a 45-minute class session for the reading booklets in each language, 45-minute class sessions for listening and writing tasks in each language, whereas a separate session was devoted to filling in the questionnaires. All tests and questionnaires were coded and assessed centrally during the summer of 2004. Oral tests, observations and interviews with EFL teachers were conducted during the week or the week following test administration.

Tests included tasks that were simple to score and assess (e.g., multiple choice items), whereas others required some standardization (short answers) and the construction of assessment tools and sophisticated training (writing and speaking tasks). For the writing tasks (one in English and two in the L1) three separate assessment scales were constructed along four criteria: task achievement, vocabulary, grammar/accuracy and text/cohesion with scales of four bands each. Assessment of speaking performance was done by means of a specially designed assessment scale constructed along the following criteria: task achievement, vocabulary, accuracy and fluency, pronunciation and intonation with a scale including five bands.

The assessors of both writing and speaking were trained in order to reliably and consistently apply the scales. Since such training has to focus on the actual tasks, four sets of training were conducted. Length of the training depended on how much time the assessors needed to standardize their criteria (between three to five hours). The inter-rater reliability of Hungarian assessors concerning L2 English tests proved remarkably high: 0.96 Cronbach's alpha, exceeding the internationally accepted 0.75 by far. It also surpassed the average of 0.84 Cronbach's alpha achieved in the IEA survey (Kádárné, 1990). This means that the difference between scores presented by the two assessors was zero to one score including all four assessment criteria.

After the assessment of papers all participating institutions received coded statistical data concerning all schools and detailed information on their students' performances in November 2004. For the purpose of this study, that is, to acquire a closer look at students' L2 proficiency

in the selected language areas, I needed to continue data analysis and processing. In the next section I describe this procedure.

### 4.7.2 Second assessment of a sample of writings

#### 4.7.2.1 Selection of compositions for reassessment

The number of writings to be chosen for reassessment had to be manageable for raters, and, at the same time, it had to be high enough for representing students' various language performances. Fifteen writings seemed to meet both requirements. So as to include as many performance levels as possible, first I selected writings which received maximum scores (32) for all criteria, then I chose compositions with decreasing scores on each criterion. To make rank ordering a bit more of a challenge, two compositions of equally 30 scores were chosen. Samples from various schools were taken, but representation of all institutions was not a point of consideration in this process. Original writings were recoded so as to hide all clues referring to school identities. Letters of the alphabet were chosen as new codes for identification excluding the overused first four letters (a, b, c, d) to prevent letters from suggesting an order of excellence. Photocopied versions of the 15 writings were handed to raters together with their task sheets. Table 4, which was not handed to raters, illustrates the rank order of selected writings based on scores awarded for each performance criterion during the analytic scoring procedure by the first rater.

Table 4: Rank order and scores of the 15 selected writings based on first assessment

| Rank 1-15 | 1 | 2-3 | 2-3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Students' code | F | V | S | G | R | M | P | H | K | E | L | O | N | T | Z |
| Task achievement | 8 | 8 | 8 | 6 | 7 | 6 | 5 | 6 | 5 | 3 | 4 | 3 | 1 | 1 | 0 |
| Vocabulary | 8 | 7 | 6 | 6 | 5 | 5 | 4 | 5 | 4 | 3 | 3 | 2 | 1 | 1 | 0 |
| Accuracy | 8 | 7 | 8 | 6 | 5 | 5 | 5 | 3 | 4 | 4 | 2 | 2 | 0 | 0 | 0 |
| Cohesion | 8 | 8 | 8 | 7 | 7 | 6 | 5 | 4 | 4 | 4 | 4 | 3 | 1 | 0 | 0 |
| Total score | 32 | 30 | 30 | 25 | 24 | 22 | 19 | 18 | 17 | 14 | 12 | 10 | 3 | 2 | 0 |

#### 4.7.2.2 Rank ordering samples of students' writing

The raters' first task was to rank order a sample of 15 student writings according to their level of English language proficiency using holistic assessment (Appendix B). Accordingly, they did not receive the analytic scale developed for the first assessment, but were to rely on their overall impression of the students' performances. They were supposed to rank the scripts starting from the best work and finishing with the poorest writing.

#### 4.7.2.3 Placement of samples along levels A1-B1 of *CEFR*

Raters' next task was to place the students' scripts in a grid with randomly placed descriptors representing levels A1-B1 of the *CEFR* (Council of Europe, 2001; Appendix C). The descriptors were related to following criteria of writing, all described in the C*EFR* (Appendix D):

(a) general linguistic range (GLR),

(b) overall written production (OWP),

(c) vocabulary range (VOC), and

(d) grammatical accuracy (GRA).

Thus, raters could not automatically follow a preset order of descriptors representing greater language proficiency as proceeding from the first grid to the last one, that is, from level A1 towards B1. Instead, they first needed to interpret descriptors, then, match them to the level of represented language proficiency. Then, they were free to assign each writing in as many boxes of the grid as they liked, though four matches were suggested as optimal; one in each criterion. Thus, if they felt unsure about the level of proficiency of the composition in relation to a certain criterion, they could choose to place it in two boxes within one category. The aim of this task was to examine how raters cope with the interpretation of the descriptors of *CEFR.* All of them had encountered and worked with this European document previously, but none of them claimed that they knew it well. No standardization of raters took place, as one of the aims of the research was the examination of raters' individual interpretation of *CEFR* descriptors on the same task.

These four criteria for the measurement of writings in *CEFR* were chosen, as they could be paired with some of the criteria of the analytic scale used in the first assessment. The areas especially overlapping with the analytic criteria were: VOC, GRA and OWP. Overall written production can be related to cohesion in the analytic scale applied in this study, as it concentrates on the use of linking devices. GLR is an overarching term suggesting that it includes all areas of language performance, although the descriptors in the *CEFR* scale focus primarily on vocabulary, social functions, and grammatical structures that learners can apply.

There was an additional point in the reassessment procedure: Rater 5 is the author of this study, thus, the results of the first analytic assessment could be compared with the same person's holistic judgment of the selected writings two years later. As considerable time passed between the two periods of assessment, students' written performances could be read without bias for the second time.

### 4.7.3 Data processing

In this section the steps of quantifying qualitative data (Creswell, 2003), and other steps of data processing will be described in detail based on the research methods introduced in section 4.4. My research questions aim to get information on students' L2 English proficiency, focusing especially on their (1) language accuracy, (2) richness of vocabulary, and (3) fluency. In order to obtain necessary statistical information on the selected areas of language proficiency, I needed to enter all students' compositions into the computer first.

#### 4.7.3.1 The selection of language forms to be investigated for language accuracy

First, I will illustrate how data on language accuracy were gained and coded. To start with, I will explain how I designated the language forms to be investigated. The task itself, which was a picture description, elicited certain language forms, as it was introduced by the model task. After reading all the compositions, and looking for characteristic patterns I could identify which forms and constructions emerged most often. They were as follows:

(1) present progressive /continuous (PRC) was the main verbal form necessary for describing the four people's actions;
(2) expletives (EXP) *there* and *it* were to be used for depicting places of objects and reporting on weather conditions;
(3) the plural form of nouns (PLU) was needed for recounting about plants and pictures in the room;
(4) the negative construction (NEG) was to be applied for stating the absence of the clock in picture A.

Thus, I selected these four language forms emerging in students' compositions most frequently. As all the selected L2 forms elicited by this writing task are introduced in course books for beginners and elementary learners generally used in Hungarian primary schools, students usually get acquainted and practise them in the first two years of their English learning. The participants were at least in the fifth year of their L2 education, so all of them must have met the selected language forms, even the pupils who did not use them.

#### 4.7.3.2 Creating the coding system

So as to answer questions concerning accuracy (RQ 6: "What can we learn about students' texts with the help of a corpus analysis? How accurately do 8th-graders use simple target language forms (PRC, EXP (there/it), NEG, PLU)?" and "What developmental patterns emerge?"), a coding system for correct forms and errors had to be constructed (see Table 5).

As discussed previously in relation to corpus linguistics (2.4), error tagging is a special type of annotation which is associated with learner corpora and aims to serve language pedagogy. In error tagging, the different types of errors occurring in learner corpora are coded. Corpora annotated for learner errors can reveal the relative frequency of error types committed by learners of different proficiency levels (McEnery, Xiaio, & Tono, 2006).

In the tagging process I did not rely on any well-known corpus annotation systems (POS-tagging or semantic annotation), as my foci of interest were varied in terms of linguistic categories: one of the investigated forms (PLU) concerned morphology, whereas the others belonged to syntax (e.g., the word order in applying 'there is' or 'She is coming'), and I had not met an automatic coding for syntax. As I could not use automatic tagging, I created an own system for separating the four grammatical forms to be examined.

So as to acquire the ratio of correct forms in the total number of forms applied, separate codes were created for accurate and erroneous forms of the selected target language constructions. Either an acronym or the first three letters of the code identified the language form investigated: PRC stood for the present continuous verb form, EXP meant expletives, PLU designated plural noun forms, and NEG represented NEG. If the scrutinised language form was applied correctly, OK was pasted to the end of the code: thus, PRCOK (e.g., *The boy is playing with a cat*) represented a verb applied correctly in the PRC. Similarly, NEGOK stood for a negative construction used correctly (e.g., *There is no clock next to the TV*). On the other hand, if the investigated language form was erroneous, a symbol was created for the type of error and it was stuck to the end of the main code. Types of errors included, for example, the use of the simple verb form (code: SV) or present simple form (code: PS) instead of present continuous tense (code: PRC). Thus, PRCSV stood for an erroneous simple verb form used instead of present continuous tense (e.g., *he sit on the chair*), and PRCAUX meant a missing auxiliary when applying present continuous tense (e.g., *she coming into the room*).

Table 5: The coding system developed for analysing L2 English accuracy

| Codes for language structures in students' writings | | Explanation of the codes and examples |
|---|---|---|
| PRCOK | | present continuous correctly used<br>(e.g., The man is reading a book.) |
| PRCNOOK | | all five types of present continuous mistakes added up:<br>prcps, prcaux, prcing, prcun, prcneg |
| | PRCSV | simple verb form is used instead of present continuous<br>(e.g., instead of '*he is reading'* '*he read'* is used) |
| | PRCPS | present simple is used instead of present continuous<br>(e.g., instead of '*he is reading'* '*he reads'* is applied) |
| | PRCAUX | missing auxiliary from present continuous tense<br>('*is*' is missing from '*he is reading'*- e.g., '*he reading'*) |
| | PRCING | missing 'ing' from present continuous<br>(e.g., '*he is read'*) |
| | PRCUN | unnecessary use of present continuous tense<br>('e.g., the *TV is turning on'*) |
| | PRCNEG | erroneous use of present continuous negative<br>(e.g., '*he don't read*') |
| PRCPAST | | both correct and incorrect use of past continuous tense<br>(e.g., in the 1st picture he *was playing* with a cat/ he *was play with*) |
| EXPOK | | the singular expletive '*there is'* correctly used<br>(e.g., *There is* a clock next to the TV.) |
| EXPPLOK | | the plural expletive *'there are'* correctly used<br>(e.g., *There are* three plants in picture A.) |
| EXPNOOK | | both incorrect uses of the expletive added up:<br>expmis and expmix |
| | EXPMIS | missing expletive (e.g., 'in picture A *is* one plant, in B *are* two plants' instead of '*there is* one plant', etc.) |
| | EXPMIX | expletives mixed ('*It is* clock' instead of '*there is a* clock', or *'there is a* rainy day' instead of *'it is a* rainy day.') |
| PLUOK | | correct plural form of the noun<br>(e.g., *'two pictures'*) |
| PLUNOOK | | all incorrect forms of plural nouns added up:<br>plsng, pluun, pludbl |
| | PLUSNG | singular noun is used where plural is needed, 's' is missing at the end of the plural noun (e.g., *'two picture'*) |
| | PLUUN | unnecessary 's'<br>(e.g.,*' there is a windows'*) |
| | PLUDBL | double marking of plural in nouns<br>(e.g., *'childrens'*) |
| NEGOK | | negation correctly applied<br>(e.g.,*' there isn't a* plant') |
| NEGNOOK | | error in negation<br>(e.g., *'there is don't* flower*',* 'the clock *don't* on the desk') |
| OTHER | | other mistakes<br>(e.g., word order error, which is not attended to at present) |

During the encoding process I noticed that some of the errors appeared in abundance, others were scarce, so not all of the codes were equally useful in the final account. To give an example, past progressive (code: PRCPAST*:* in the first picture the girl was playing with a doll) was used by few students, but at its first appearance it could not be foreseen how frequently it would emerge, so it was included in the coding scheme.

Students committed several types of mistakes (e.g., word order /verb agreement errors) besides the errors concerning the four selected language forms investigated in this paper. At the time of tagging I thought it to be crucial to involve these additional mistakes in the count of errors, as RQ 7 investigated the relationship of points in criterion-based assessment with linguistic analyses of texts. As judges using the analytic scales took into consideration all kinds of errors, not only those of the four selected language forms, I presumed that all important additional mistakes also had to be tagged and counted, If I ignored these errors (word order /verb agreement, and the rest) in tagging, it would prevent me from the possibility of comparing the results of the two methods of assessment reliably.

Therefore, at the initial stage of the tagging process all errors other than the ones related to the selected forms were coded separately, as follows: word order error (WO), verb agreement error (VERBAE), missing verb (MISVER), the use of '*on* the picture' (ONPIC) for expressing '*in* the picture', the lack of the preposition *with* (WITHMIS) in '*play with'*, and vocabulary misuse (LANG). The incorrect use of articles was flagged in the OTHER category.

First, a specially designed XML tool similar to POS-tagging was used to separate selected language forms in the text, but later it was found that, in addition to the extremely time consuming tagging procedure, the coding programme had limitations in producing the necessary data. Let us see an example for the limitations.

If there were two or more mistakes in a word / expression, this system, which embraced coded language items by an initial and a closing tag, failed to work, as two different tags cannot be used in parallel. The sentence: **There isn't any plants,* was coded as:

<s>There <verbae>isn't</verbae> any <pluok>plants</pluok> </s>,

where <*s*> meant the beginning of a sentence, and <*/s*> stood for signalling the end, whereas <*verbae*><*/verbae*> showed the verb agreement error, and <pluok> and </pluok> embraced the correct PLU form of *plant*. Apart from the typing burden of the long tags for each mistake, in case of two different foci of investigation occurring in one word (e.g., *isn't*), one of them could not be signalled. Here, I did not flag the occurrence of NEG, as I had to choose between tagging the mistake of using PLU or signalling NEG.

Due to these limitations a new encoding procedure was chosen: all students' texts were copied into a Microsoft Office Excel table so that each line contained only one clause with a main verb of interest for our investigation. For example:

In the first picture there's a plant next to the girl,

but in the second there isn't a plant.

Then, each line (N=3,348) was scrutinised and coded for the selected grammatical features (22) one by one, both correct and erroneous forms being flagged (Appendix E on CD). Each error type was registered and counted in separate columns (e.g., in Table 6). Later, for the statistic analysis, the incorrect forms of each language form were added up. In other words, if a student made two PRCSV and three PRCAUX mistakes, it was summarised in column PRCNOOK as five; likewise, EXPMIS and EXPMIX mistakes were counted together in EXPNOOK. Correctly applied EXPs, either singular or plural, were also added. Thus, the totals of correctly and incorrectly used language forms were used in calculating ratios.

I will illustrate the encoding process by three examples (Tables 6, 7 and 8), first by showing the tagging of the second clause of the sentence above "*but in the second there isn't a plant."* (Table 6): The EXP is applied correctly in this clause, so EXPOK is ticked; and as it is in negative form, NEGOK is also flagged.

Table 6: Tagging of the clause "*but in the second there isn't a plant."*

| **PRC OK** | **PRC NOOK** | PRC SV | PRC AUX | PRC ING | PRC UN | PRC NEG | PRC PAST | **EXP OK** | **EXP PLOK** | **EXP NOOK** | EXP MIS | EXP MIX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | 1 | | | | |

| **PLU OK** | **PLU NOOK** | PLU SNG | PLU UN | PLU DBL | **NEG OK** | **NEG NOOK** | LANG | WO | ON PIC | WITH MIS | MIS VER | VERBAE | **OTHER** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | 1 | | | | | | | | |

The erroneous sentence * "*The boy play the cat."* was tagged like this (Table 7):

The present continuous form was incorrectly substituted by a simple verb form of *play,* so the category PRCSV was ticked. When summing up mistakes, it was counted in the PRCNOOK column. The preposition *with* was also missing in the expression of *play with the cat,* this was first registered in the WITHMIS category, then, it was placed in OTHER.

Table 7: Tagging of the sentence "*The boy play the cat."*

| **PRC OK** | **PRC NOOK** | PRC SV | PRC AUX | PRC ING | PRC UN | PRC NEG | PRC PAST | **EXP OK** | **EXP PLOK** | **EXP NOOK** | EXP MIS | EXP MIX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **1** | 1 | | | | | | | | | | |

| **PLU OK** | **PLU NOOK** | PLU SNG | PLU UN | PLU DBL | **NEG OK** | **NEG NOOK** | LANG | WO | ON PIC | WITH MIS | MIS VER | VERBAE | **OTHER** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | 1 | | | 1 |

Another example illuminates the coding procedure in case of several erroneous forms (Table 8): *"*"*On picture A are on the wall two pictures."*

*On picture A* was erroneous, as the preposition *on* was incorrect, it was coded as OTHER mistake. The depiction of the location of pictures (*are on the wall two pictures)* was placed in the middle of the sentence instead of the possible beginning or end positions, this word order mistake was first ticked in the WO column, then placed in the OTHER category. They were counted as two errors. For the missing EXP '*there'* (*are two pictures)* EXPMIS was ticked, which then was placed in the EXPNOOK, *two pictures* was a correctly implemented *plural form* of *picture,* thus labelled as PLUOK.

Table 8: Tagging of the sentence *"On picture A are on the wall two pictures."*

| **PRC OK** | **PRC NOOK** | PRC SV | PRC AUX | PRC ING | PRC UN | PRC NEG | PRC PAST | **EXP OK** | **EXP PLOK** | **EXP NOOK** | EXP MIS | EXP MIX |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | 1 | 1 | |

| **PLU OK** | **PLU NOOK** | PLU SNG | PLU UN | PLU DBL | **NEG OK** | **NEG NOOK** | LANG | WO | ON PIC | WITH MIS | MIS VER | VERBAE | **OTHER** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | | | | 1 | 1 | | | 1 | ll |

As the scope of this study did not permit the analysis of the abundance of data gained from this detailed tagging, the main RQs were focused on, and OTHER mistakes were counted in sum (LANG, WO, ONPIC, WITHMIS, MISVER, & VERBAE). Table 5 illustrates the codes that finally remained.

Counting additional mistakes (OTHER) produced surprising results in statistics. Unfortunately, only at the stage of investigating the relationship between criterion-based and corpus-based measurement by a multiple regression analysis (4.8.3) did the first doubt emerge concerning the reliability of this coding method. Later, when I analyzed a few selected scripts qualitatively (4.8.5), it became crystal clear what caused the disparity in results. The mystery will be unveiled in those sections.

#### 4.7.3.3 Defining the ratio of accuracy

When the coding process was completed, I needed to identify the number of times the selected language forms occurred in the text, both in correct and erroneous forms. In this phase, all correctly used PRC forms, EXPs, PLUs and NEGs were summed respectively, as well as incorrect versions of the same constructions.

The method for defining the ratio of accuracy of certain language forms was based on the approach of obligatory occasion analysis (OOA, Brown, 1973). Although originally *OOA* was used mainly for assessing mother tongue oral performances, it could be adapted to conveniently measure the accuracy of L2 written performances. The obligatory minimum of occurrences of a scrutinised language form was set differently from the original *OOA* method, though. In this study, if an investigated language form was used once, either correctly or erroneously, it was counted, whereas in *OOA* fewer than three occurrences of forms are usually excluded from the examination. In our case PLUs and NEGs were elicited by the task in fewer numbers, thus the exclusion of one or two occurrences would have limited the scope of the investigation.

In oral performances the procedure of calculating the ratio of accuracy implies dividing the number of correct suppliances of a language form by total obligatory occasions. All occasions where the use of a certain language form is required in the context are called obligatory occasions. Following this method, I divided the number of correct applications of a certain language form by all trials of the respective form. To provide an example, if the student used EXPs six times, but three of them were incorrect, the ratio was 0.5 (3:6).

Although, based on the prompts, an optimal number of applications of certain language forms could have been expected, and students could have been penalised for not using them in the required number, this approach would have contradicted the main credo of communicative language teaching and testing (Bachman & Palmer 1996; Breen & Candlin, 1980; Hymes, 1972; Widdowson, 1978). Test-takers would have been punished for not doing something instead of being appreciated for their accomplishments. Thus, based on the theory of communicative language teaching, the task achievement rubrics catered for rewarding all required items covered in the composition, while the grammatical forms to be applied depended on the students' choices.

When encoding was completed and the theoretical decisions on processing methods were made, SPSS 14.0 statistical software was applied to gain exact information on sums, ratios, frequencies and correlations in the dataset. Nevertheless, based on the assumption that statistical data illuminate only general features of students' L2 competence, I also investigated a sample of writings (5 scripts) qualitatively providing a detailed analysis of their task achievement, vocabulary, accuracy, and text cohesion.

#### 4.7.3.4 Measuring vocabulary features

For scrutinising students' vocabulary I used Laufer and Nation's (1995) *Lexical Frequency Profile*. It is based on the GSL (West, 1953), and it shows the relative proportion of words from different frequency levels in any sample of writing.

First, I introduce the software designed for measuring various features of vocabulary, then, I will explain the concepts used in the description of the measurement by examples.

Heatley *et al.* (2002) developed the software for Windows-based PCs, named RANGE, which can do the analysis based on three to fourteen pre-set word lists (WLs), and it also provides the possibility for the development of own WLs. The first of the fourteen lists (BASEWRD1.txt) includes the most frequent 1,000 words of English. Base list two includes the second most frequent one thousand words, and the third one comprises words not in the first 2,000 words of English but which are frequent in secondary school and university texts from a wide range of subjects. All of these base lists include the base forms of words together with derived forms. The first 1,000 words thus consist of around 4,000 forms, called types. The types 'come/comes/coming/came' are all members of the 'come' family. The sources of these lists are the GSL of English Words by West (1953) for the first 2,000 words, and The Academic Word List by Coxhead (1998, 2000) containing 570 word families.

To provide an example, I will show a student's results from our cohort (Table 9) on whose text five WLs of Range Test were run (Nation, 2005). In the WL column, one, two, three, etc., refer to each of the base lists. The table includes types, tokens and families of lexical / content words and function words (articles, auxiliaries, prepositions, pronouns), and it also shows the percentage of these in each WL.

Table 9: Tokens, types and word families in a student's writing

| WORD LIST | TOKENS/% | TYPES/% | FAMILIES |
|---|---|---|---|
| One | 147/92.45 | 51/83.61 | 49 |
| Two | 8/ 5.03 | 6/ 9.84 | 6 |
| Three | 1/ 0.63 | 1/ 1.64 | 1 |
| Four | 1/ 0.63 | 1/ 1.64 | 1 |
| Five | 1/ 0.63 | 1/ 1.64 | 1 |
| Not in the lists | 1/ 0.63 | 1/ 1.64 | 1 |
| **Total** | **159** | **61** | **58** |

This student used 61 different words in a 159-word-long text, where 51 of the running words belonged to the first word list, making up 83,6 percent of the whole script, whereas about 16 percent of lexemes came from the second to fifth lists, in other words, the less frequently used

vocabulary groups. Words counted in the 'not in the list' category either come from a less frequent WL, or have spelling mistakes, and thus the programme automatically placed them here.

Based on the comprehensive measures of lexical richness defined by Read (2000) the following four calculations were done on a small sample of the texts: (1) Lexical variation / type-token ratio (TTR), (2) variation of lexical words ($TTRL_{)}$, (3) lexical density (LD), and lexical sophistication (LS). I will introduce these terms next:

Lexical variation equals the type-token ratio: the number of different words in the text is divided by the total number of words used in the text. Thus the most frequent word 'the' is counted as one type, but as many tokens as the number of its appearance in the writing.

A second type of calculation was also done to calculate the type-token ratio of the lexical words (as distinct from grammatical or function words) in the text. In Laufer's study (1998, cited in Read, 2000, p. 203) a lexeme is a "single lexical item which may consist of more than one form." She identified inflected forms of the verbs as one lexeme. Thus come/comes/coming/came is identified as the same lexeme. In the Range programme these verb forms belong to the same family, so they are counted as one family, but four types. The highest possible type-token ratio is one indicating that all lexemes used in the text come from different word families. The widest range of different words the text contains, the nearest this figure is to one; a low number indicates that the writer has relied on a small stock of words whose forms are frequently changed.

The third type of investigating vocabulary richness is the measuring of LD. In this process the total number of lexical words is divided by the total number of words in the text. Lexical sophistication is the fourth measure; it shows the rate of types from WL 1 to those belonging to less frequent WLs in the text (WLs 2, 3, & 4).

Due to the limitations of using this method for measuring vocabulary features of texts of unequal length (Read, 2000; Richards, 1987; Vermeer, 2004), I intend to do the calculations only in a restricted number of writings (five). Then, I will also qualitatively analyze those texts, thus quantitative and qualitative methods of measurement can be compared for the purpose of triangulation.

## *4.8 Results of the writing test*

### 4.8.1 Results by criterion-based assessment

#### 4.8.1.1 Overall results by skills compared with the results of a national survey

In this section I will answer the following RQs:

RQ 1: How does Baranya students' writing proficiency compare with their other EFL language skills?

RQ 2: How does Baranya primary-school learners' L2 proficiency compare with the L2 proficiency of the same age-group in a national sample?

In order to better illuminate results in writing, I will start by presenting data of all language skill areas measured in the Croatian-Hungarian research project (Mihaljević-Djigunović et al., 2006). Thus, we can proceed from a general view on students' L2 English proficiency towards the narrower focus on their writing achievements. Accordingly, first, let us see the mean scores students attained for their L2 performance in all areas of investigation. Table 10 illustrates the summary statistics of participating students' L2 performance for various skills and for two total scores. TOTAL includes data on listening, reading and writing, whereas TOTALS includes results on the speaking test in addition to the first three skills. N represents the number of students whose data proved valid while running statistics.

Table 10: Summary statistics for students' L2 performance by skills, mean, and standard deviation (Source: Mihaljević-Djigunović, J. et al., 2006. p. 179)

| | N | Max. | Mean | Mean (%) | SD | SD (%) |
|---|---|---|---|---|---|---|
| LISTENING | 216 | 20 | 17.62 | 88.10 | 3.43 | 17.15 |
| READING | 231 | 46 | 26.94 | 58.56 | 11.67 | 25.37 |
| WRITING | 231 | 32 | 17.26 | 53.93 | 11.03 | 34.47 |
| SPEAKING | 35 | 48 | 29.09 | 60.60 | 14.21 | 29.6 |
| PRAGMATICS | 230 | 12 | 7.18 | 59.83 | 2.68 | 22.33 |
| TOTAL | 202 | 98 | 62.22 | 63.48 | 24.62 | 40 |
| TOTALS | 25 | 146 | 89.00 | 60.95 | 39.01 | 26.7 |

Figure 3 also illustrates the results in percentages. The overall mean of scores for three skills is 62.22, equalling 63.48 percent achievement. Investigating the table and figure by skills, we can observe that the listening task was the easiest one, as the mean is 88.1 percent of the maximum score, and standard deviation is low (17.15%). This means that there were no substantial differences in participants' listening scores, and a ceiling effect can be observed. The low standard deviation also indicates that the listening tests failed to place participants along a scale, as their scores were skewed to the right.

Reading, pragmatics, and speaking performances fall near each other reaching a mean of about 60 percent. Students' mean scores in reading are 26.94 (58.56%) with 25.37 percent standard deviation. Students also reached a 59.83 percent mean of correct answers in pragmatics, though in this area scores are not as varied (SD=22.33%) as those of reading. Speaking achievements of the random sample also reach 60.6 percent, although with higher standard deviation of 14.21 (29.6%).

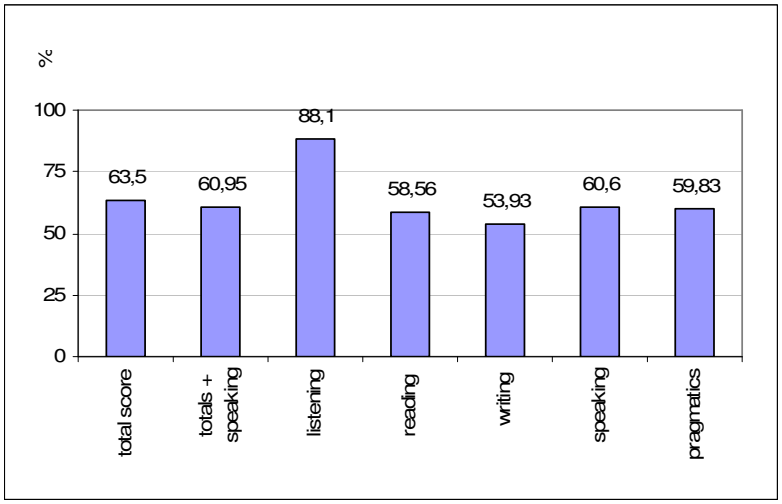



Figure 3: L2 summary statistics by language skills

The whole cohort's mean of writing, as already anticipated from results of previous surveys (Bors, Nikolov, Pércsich, & Szabó, 1999; Bukta & Nikolov, 2002; Nikolov, 2003; Tagányiné, 2001a, b; Sturman, 2001), is the lowest number: 53.93 percent, and the differences between scores are the highest here (SD=11.03; 34.47%). Thus, the answer to RQ 1 is that students' writing skill is the least developed one among the four skills. Nevertheless, the difference between writing and the other skills is not as high as in previous studies; it reaches only 6 percent less than reading, pragmatics or speaking. This may be partly explained by the slight overrepresentation of city-dwellers (Schools 1, 2, & 4) learning in relatively good socio-educational circumstances. In two of these institutions teacher trainees practise for their future jobs, so the teacher-trainers working there need to be up-to-date concerning educational methodology, and they probably develop all four skills in a more balanced way than their peers in other schools.

As these results can be related to data of the 2002 national survey (Csapó & Nikolov, in press, Table 11), RQ 2: "How does Baranya primary-school learners' L2 proficiency compare with the national L2 proficiency of the same age-group?" can be answered. Comparisons can

be made, as tasks and assessment methods were identical in the two projects. All differences between the results of the two cohorts are significant ($p<.01$); listening and reading skills show similar levels in the national and Baranya studies. In the national survey the poorest achievements characterised students' writing proficiency, the national mean (2002) of 3,653 8th-graders being as low as 33.7 percent. The writing results (53.93%) of students tested in this survey surpass the findings of the national survey by far. This may be partly due to the slight overrepresentation of students of high SES learning in city-schools in this study more details on whom will be provided in the next section. The other reason for this great disparity may lie in the compulsory participation of schools in the national survey, whereas in the Baranya research I needed to persuade teachers to accept the challenge of participation. The ones who took it up were teachers who must have been more self-confident due to their previous results in teaching. Thus, teachers' motivation played an accountable role in participating in our project and may partly explain better results.

Table 11: A comparison of the results of Grade 8 in the national survey (2002) and the results of this study

| | National survey (N=3,653) | | Baranya (N=231) | |
|---|---|---|---|---|
| Skill | Mean (%) | SD (%) | Mean (%) | SD (%) |
| Listening | 85.66 | 16.67 | 88.1 | 17.15 |
| Reading | 53.19 | 25.89 | 58.56 | 25.37 |
| Writing | 33.70 | 27.58 | 53.93 | 34.47 |

After the investigation of achievements in all skill areas of the whole cohort, I will turn my attention to students' writing performances, examining them by institutions.

#### 4.8.1.2 Means of total scores by school

In this section I aim to find answers to the following RQs:

RQ 3 What level of writing proficiency do Baranya primary school students reach in English by the end of their primary education?

RQ 4 What is the relationship between students' writing proficiency and the curricular achievement targets?

RQ 5 What is the relationship between students' socio-educational background and their writing proficiency?

One point of interest concerns students' L2 writing competence, the other focal point requires the comparison of their proficiency measures with curricular achievement targets. The focus

of RQ 3, EFL writing proficiency, will be investigated by two methods: rater assessment of students' tests, which is a criterion-based, qualitative way of evaluation, and secondly, by computerized text analysis, which is intended to be an objective, quantitative measuring instrument. Then, the two separate measurements, that is, the rater's scores and data on language use will be correlated, thus the reliability of expert assessment will be monitored.

First, I aim to show what criterion-based assessment tells us: I will explore the results of all schools, based on the total mean scores attained for their students' writings, then, totals will be broken down to the mean scores of task achievement, vocabulary, accuracy, and cohesion, and these four areas will be compared. Table 12 shows the summary statistics of scripts by school concerning the number of participants, range, minimum score, maximum score, mean, and standard deviation.

Table 12: Descriptive statistics concerning total scores on L2 writing broken down by school

| | | Total scores | | | | |
|---|---|---|---|---|---|---|
| Schools | N | Range | Min | Max | Mean | SD |
| 1 | 55 | 31 | 1 | 32 | 24.25 | 7.397 |
| 2 | 44 | 28 | 4 | 32 | 26.09 | 5.202 |
| 3 | 10 | 23 | 2 | 25 | 11.40 | 8.746 |
| 4 | 26 | 31 | 0 | 31 | 22.50 | 8.320 |
| 5 | 13 | 29 | 2 | 31 | 20.38 | 11.117 |
| 6 | 18 | 13 | 0 | 13 | 5.28 | 3.723 |
| 7 | 20 | 23 | 0 | 23 | 3.95 | 5.735 |
| 8 | 11 | 19 | 0 | 19 | 5.36 | 5.954 |
| 9 | 11 | 21 | 0 | 21 | 10.27 | 8.088 |
| 10 | 20 | 19 | 0 | 19 | 9.20 | 5.699 |
| 11 | 3 | 10 | 0 | 10 | 3.33 | 5.774 |

As illustrated in Table 10, the average score of all L2 writings was 17.26 out of the maximum 32, which equals 54 percent performance. From Table 12 and Figure 4 we can see that this result is not balanced for all institutions. Four schools (1, 2, 4, and 5) outperformed the rest, with 64-82 percent achievement on writing tests. School 2 excelled by an average of 26.09 scores (82%), and School 1 also surpassed the rest by a 76 percent mean. As 99 students (43%) out of 231 attend these two schools, no wonder that the results of Baranya exceed that of the national survey taken in 2002. Nevertheless, not all institutions can boast of their students' achievements. Four other schools (6, 7, 8, and 11) did not even reach six scores out of 32. This represents less than one fifth of the maximum achievement. Schools 9, 10 and 3 also performed weakly by achieving 9.2-11.4 means, equalling about one third (29%-36%) of

the best possible result. Standard deviation is the highest (SD=11.117) in School 5, and it is also very high in Schools 3 (SD=8.746), 4 (SD=8.32), and 9 (SD=8.088).

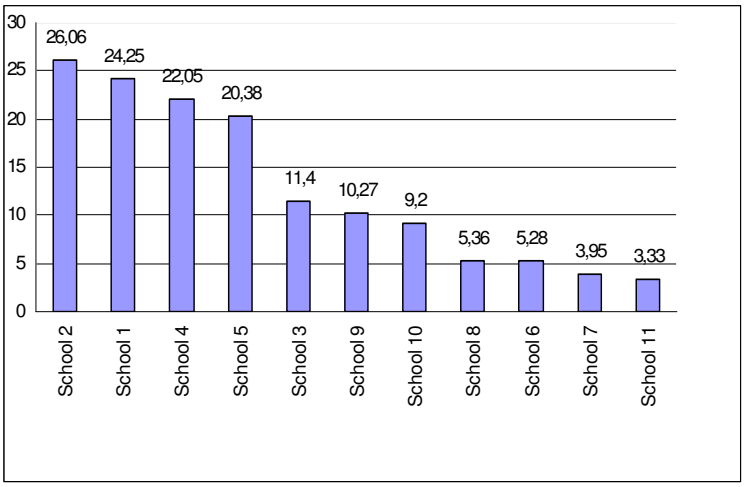



Figure 4: Mean of total scores on L2 writing by school

Returning to and surveying Table 1 under 4.5.1, these results are not very surprising if we consider the socio-educational background of the institutions. Out of the best-achieving four schools three are located in the large town (Schools 1, 2, & 4), and their students' SES exceeds that of all other participants'. The students of these institutions actively and successfully take part in county and national EFL competitions. The fourth (5) school is situated in a town near the Croatian border. It is characterised by well-developed international tourism, where high FL proficiency can be easily used in everyday situations. The institutions falling in the lowest achievers' category belong to the town (6, 7) or village (8, 10) group, but one of them (11) is a city school, with especially low SES children. The other two low-achieving educational institutions belong to the village-dweller (9) and large-town-institution (3) category. Regarding the large-town-school (3), a similarly low SES and lack of motivation of their students provides challenges in their everyday work as the one mentioned in connection with School 11.

Other background variables, such as the starting age of L2 learning, the number of L2 lessons per week are also different in these schools, but this study does not aim to investigate time factors. Children's SES is latently involved in their achievements, as the need for the selection of a prestigious institution for their children arises mostly with parents of high SES. An abundance of studies finds that children's SES has a strong influence on their access to FL learning (Vágó, 2007), which, in turn, affects language proficiency levels measured in these

investigations (Andor, 2000; Bors, Lugossy, & Nikolov, 2001; Bors, Nikolov, Pércsich, & Szabó, 1999; Bukta & Nikolov, 2002; Csapó, 1998; 2002).

In this section I have compared means of total scores on writing regarding participating schools. The results show huge differences between schools according to their location: large town institutions where students' SES is favourable (1, 2, & 4) performed extremely well, two large town schools with low SES pupils (3, & 11) demonstrated poor language achievement levels, whereas students of town schools (5, 6, & 7) displayed large variance in their performance. Village institutions (8, 9, & 10) generally had poor results, but within this category also large differences were shown.

The next section intends to take a close look at the mean scores of institutions along the four criteria of assessment: task achievement, vocabulary, accuracy, and cohesion; and investigates the relationship between the achievements on these four criteria. The results will also be related to the *NCC* requirements.

#### 4.8.1.3 Mean scores of task achievement, vocabulary, accuracy, and cohesion by school

Let us overview the mean scores in the four areas (Figure 5) in general first, then, a detailed analysis follows. The dotted line indicates the pass-level (3 scores), which approximately corresponds to the minimum requirement prescribed in the *NCC,* that is, covers the competences described by A1 level in the *CEFR* (Council of Europe, 2001). The best-achieving schools as regards mean scores (1, 2, 4, & 5) have the highest scores in all four areas, task achievement reaching the top-score among the four performance criteria (Table 14). Students of Schools 3, 9, and 10, displayed very similar language levels, which nearly reaches the requested minimum, level A1. Participants from Schools 6, 7, 8, and 11 performed under the required minimum achievement level.

At first glance it is evident that the four sub-scores constituting the total score are in strong relation with the total score, and also with one another. Vocabulary results (Table 15) are close to those of accuracy, except for School 3. Means for accuracy / grammar performances (Table 16) go hand in hand with cohesion (Table 17), cohesion being a bit higher in most institutions with the exception of School 6.

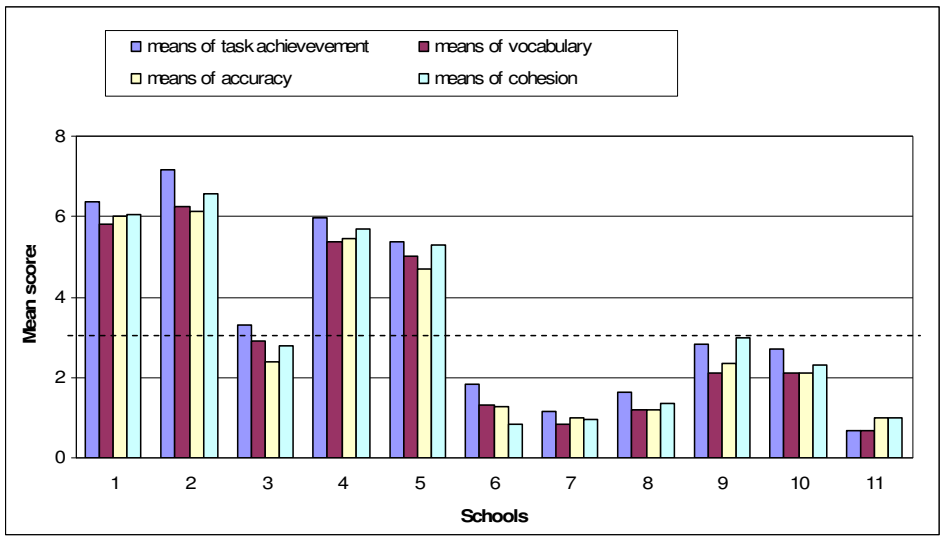



Figure 5: Mean scores of task achievement, vocabulary, accuracy, and cohesion by school

These statements made at first sight were tested by investigating correlations between the results along the four criteria (Table 13). Correlations between all criteria proved significant ($p<.01$) and extremely high, ranging between .942 and .986. The strongest relationships occur between the total score and each of the four criteria, all of them being above 98 percent. Examining the criteria one by one, vocabulary results show the highest correlation with the total score (.986), then come cohesion (.983), task achievement (.982) and accuracy (.980), all of them representing very high correlation. Task achievement and vocabulary are also in close relation, which is reflected in the .966 correlation. Vocabulary also strongly correlates with accuracy and cohesion (.956). Even the lowest correlation between task achievement and accuracy is very high (.942). These results may reflect that most students' L2 skills in various linguistic areas develop in tandem, or, at least the assessor judged so. We have seen a sample of a top achievement, and Appendix F shows a performance at the other end of the linguistic achievement scale.

Table 13: Correlations between mean scores of task achievement, vocabulary, accuracy, and cohesion

| N=231 | | Task | Vocabulary | Accuracy | Cohesion |
|---|---|---|---|---|---|
| Vocabulary | Pearson Correlation | .966(**) | 1 | | |
| Accuracy | Sig. (2-tailed) | .942(**) | .956(**) | 1 | |
| Cohesion | | .951(**) | .956(**) | .955(**) | 1 |
| Total Score | | .982(**) | .986(**) | .980(**) | .983(**) |

** Correlation is significant at the 0.01 level (2-tailed).

Having compared the mean scores on task achievement, vocabulary, accuracy and cohesion in general, I aim to demonstrate the results in the four performance criteria attained by institutions in more detail.

As for task achievement (Table 14 & Figure 6), School 2 outperformed the rest by a 7.16 (90%) result, which is the highest mean score among all institutions along all criteria. School 1 achieved a similarly high mean by 6.36 (80%). This reflects that students' texts covered nine or ten things relevant to pictures A and B. School 4 with its 75 percent performance and School 5 with 67 percent are also among the top achievers. Their scripts described either seven to eight items in the pictures relevantly, or more items, but partly relevantly to the prompt. School 3 is just in (3.3=41%) the accepted band, whereas Schools 9 and 10 are just below the pass-level score (2.82< 3, 35%; 2.7< 3, 34%).

Table 14: Descriptive statistics concerning task achievement scores on L2 writing broken down by school: N, range, minimum, maximum, mean, standard deviation

| School | N | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| 1 | 55 | 1 | 8 | 6.36 | 1.947 |
| 2 | 44 | 2 | 8 | 7.16 | 1.275 |
| 3 | 10 | 1 | 6 | 3.30 | 2.312 |
| 4 | 26 | 0 | 8 | 5.96 | 2.341 |
| 5 | 13 | 1 | 8 | 5.38 | 2.785 |
| 6 | 18 | 0 | 5 | 1.83 | 1.150 |
| 7 | 20 | 0 | 7 | 1.15 | 1.694 |
| 8 | 11 | 0 | 5 | 1.64 | 1.502 |
| 9 | 11 | 0 | 7 | 2.82 | 2.442 |
| 10 | 20 | 0 | 5 | 2.70 | 1.559 |
| 11 | 3 | 0 | 2 | .67 | 1.155 |

The rest of the institutions (6, 7, 8, 11) representing 52 students (23% of the cohort), belong to the under-achiever category with results ranging between 8 percent and 23 percent. According to the descriptors, their texts inform the reader on two or three things relevant to the pictures, or relate to more items but only partly relevant. Scripts falling in this band usually describe either picture A, or B. SD is the highest in Schools 5 (SD=2,785) and 9 (SD=2,442).

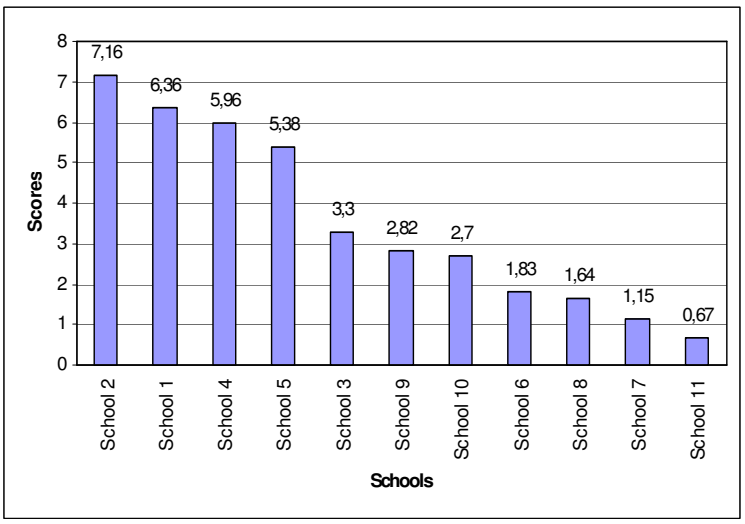



Figure 6: Means of task achievement scores by school

Table 15 and Figure 7 report on vocabulary means: here again Schools 1 (5.82=73%) and 2 (6.25=78%) outperformed the rest and standard deviation is relatively low, especially for School 2 (SD=1.26, 16%). The average results of Schools 4 (5.38=67%) and 5 (5.00=63%) fall above the middle band (Band 4), which represents a wide scale of vocabulary that is mostly appropriate to the task. Students of Schools 6 (1.33=17%), 8 (1.18=15%), 9 (2.09=26%), and 10 (2.1=26%), showed a limited scale and choice of vocabulary on average, or the words they applied were often inappropriate in the context. The participants of School 3 performed poorly (2.9=36%), but their average result still falls in the pass-band.

Table 15: Descriptive statistics concerning vocabulary scores on L2 writing broken down by school: N, range, minimum, maximum, mean, standard deviation

| School | N | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| 1 | 55 | 0 | 8 | 5.82 | 1.775 |
| 2 | 44 | 2 | 8 | 6.25 | 1.260 |
| 3 | 10 | 1 | 6 | 2.90 | 1.792 |
| 4 | 26 | 0 | 8 | 5.38 | 2.021 |
| 5 | 13 | 1 | 8 | 5.00 | 2.739 |
| 6 | 18 | 0 | 3 | 1.33 | 0.907 |
| 7 | 20 | 0 | 5 | 0.85 | 1.226 |
| 8 | 11 | 0 | 4 | 1.18 | 1.328 |
| 9 | 11 | 0 | 4 | 2.09 | 1.758 |
| 10 | 20 | 0 | 4 | 2.10 | 1.210 |
| 11 | 3 | 0 | 2 | 0.67 | 1.155 |

The poorest knowledge of vocabulary was demonstrated in Schools 7 (0.85=11%) and 11 (0.67=8%), these averages show that vocabulary points achieved were near zero. In School 11 the highest score received was two out of eight, whereas in School 6 the best achievement was

three scores out of the maximum of eight. Considering that the achievement of three scores means the minimal acceptable performance, the results indicate large differences between schools: students in five of the institutions performed well, two are near the pass-line, and four of them are lagging far behind.

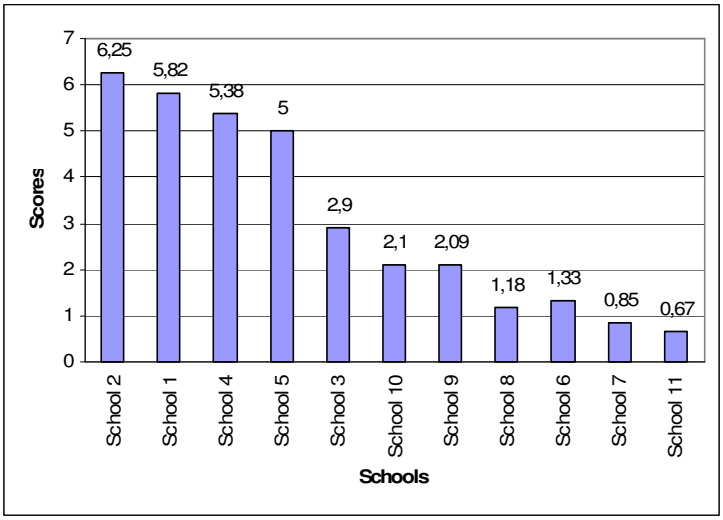



Figure 7: Means of vocabulary scores by school

Accuracy means presented in Table 16 and Figure 8, are undeniably characterised by the same variance as the previously described areas. The performance of the cohort is split in two clearly separable language proficiency levels when considering group standards; the upper level covers Band four (Schools 1, 2, 4, & 5) with achievements ranging between 59 and 77 percent, while the groups producing lower levels are mostly situated in the second band (Schools 6, 7, 8, & 11). Their results distributed in a narrower range between 13 percent and 16 percent. Some of the schools approximate the lower score of band three by producing 26-30 percent accuracy means (Schools 3; 10; 9). This shows that there are large differences between students' L2 proficiency levels, and the average performance of 54 percent covers huge differences.

Table 16: Descriptive statistics concerning grammar/accuracy scores on L2 writing broken down by school: N, range, minimum, maximum, mean, standard deviation

| School | N | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| 1 | 55 | 0 | 8 | 6.02 | 1.958 |
| 2 | 44 | 0 | 8 | 6.14 | 1.549 |
| 3 | 10 | 0 | 6 | 2.40 | 2.119 |
| 4 | 26 | 0 | 8 | 5.46 | 2.024 |
| 5 | 13 | 0 | 8 | 4.69 | 2.720 |
| 6 | 18 | 0 | 3 | 1.28 | 1.074 |
| 7 | 20 | 0 | 6 | 1.00 | 1.589 |
| 8 | 11 | 0 | 5 | 1.18 | 1.662 |
| 9 | 11 | 0 | 5 | 2.36 | 1.859 |
| 10 | 20 | 0 | 5 | 2.10 | 1.586 |
| 11 | 3 | 0 | 3 | 1.00 | 1.732 |

Having a closer look at accuracy means by institutions, we can see that School 2 is again the top achiever with 6.14 (77%) means, while Schools 1 (6.02=75%) and 4 (5.46=68%) also belong to the top party. School 5 approaches the best achievers with an average score of 4.69 (59%). It is also obtrusive that all other schools have very low results concerning this criterion. Schools 3 (2.4) and 9 (2.36) reach (30%), which can be interpreted as the lowest achievement in the pass band. Schools 6 (16%), 7 (13%), 8 (18%), 10 (26%), and 11 (13%) belong to the underachiever group concerning target language grammar results. This means that in these poor compositions a lot of mistakes occurred, and only part of the text was comprehensible. SD was highest in School 5 (SD=2.72; 34%) and lowest in School 6 (SD=1.074; 13%). The latter data illustrates that all 18 participating pupils attained scores ranging between 0 and 3, that is, they performed at a remarkably low level.

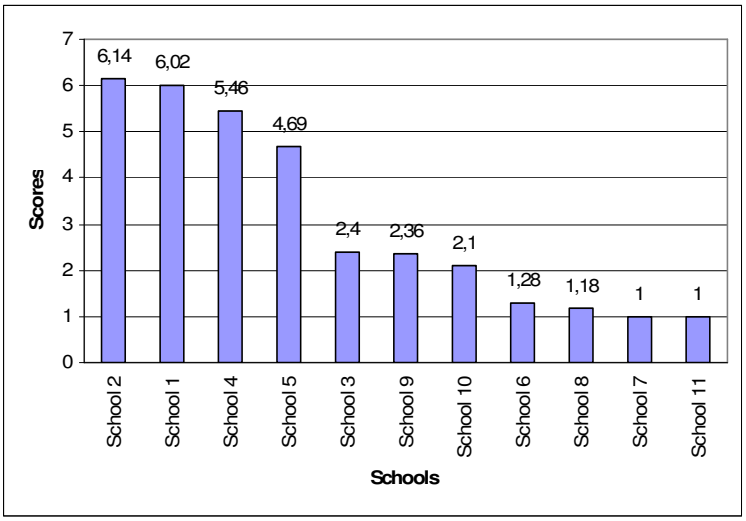



Figure 8: Means of accuracy scores by school

The same observations can be made about cohesion means, which are shown in Table 17 and Figure 9. The best results belong to Schools 2 (6.55=82%) and 1 (6.05=76%). These outstanding averages indicate that students' texts are well structured, logically built, and parts on different things are separated. The best pair of institutions is followed by Schools 4 (5.69= 71%) and 5 (5.31=66%). Their texts are less structured in general: there are some links between the sentences, and minimum three sentence types vary. Institutions 3 (2.8= 35%) and 9 (3=38%) can again be paired according to their performances, which result places them in the lower part of the pass-band.

Table 17: Descriptive statistics concerning cohesion scores on L2 writing broken down by school: N, range, minimum, maximum, mean, standard deviation

| School | N | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| 1 | 55 | 0 | 8 | 6,05 | 1,985 |
| 2 | 44 | 0 | 8 | 6,55 | 1,577 |
| 3 | 10 | 0 | 7 | 2,80 | 2,700 |
| 4 | 26 | 0 | 8 | 5,69 | 2,311 |
| 5 | 13 | 0 | 8 | 5,31 | 3,011 |
| 6 | 18 | 0 | 3 | 0,83 | 1,043 |
| 7 | 20 | 0 | 5 | 0,95 | 1,317 |
| 8 | 11 | 0 | 5 | 1,36 | 1,690 |
| 9 | 11 | 0 | 5 | 3,00 | 2,280 |
| 10 | 20 | 0 | 5 | 2,30 | 1,593 |
| 11 | 3 | 0 | 3 | 1,00 | 1,732 |

The mean scores for cohesion show very low achievements in Schools 6, 7, 8, 10, and 11. Means range between 0.83 and 2.3, equalling 10-29 percent average success in constructing a cohesive text. It indicates that in poorer writings the text was unstructured and incomprehensible, whereas in slightly better ones the same sentence type was repeated monotonously, or the text consisted of a sequence of sentences without cohesive devices. SD is highest in Schools 5 (SD=2.7), 3 (SD=3.011), and 4 (SD=2.311), whereas students in School 6 differed the least in their achievements (SD=1.043), their scores ranging between 0 and three.

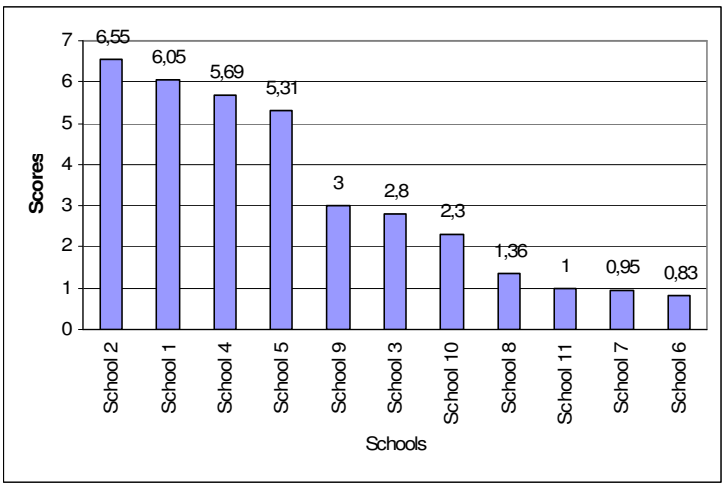



Figure 9: Means of cohesion scores by school

Summing up average school performances on the four criteria (task achievement, vocabulary, accuracy and cohesion), and relating the results to the requirements of the *Core Curriculum*, I can answer RQ 4 concerning the relationship between students' writing proficiency and the curricular achievement targets.

It is transparent from the data that on an institutional level only five schools met the requirements. Two of these five institutions performed between 73 percent and 90 percent on average (Schools 1, & 2; N=99), two of them exceeded 59 percent means, but achieved lower than 76 percent on the four linguistic criteria (Schools 4, & 5; N=39). The means of one institution were just about pass-level (School 3; N=10) with performance measures between 30 percent and 41 percent.

These five schools participated in the project with 148 pupils, so approximately 64 percent of the students met the curricular requirements calculated on an institutional level. Nevertheless, this statement is true in general only, as mean results always cover the discrepancies between high and low achievements. So, this number evidently includes some students who did not reach the prescribed minimum level (12 scores).

To get more accurate data on the number of pupils who performed at or beyond the required curricular level, I checked the database and identified scripts belonging in the top three bands regarding total scores. I also added a few writings of slightly poorer performance, those achieving between nine and eleven scores (Table 18), as during the qualitative analysis of scripts a slight adjustment of the estimated A1 level proved necessary (see 4.8.5). We can see that 149 students' scripts (64%) fall in the top three bands (12-32 scores), whereas the number of slightly weaker scripts (15) add another 7 percent to the sum, so approximately 71

percent of the participants reached the required language level. In the lowest bands 44 students received scores one to eight, and 21 of them (9%) was rated zero points.

Table 18: Number of students in the top three bands of the rating scale based on their total scores

| Total scores | Number of students | % |
|---|---|---|
| 25-32 | 88 | 38 |
| 17-24 | 37 | 16 |
| 12-16 | 24 | 10 |
| **12-32** | **149** | **64** |
| 9-11 | 15 | 7 |
| **9-32** | **164** | **71** |

As for under-achievers, their results are also distributed along a wide scale. School 9 produced percentiles between 26 and 38, whereas School 10 achieved results between 26 and 34 percent, showing lower performance on accuracy and cohesion than School 9. The lowest linguistic levels characterized Schools 6, 7, 8, and 11, as they achieved only 8-26 percent of the possible maximum mean scores. These results are definitely below the curricular requirements. The widest range (8-26%) was demonstrated in vocabulary means, while the lowest differences appeared in accuracy results (13-16%).

In this section I have discussed overall means for L2 performances and the results of schools separately along four criteria, based on rater assessment. I will turn to discuss the quantified data on language accuracy gained from counting students' errors and correct solutions concerning four selected language forms. Accuracy measures on the four encoded language forms will be calculated, as well as measures of overall accuracy.

### 4.8.2 Results on accuracy gained from a corpus analysis

#### 4.8.2.1 Results related to the whole corpus

In this section I aim to find answers to RQ 6: What can we learn about students' texts with the help of a corpus analysis? How accurately do 8$^{th}$ graders use simple target language forms (PRC, EXP, (there/it), NEG, PLU)? What developmental patterns emerge?

We have seen that the language forms for closer observation were chosen based on their frequency of emergence in students' texts. For defining the ratio of accuracy of these language forms the method of obligatory occasion analysis (OOA, Brown, 1973) was adapted to written language use. The procedure of calculating the ratio of accuracy is as follows: the

number of correct suppliances of a language form is divided by total obligatory occasions. Following this method, the ratio of accurate usage and all (correct/incorrect) usage was calculated. To illustrate it by examples: if a student used ten correct forms out of ten, the ratio was 10/10=1. If s/he succeeded in producing four correct present continuous verb forms and six incorrect ones, the ratio was 4/(4+6)=0.4, expressing that only four correct forms were applied out of ten trials, which also means that the ratio of success is 40 percent in this case. The minimum of correctness is zero, whereas the maximum is 1. In order to make results more comprehensible, they will be turned into percentiles. Before investigating the ratios of correctness regarding the chosen language forms, the number of students attempting to use these forms will be compared.

According to descriptive statistics (Figure 10), present continuous tense (N=216=94%) and PLU forms (N=200=87%) were used by the highest number of students, while EXPs were included in 176 writings (76%), and negative constructions were attempted by 134 participants (48%).

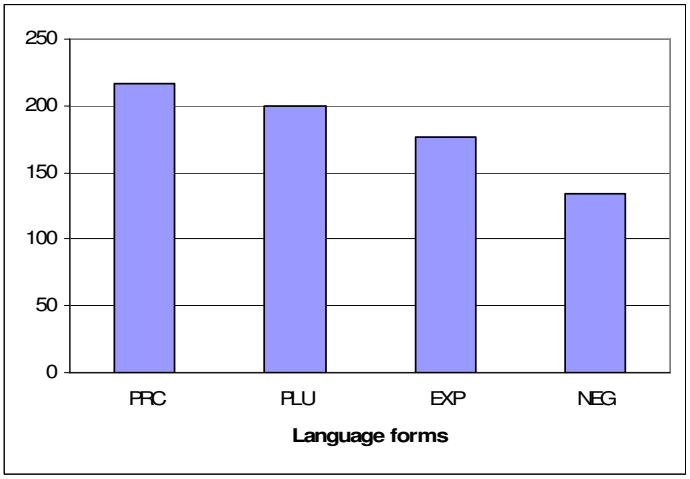



Figure 10: Number of students applying the analyzed language forms (N=231)

Figure 11 illustrates the total numbers (N=2923, correct and incorrect) of the four examined language forms (PRC, EXP, PLU, & NEG) used in the corpus. It demonstrates that PRC forms cover 52 percent of the usage of the four forms; EXP constructions add another 24 percent, whereas PLU is represented in nearly 16 percent, and NEG comprises almost 8 percent of the analyzed language forms.

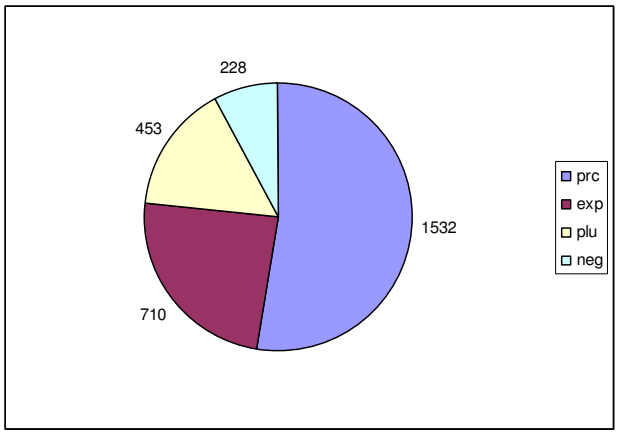



Figure 11: The total number of each of the four language forms (PRC, EXP, PLU & NEG) used in the corpus (total N=2923)

These ratios are similar to that of the ratio of language forms used in the model script (Figure 12), where ten PRCs, six EXPs, two PLUs, and one NEG comprise 52, 32, 11, and 5 percent of the examined forms respectively. The biggest difference occurred in the use of EXPs, as students applied them in smaller ratio in favour of using PLU and NEG.

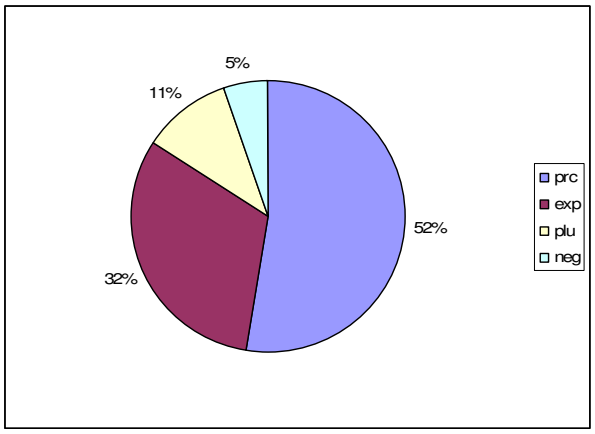



Figure 12: The percentage of the four language forms (PRC, EXP, PLU & NEG) used in the model text

Table 19 and Figure 13 show the summarized ratio of correctly applied language forms in the four investigated areas: present continuous tense, expletives, plural and negative, concerning all 231 writings. N stands for the total number of compositions out of all writings, in which the respective language form appeared. In the last line of the table 'valid listwise N' means that altogether 104 pupils (45%) applied all four investigated language forms in their compositions. Means indicate that NEG is the most successfully employed construction (0,79) in the whole corpus, PLUs come second (0.65), while EXPs reach only a 54 percent

correctness rate and PRC is the least accurately managed area with 48 percent successful trials.

Table 19: Descriptive statistics: ratio of correct language forms: PRC, EXPs, PLU, NEG, N, percentile compared to total N, mean, standard deviation

| | N | Percentile | Mean | SD |
|---|---|---|---|---|
| prc | 216 | 94 | .4775 | .43222 |
| exp | 184 | 80 | .5378 | .48197 |
| plu | 200 | 87 | .6477 | .41082 |
| neg | 134 | 58 | .7917 | .38314 |
| Valid N (listwise) | 108 | 47 | | |

There seems to be a contradiction between the attempt ratio and rate of success in the application of the analyzed language forms. While present continuous was the form attempted by most of the students (n=216=94%), it was the least successful one (48%). NEG, while being the least frequently attempted construction (n=134=58%) of the four evaluated forms, was applied most correctly (79%). PLU was used by 87 percent of the children (n=200) with relative success (65%). EXPs were attempted by 80 percent of the students with 54 percent accuracy. Standard deviation is high in three areas, thus, NEG seems to be the least challenging and best acquired language form in the corpus.

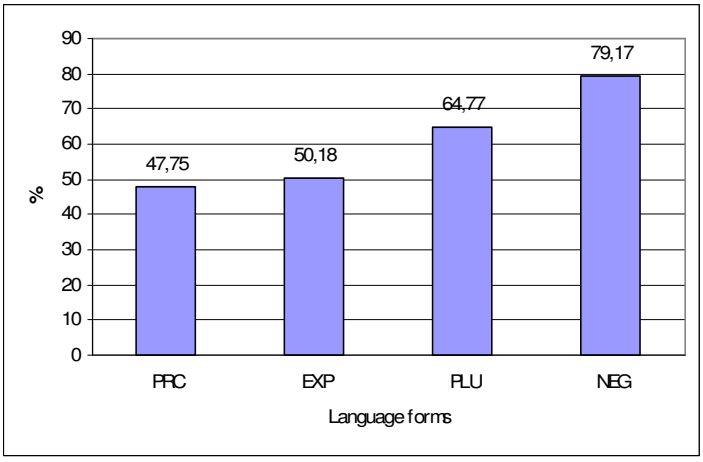



Figure 13: Ratio of correct language forms (PRC, EXP, PLU & NEG) for the whole corpus (percentile)

Nikolov and Krashen's (1997) study on Hungarian learners of the same age can be related to the present research. Although their data were collected in oral interviews, several similarities characterise the two projects. These similarities include the age group (7th and 8th-graders), and the venue for data collection, as the school is also involved in the present study where

Nikolov and Krashen gained their data (Nikolov, 2007, personal communication). The language forms investigated also overlap, and, in addition, the method of linguistic analysis (OOA) is the same. In Nikolov and Krashen's (1997, p. 198) research the following grammatical structures were investigated among others:

(1) plural (short plural only, as in dog*s*)

(2) progressive -ing (as in 'he is smok*ing* a cigarette')

(3) auxiliary (as in 'she *is* dancing', 'they *are* singing'.)

Accuracy results for grammatical morphemes are shown in Table20.

Table 20: Accuracy results for grammatical morphemes (Nikolov & Krashen, 1997, p. 199)

| grammatical morphemes | Experimental grade 7 | Exp. grade 8 | Traditional grade 8 |
|---|---|---|---|
| plural | 0.767 | 0.885 | 0.888 |
| progressive | 0.403 | 0.627 | 0.786 |
| auxiliary | 0.456 | 0.544 | 0.359 |

Although the main aim of their research was to investigate the influence of communicative, content-based approach to teaching EFL regarding students' accuracy, we can make conclusions from Nikolov and Krashen's (1997) data about students' acquisition order related to certain language forms. According to their data, PLU was the most accurately acquired form, thus, it can be deduced that it was the first acquired form of all. It was followed by the progressive marking '-ing', then the auxiliary 'be'. NEG was not explored in their study.

Comparing results it stands out that the PLU form was used with lower success (71%) among the 44 pupils of the same school (2) in the present study than in Nikolov & Krashen's (89% in grade 8). Regarding the present continuous form, which is constructed of the respective form of the auxiliary 'be' and the progressive form of the verb, participants of School 2 in this study achieved 62 percent accuracy. This includes applying 226 correct present continuous forms (PRCOK), 58 progressive markings (PRCAUX), 6 forms where the auxiliary 'be' was used (PRCING), and 67 simple verb forms (PRCSV), adding up to 357 attempts to use PRC. According to these data, progressive marking was 80 percent correct (PRCOK+PRCAUX), whereas the auxiliary 'be' was well used in 65 percent (PRCOK+PRCING) of PRC occurrences. As for the results of the oral interview, pupils produced 63-78 percent accuracy on progressive forms, and 36-54 percent on the auxiliary 'be', so in this respect the students of the present study slightly outperformed those of Nikolov and Krashen's.

These data cannot be easily compared, as the method of data collection was different (oral interview vs. written performance). Nevertheless, the order of acquisition seems to be similar

in the two studies with PLU forms produced more accurately than verbal constructions, and, also, with smart progressive marking preceding the correct application of auxiliary 'be' .

After investigating the ratio of correct language forms regarding the whole corpus, and comparing the results of School 2 with the data of a previous study (Nikolov & Krashen, 1997), let us examine the accuracy rate school by school.

#### 4.8.2.2 Corpus-based results of accuracy by school

The tables displaying descriptive statistics of school results convey how many students attempted to use the respective language constructions, show the mean of the correctness ratios for the four investigated language areas and the standard deviation of results. In institutions where the minimum ratio achieved by any participating pupil is zero, or the maximum is one, the tables will not contain the data on minimum or maximum ratios. In case of those schools where the minimum ratio calculated is more than zero, or the maximum ratio attained by any of the students is lower than one, these data will be shown in the tables.

The results of School 1 considerably exceed those of the whole corpus (Table 21), and their achievements concerning the four language forms are well balanced, as all correctness ratios fall between 70-95 percent. Out of 55 students 54 used PRC with relatively high correctness (70%), EXPs were attempted by two thirds of pupils reaching even higher standards (77%), exceeding the overall mean by 24 percent. PLUs are in line with EXPs regarding accuracy (77%), while the negative construction seems to be fully acquired (95%) by the students who used it (Brown, 1973). The ratio of students using several language forms to express their ideas is 47 percent; it is the lowest in the case of NEG. Standard deviation is lower than that of the whole sample, especially regarding NEG (.188).

Table 21: Descriptive statistics of School 1: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 1 | N=55 | Mean | SD |
|---|---|---|---|
| prc | 54 | .7049 | .38990 |
| exp | 45 | .7744 | .39135 |
| plu | 51 | .7712 | .38436 |
| neg | 32 | .9505 | .18797 |
| Valid N (listwise) | 26 | | |

School 2 (see Table 22) also outperformed most of the other participating ones by 62 percent correct usage of PRC, which was endeavoured by all students of the school. EXPs were applied slightly more successfully (78%) than in School 1, while PLUs reached 71 percent

and NEG was also well accomplished by 89 percent. The ratio of pupils (80%) who applied all four constructions is noteworthy in this school. Standard deviation (.265-.401) is also a bit lower in their results than in the whole sample.

Table 22: Descriptive statistics of School 2: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 2 | N=44 | Mean | SD |
|---|---|---|---|
| prc | 44 | .6169 | .39690 |
| exp | 41 | .7772 | .38322 |
| plu | 43 | .7112 | .40104 |
| neg | 36 | .8935 | .26473 |
| Valid N (listwise) | 35 | | |

Students of School 3 (Table 23) displayed a very unbalanced performance regarding the four investigated language forms, their results fall between 11 and 67 percent. They achieved a very low correctness ratio in the usage of PRC (18%) and EXPs (11%), whereas PLUs were used with nearly 40 percent accuracy. NEG reached 67% correctness, though only four students applied it, and standard deviation is also high (.471). In the case of PRC none of the ten pupils achieved a 100 percent performance; 67 percent of correct usage being the best achievement. Altogether three students applied all four investigated constructions.

Table 23: Descriptive statistics of School 3: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 3 | N=10 | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| prc | 9 | .00 | .67 | .1764 | .22899 |
| exp | 9 | .00 | 1.00 | .1111 | .33333 |
| plu | 9 | .00 | 1.00 | .3889 | .42492 |
| neg | 4 | .00 | 1.00 | .6667 | .47140 |
| Valid N (listwise) | 3 | | | | |

School 4 (Table 24) belongs to the leading party as well with accuracy ratios distributed along a wide scale between 48%-90%. The results of PRC are ten percent higher than the overall results (58%); EXPs (75%) and NEG (90%) also show high correctness ratios compared to the means of the eleven schools. PLUs are less adequately performed (48%), which falls below the average, though the volunteering ratio is high, almost every student (24 out of 26) gave it a try. All scrutinised language forms are used by 54 percent of the pupils. Standard deviation (.28-.45) shows similarities with the overall average.

Table 24: Descriptive statistics of School 4: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 4 | N=26 | Mean | SD |
|---|---|---|---|
| prc | 26 | .5826 | .45396 |
| exp | 22 | .7485 | .42311 |
| plu | 24 | .4759 | .41385 |
| neg | 15 | .9000 | .28031 |
| Valid N (listwise) | 14 | | |

School 5 (Table 25) also joins the successful half of the participating institutions with accuracy ratios falling between 55-84 percent (Table 15). Their achievements display a similar trend in accuracy of the respective language forms to the means of the whole corpus. The only exception is NEG, which, compared to PLUs, is used slightly less correctly. PRC seems to be the least acquired form (55%), EXPs reach 68 percent correctness. NEG and PLUs are close in correctness with 81 and 84 percent, respectively.

Table 25: Descriptive statistics of School 5: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 5 | N=13 | Mean | SD |
|---|---|---|---|
| prc | 13 | .5463 | .42580 |
| exp | 10 | .6800 | .47329 |
| plu | 10 | .8350 | .32494 |
| neg | 9 | .8148 | .37680 |
| Valid N (listwise) | 8 | | |

Students in School 6 (Table 26) excelled in applying NEG; their 75 percent correctness is the highest out of the four language forms, though only four of the 18 students used it in their texts. PLUs reach 67 percent with two thirds of the pupils applying them. Their achievements in the usage of present continuous tense are lower than required (25%), and EXPs seem to be beyond the present capabilities of these students with a 16 percent correctness ratio. In this institution the accuracy ratios are also distributed along an extremely wide scale falling between 16-75 percent, representing a developing interlanguage, in which the respective language forms are at different stages of acquisition.

Table 26: Descriptive statistics of School 6: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 6 | N=18 | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| prc | 16 | .00 | 1.00 | .2492 | .31112 |
| exp | 14 | .00 | 1.00 | .1571 | .36101 |
| plu | 13 | .00 | 1.00 | .6679 | .29818 |
| neg | 4 | .50 | 1.00 | .7500 | .28868 |
| Valid N (listwise) | 4 | | | | |

In School 7 (Table 27) PLUs seem to be the most successfully acquired forms for 16 pupils out of 20 with a 54 percent accuracy rate. Other language forms are far from being acquired, as PRC achieves 33 percent correctness, whereas EXPs reach only 13 percent. The best achiever in this institution can use EXPs with 70 percent accuracy, which is lower than the average of Schools 1 and 2. While NEG is the best acquired language construction regarding the whole corpus, in this school it seems extremely underdeveloped with 22 percent of correct usage. Besides the low accuracy ratio SD is very high (.441) in this area. Interlanguage features may lie in the background or the amount of language input may differ from others.

Table 27: Descriptive statistics of School 7: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 7 | N=20 | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| prc | 13 | .00 | 1.00 | .3326 | .39463 |
| exp | 6 | .00 | .70 | .1282 | .31404 |
| plu | 16 | .00 | 1.00 | .5395 | .43080 |
| neg | 9 | .00 | 1.00 | .2222 | .44096 |
| Valid N (listwise) | 3 | | | | |

In School 8 (Table 28) the correctness ratios of the examined language forms fall near each other (11-45%), but, unfortunately, most of them are below the nationally required level of proficiency. The most successfully used construction is the PLU form of nouns, as nine of the eleven students applied it with 45 percent accuracy ratio in their scripts. This is still 20 percent lower than the average of the whole cohort. Negatives are also at the lower end of the IL continuum (33%) compared to the 79 percent accuracy ratio of the cohort. SD shows high values concerning both PLUs and NEG. The correctness ratios of present continuous (13%) and EXPs (11%) are far from reaching the minimum requirements of the national curriculum. Four out of eleven (36%) students applied all four language forms in their texts.

Table 28: Descriptive statistics of School 8: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 8 | N=11 | Mean | SD |
|---|---|---|---|
| prc | 10 | .1310 | .31211 |
| exp | 9 | .1111 | .33333 |
| plu | 9 | .4481 | .46161 |
| neg | 6 | .3333 | .51640 |
| Valid N (listwise) | 4 | | |

The mean for NEG (71%) nearly reaches the average of the cohort (79%) in School 9 (Table 29), while SD (.488) is high, representing great differences between students. EXPs are produced a bit more correctly than the average level of accuracy (57%) regarding all 231 pupils, whereas the correctness ratio of PLUs (.56) is 9 percent lower than that of the whole sample. The least successfully applied language form is present continuous (33%) among the students of this institution, and it is below the expected curricular standard. 46 percent of pupils volunteered to use all four language constructions.

Table 29: Descriptive statistics of School 9: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 9 | N=11 | Mean | SD |
|---|---|---|---|
| prc | 10 | .3333 | .38686 |
| exp | 7 | .5714 | .53452 |
| plu | 9 | .5556 | .44096 |
| neg | 7 | .7143 | .48795 |
| Valid N (listwise) | 5 | | |

The performance of students in School 10 (Table 30) shows an unexpected variety concerning the four language constructions, all of which was used by merely 28 percent of pupils. PLUs (73%) and NEG (71%) are acquired at similar, relatively high levels, whereas present continuous is used mostly inaccurately with a 6 percent correctness ratio. EXPs, while applied by 18 out of 20 students, reached zero correctness; there were no accurately used forms.

Table 30: Descriptive statistics of School 10: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 10 | N=20 | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| prc | 18 | .00 | 1.00 | .0556 | .23570 |
| exp | 18 | .00 | .00 | .0000 | .00000 |
| plu | 14 | .00 | 1.00 | .7143 | .37796 |
| neg | 11 | .00 | 1.00 | .7273 | .46710 |
| Valid N (listwise) | 5 | | | | |

The three students in School 11 (Table 31) displayed an extremely weak L2 performance with PRC being the only language form used above zero correctness (5%). As described under "Mean scores on task achievement" (4.8.1.3), in this school the highest score received for task achievement was two out of eight, which indicates that pupils did not write about the pictures, or wrote irrelevant things. These low achievements may be due to both poor language proficiency, and low motivation to write the test.

Table 31: Descriptive statistics of school 11: ratio of correct language forms; PRC, EXPs, PLU, NEG, N, mean, standard deviation

| School 11 | N=3 | Min | Max | Mean | SD |
|---|---|---|---|---|---|
| prc | 3 | .00 | .14 | .0476 | .08248 |
| exp | 3 | .00 | .00 | .0000 | .00000 |
| plu | 2 | .00 | .00 | .0000 | .00000 |
| neg | 1 | .00 | .00 | .0000 | .00000 |
| Valid N (listwise) | 1 | | | | |

In this section I have scrutinised how the students performed in the four language constructions in each school. Now, the results will be examined from the point of view of morpheme acquisition and also areas of strength will be shown in each institution.

The question concerning the accuracy level at which a particular morpheme can be regarded acquired was traditionally answered (Brown, 1973): a morpheme is considered 'acquired' if a learner achieves the accuracy score of 90 percent or higher. In other studies (McDaniel, McKee, & Cairns, 1998) the cutoff point is 75 percent correct use in obligatory contexts. I aim to explore which language forms can be considered acquired on a group level by the students of the eleven institutions, based on the more lenient viewpoint of McDaniel et al.

In School 1 the results of accuracy in NEG (95.05%) and the correct use of PLU nouns (77.12) surpasses the preset acquisition level, whereas the correctness ratios of both PRC

(70.49%) and EXPs (77.44%) are very close to this level. Pupils in School 2 excel in applying NEG (89.35%) and EXPs (77.72%), with PLUs (71.12%) being close to the pass-line. NEG is well acquired (90%) in School 4, whereas EXPs (74.85%) are at the preset point of acquisition. Students in School 5 acquired PLU forms (83.5%) best, while NEG (81.48) is also above the required level of accuracy. In School 6 merely one of the correctness ratios, that of NEG (75%) achieves the threshold. In Schools 9 and 10 the use of NEG (71.43%; 72.43%) is close to the required correctness ratio.

The above data suggest that NEG is the first acquired language construction in the investigated schools, whereas PLUs seem to come next, followed by EXPs. Present continuous tense appears to be acquired last of the four examined forms.

Expletives and PRC, however, show varying and somewhat contradictory correctness ratios in the examined institutions (Figure 14). In schools that outperformed their counterparts (Schools 1, 2, 4, 5, & 9) EXPs are produced more accurately than are present continuous forms. In schools underachieving in most language areas (Schools 3, 6, 7, 8, 10, & 11), however, PRC shows higher correctness ratios compared to EXPs. In Schools 10 and 11 no correct EXP occurred.

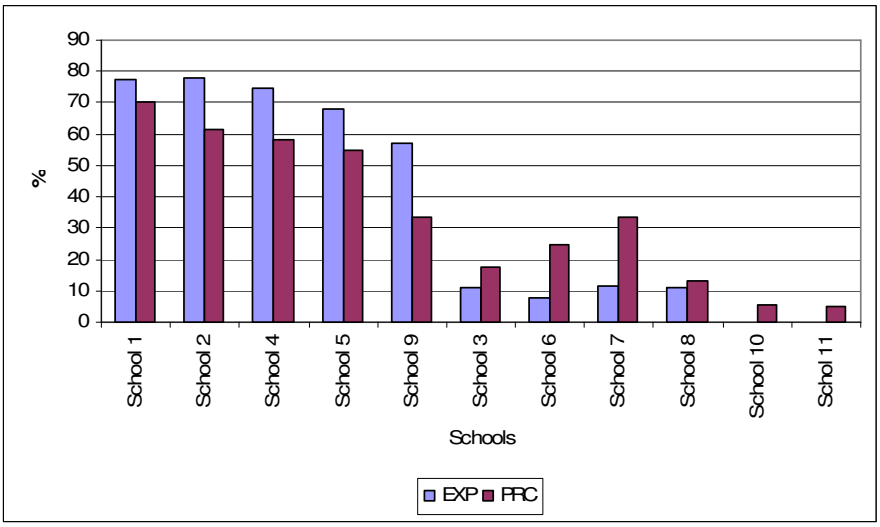



Figure 14: The correctness ratios of EXP and PRC in schools (%)

These data can be related to, and explained by, the transitional feature of interlanguage development. As new linguistic forms are constructed either externally from input or internally through cognitive processes, learner language construes an interlanguage continuum reflecting stages of development. Accordingly, the level of acquisition of a morpheme can be misjudged, as some morphemes go through a U-shaped development: first, they are produced correctly, then, due to overgeneralization, they are replaced by incorrect

forms. Finally, the correct form is acquired. Although it is hard to draw conclusions from a cross-sectional study, based on previous research (Bailey, Madden, & Krashen, 1974; Larsen-Freeman, 1976) I will attempt to do so.

The relatively correct usage of present continuous forms by low-achieving students may illustrate the first stage of development, where the construction is used as a memorised chunk of language. Later, at the next stage of its development, when a variety of verb aspects and tenses have already been 'introduced', and many verb forms are salient for students, they may mix the formal features of the continuous aspect with that of the simple aspect of the verb, as the language construction is not acquired yet. Considering the complex construction of PRC, at least three mistakes can occur while producing it: the omission of the auxiliary 'be', the omission of the morpheme '-ing' at the end of the verb, or both errors in parallel. Another explanation can be the later start, but faster rate of the development of EXPs. In a cross-sectional study we can only speculate as to the possible cause, longitudinal research could tell more about the phases of development.

Having observed the ratio of accurate usage of the four language forms by school, I aim to investigate the relationship between rater assessment and computer-based analysis.

### 4.8.3 Relationship between criterion- and corpus-based measurement: Accuracy scores versus ratios of correctness

In this section RQ 7: What is the relationship between criterion-based and corpus-based measurement of accuracy? will be answered.

First, the relationship between scores given for accuracy and the ratio of correct application of grammatical forms is explored for the whole corpus. To measure this, Pearson product-moment correlation coefficients were computed. As Table 32 illustrates, the number of correct forms have a strong (.797), significant ($p<0.01$), positive relationship with accuracy scores, whereas these scores have a modest, negative (-.318) relationship with the number of errors committed. This negative relationship is also significant ($p<0.01$). This result proves that the higher the number of correct language forms in a composition was, the higher accuracy score was awarded for the writing. This did not work the same way with the growing number of errors, as the negative relationship between errors and scores is a lot weaker. This moderate relationship may indicate that the assessor considered the number of repeatedly occurring errors, and also the influence of the mistake on the communicative function. So she did not deduct scores for each mistake. The number of correct and erroneous forms also shows a moderate, negative connection (-.331), and this relationship is significant at $p<.01$

level. The negative connection seems to be obvious, as it shows that students using a lot of correct language forms tend to make fewer mistakes. Moderate relationship indicates that there are students who apply an abundance of both correct and erroneous forms in their texts.

Table 32: Correlations between accuracy scores, errors and correctly used forms

| N=231 | | Accuracy score | Errors |
|---|---|---|---|
| Errors | Pearson Correlation | -.318(**) | 1 |
| Correct forms | | .797(**) | -.331(**) |

** Correlation is significant at the 0.01 level (2-tailed).

In an attempt to explain the effects of the two independent variables (correct forms and errors) on the accuracy score as a dependent variable, a multiple regression model was built. Both independent variables were entered into the model simultaneously. The main aim was to find the explanatory force of both correct and erroneous language forms. The result of the analysis indicates that the two variables contribute to the variation by .799, whereas the $R^2$ value shows the proportion of variation explained by these variables. According to this, 64 percent of the variance in accuracy scores is explained by the number of correct and erroneous language forms, whereas the other 36 percent is explained by other factors not included in this model.

Table 33 presents the regression model, which includes the independent variables (correct forms and errors) and shows how they contribute to the prediction of the dependent variable, that is, the accuracy score. The table displays the beta coefficients with their significance levels. Beta coefficients indicate the relative importance of a variable in predicting the accuracy score. According to this, errors have no significant relationship (p=.149) with accuracy scores, whereas correct forms show a significant relationship with grammar scores at p<.001. This fact shows that the assessor focused on communicative competence as opposed to the erroneous language forms, and rewarded the linguistic forms the students 'could do'.

Table 33: Results of the multiple regression analysis, dependent variable: accuracy score; predictors: correct forms, errors

| Model | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. |
|---|---|---|---|---|---|
| | B | Std. Error | Beta | | |
| Errors | -.015 | .010 | -.061 | -1.448 | .149 |
| Correct forms | .341 | .019 | .777 | 18.405 | .000 |

The same multiple regression model was built to explain the effects of the four investigated language forms (PRC, EXPs, PLU, & NEG) as independent variables on the accuracy score as a dependent variable. The $R^2$ value in Table 34 shows that the total ratio of variance in the dependent variable explained by the model is 75 percent; when treated separately, correctness values of the four language forms explain three quarters of accuracy scores.

Table 34: Results of the multiple regression analysis, dependent variable: accuracy score; predictors: the four investigated language forms together

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .866(a) | .750 | .741 | 1.205 |

As we have seen, 64 percent of the variance in accuracy scores is explained by the number of tagged correct and erroneous language forms, whereas 75 percent of the variance in the same scores is due to correctness ratios of the four scrutinised language constructions (PRC, EXP, PLU & NEG).

The difference between the effects of various independent variables on accuracy scores raises the question of the reasons of the discrepancy. A possible answer points to the unbalanced methodology of the tagging process, which I was not aware of at the time of the procedure.

The reasons may lie in this: on the one hand, each occurrence, either correct or erroneous, of the four language forms (PRC, EXP, PLU and NEG) was tagged, thus, the multiple regression model could rely on exact data of language usage regarding these areas.

However, as RQ 7 investigated the relationship between criterion-based assessment and linguistic analyses of texts, an extra coding category was developed for the investigation of correct language forms versus erroneous forms. This category seemed to be necessary in order to include all important errors other than the four investigated constructions in the count, as raters using the analytic scales considered all kinds of errors, not only the four selected language forms. Thus, word order errors, verb agreement errors, missing verbs, the use of '*on* the picture', missing prepositions '*with'*, and vocabulary misuse were all tagged. Nevertheless, the correct forms of the above constructions were not counted, though raters also assess these language forms. Therefore, the calculation of correct forms involved merely the accurately performed forms of PRC, EXP, PLU and NEG, whereas erroneous forms included both the errors in these four constructions and *other* mistakes. Thus, the 64 percent variance in accuracy scores explained by correct and incorrect language use is not surprising.

Looking back, it seems to have been unnecessary to tag all OTHER mistakes and involve them in the calculations.

Table 35 displays the ratio of contributions of correctly used language forms to the accuracy score. It demonstrates that the number of correct PLU forms do not significantly contribute to the accuracy score (p=.755). Although correctly used negatives have a significant effect (p=.017) on accuracy results, this effect is inferior to that of the other two forms: PRC and EXPs (p<.001). PRC and EXPs significantly contribute (p<.001) to the judgment regarding accuracy. Their partial contribution to the correlation is 0.533 and 0.629 respectively; thus EXPs are the best predictors of accuracy scores. Beta coefficient multiplied by the value of the partial correlation shows how much of the variance is explained by the correct usage of the respective language form. Thus, PRC explains 20 percent of the variance, whereas EXPs are responsible for 32 percent. The correctness level of NEG predicts merely 3 percent of the accuracy measurement value.

This result does not indicate, however, that the task triggered the usage of a significantly higher number of EXPs than present continuous forms, NEG or PLU of nouns. Figure 11 (under 4.8.2) illustrates the total numbers of the investigated constructions in the whole corpus. It shows that PRC forms provide 52 percent, while EXP adds another 24 percent, whereas PLUs appear in 16 percent of all four constructions, and NEG comprises almost 8 percent of the analyzed language forms.

Table 35: Results of the multiple regression analysis, dependent variable: accuracy score; predictors: PRC, EXP, PLU and NEG

| | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Correlations | |
|---|---|---|---|---|---|---|---|
| | B | Std. Error | Beta | | | Zero-order | Partial |
| prc | 2.127 | .333 | .384 | 6.392 | .000 | .717 | .533 |
| exp | 2.683 | .327 | .510 | 8.203 | .000 | .784 | .629 |
| plu | .093 | .298 | .016 | .313 | .755 | .251 | .031 |
| neg | .981 | .404 | .137 | 2.430 | .017 | .519 | .233 |

These results suggest that the rater relying on the analytic scales found the correctness of EXPs the most important constituent of reaching communicative purposes. Errors in using EXP may cause misunderstanding of the message more easily than errors in PRC. The accuracy features of PRC are still important for communication, but the incorrect forms might not hinder communication to an extent similar to erroneous EXPs. The mistakes committed in applying PLU forms might not be considered crucial, as the numbers before the noun already

indicate the concept of PLU. As for NEG, the presence of 'no' or 'not' do suggest the concept of NEG, this might have caused the low percentage of variance explained by the correctness ratio of this construction.

In this section I have examined the interrelations of criterion-based and corpus-based measurement of students' language accuracy, evaluated by Rater 5. The next section will visit assessment issues again, this time the relationship among five raters' judgments will be explored, so as to answer RQs 8, 9, and 10.

### 4.8.4 Relationship among raters' opinions

#### 4.8.4.1 The aims of reassessment

In this section, I will answer the following research questions:

RQ 8: How are the raters' opinions of sample tests related to each other?

RQ 9: How does criterion-based assessment compare with holistic assessment?

RQ 10: How do raters' interpretations of *CEFR* scales compare?

The expressions 'assessor', and 'rater' will be used interchangeably, both of them denoting the five persons who undertook the job of reassessing a sample of students' compositions.

The aim of multiple assessments is generally the ensuring or control of the reliability of measurement. In this study, the aims are twofold: on the one hand, the two main types of assessment, holistic and analytic methods will be compared; on the other hand, the comprehensiveness, transparency and coherence of the *CEFR* scales (Council of Europe, 2001) will be investigated.

#### 4.8.4.2 Raters' tasks

First, I summarize what the raters' tasks involved. As introduced under 4.6.4.2, they were asked to rank order a sample of students' writings according to their level of English language proficiency. Thus, they were to apply holistic assessment, or rather, general impression marking. This means that they were required to express their overall opinion of the students' performances by ranking scripts, without receiving any criteria for assessment. First, they were supposed to rank order the fifteen selected scripts proceeding from the best work to the poorest writing (Appendix B).

The second task involved criterion-based assessment in which raters needed to allocate students' compositions in a grid with randomly placed descriptors representing levels A1-B1 of *CEFR* (Appendix C). Placement was to be arranged along four different criteria of writing described in the European document (Appendix D):

(a) General linguistic range (GLR),
(b) Overall written production (OWP),
(c) Vocabulary range (VOC),
(d) Grammatical accuracy (GRA).

Thus, raters first needed to interpret descriptors and rank them according to the level of represented language proficiency. The aim of this task was the examination of raters' individual interpretation of *CEFR* descriptors without standardisation of criteria. Table 36 provides an example by showing part of the grid related to OWP. This criterion involved three proficiency levels: A1, A2 and B1, the random order of which in the grid is A2, B1, and A1, starting from the top. Identifying the levels was not very difficult in this case, as there were merely three of them, and the clues regarding lexical complexity (e.g., isolated phrases; series of simple phrases; connected texts) were easy to identify.

Table 36: Part of raters' grid for assessing overall written production of 15 texts

| Overall written production descriptors | Students' codes | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | E | F | G | H | K | L | M | N | O | P | R | S | T | V | Z |
| Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but', and 'because'. | | | | | | | | | | | | | | | |
| Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. | | | | | | | | | | | | | | | |
| Can write simple isolated phrases and sentences. | | | | | | | | | | | | | | | |

The identification of levels was not simple in all knowledge areas, though, as the number of defined language levels in *CEFR* differed from criterion to criterion: main levels (A1, A2 and B1) were further divided into two sub-levels (e.g., A2.1 and A2.2). Table 37 shows the randomly placed descriptors concerning the other three areas, GLR, VOC, and GRA.

Table 37: Descriptors in raters' grid for assessing general linguistic range, vocabulary range, and grammatical accuracy of 15 texts

| **General linguistic range** |
|---|
| Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words. |
| Has a very basic range of simple expressions about personal details and needs of a concrete type. |
| Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interest, work, travel, and current events, but lexical limitations cause repetitions and even problems with formulation at times. |
| Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information.<br>Can use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulae about themselves and other people, what they do, possessions etc.<br>Has a limited repertoire of short memorised phrases covering predictable survival situations. |
| Has sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films. |
| **Vocabulary range** |
| Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations. |
| Has a sufficient vocabulary for<br>a) The expression of basic communicative needs.<br>b) Coping with simple survival needs. |
| Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events. |
| Has a sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics. |
| **Grammatical accuracy** |
| Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is clear what he/she is trying to express. |
| Uses simple structures correctly, but still systematically makes basic mistakes – for example tends to mix up tenses and forget to mark agreement; nevertheless, it is usually clear what he/she is trying to say. |
| Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire. |
| Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations. |

Within the four criteria for the assessment of scripts OWP was described in three bands (A1, A2, B1), VOC (A1, A2.1, A2.2, B1) and GRA (A1, A2, B1.1, B1.2) had four bands of descriptors each, whereas GLR contained five distinct levels (A1, A2.1, A2.2, B1.1, B1.2). It proves further inconsistency of the Framework that for the characterization of GRA it provides two separate bands within level B1, whereas in relation to VOC it offers two distinct bands within another language proficiency level, A2. With this variety of the numbers of

bands for separate criteria the comprehensiveness and transparency (2.2.1) of the reference scales seem to be weakened. Let us examine how raters coped with the complex task of the interpretation of descriptors.

#### 4.8.4.3 Reliability measures concerning holistic and criterion-based assessments

First, I will examine the reliability of raters' opinions applying Cronbach's alpha reliability measure, then, by using Krippendorff's alpha reliability statistics (Table 38).

Cronbach's alpha demonstrates strong relationships between raters' decisions concerning all criteria, as it is over 0.83 in each area of assessment. Agreement seems to be the highest in overall ranks (Cronbach's alpha=.969) and vocabulary measures (Cronbach's alpha=.913). Interrater reliability for GLR reached 0.867, whereas for OWP it was 0.861. GRA features distributed raters' opinions the most, but still produced a very high 0.835 reliability value.

According to Hayes and Krippendorff (2007), however, Cronbach's alpha does not measure concrete agreement, it merely quantifies the consistency by which assessors judge units on an interval scale without being sensitive to how much the observers actually agree in their judgments. In our case, for example, if one of the raters placed students' scripts consistently in the lowest and second lowest bands, whereas the other assessor allocated the same writings in the second and third bands with the same consistency, Cronbach's alpha would identify strong relationship between the two assessors. Krippendorff's alpha, on the other hand, measures actual agreement between assessors, so in the above example the relationship would not be so strong. Let us examine the divergence in reliability measures produced by these two different statistical methods.

Table 38 also illustrates reliability coefficients expressed by Krippendorff's alpha. Relationships are much weaker according to this measure; only overall ranking shows strong (.8665) interrater reliability. The assessment of other knowledge areas is moderately interrelated; reliability coefficients range between 0.5175 and 0.5906, VOC being the second strongest area of interrater reliability (Krippendorff's alpha=0.5906). The reliability measure of other three areas is around 0.52, which is not surprising, as Krippendorff's alpha represents concrete agreements of data.

Table 38: Reliability statistics of raters' holistic and criterion-based assessment

| Knowledge area / rank | Cronbach's Alpha | Krippendorff's Alpha | Number of Items |
|---|---|---|---|
| Grammatical accuracy | .835 | .52 | 5 |
| General linguistic range | .867 | .5211 | 5 |
| Overall rank | .969 | .8665 | 5 |
| Vocabulary range | .913 | .5906 | 5 |
| Overall written production | .861 | .5175 | 5 |

Summing up the results from the two different reliability measures, both of them showed that the most interrelated areas of raters' viewpoints were overall ranking and VOC, although great discrepancy was presented between the levels of agreement in the two separate measures.

So as to investigate the details of inter-relations in raters' opinions, Spearman rank-order correlation coefficients were computed. Tables 39-43 illustrate the results in each observed knowledge area.

#### 4.8.4.4 Correlation between holistic judgments

Task one required raters to rank order samples of students' scripts. Table 39 illustrates Spearman rank-order correlation coefficients between raters concerning the rank order of students' written performance. All correlations are significant at the $p<.01$ level, and correlation coefficients range between 0.803 and 0.928. Rater 2 (R2) has the highest correlation values with other raters, whereas R4 shows the lowest interrelation measures with his colleagues, which range between 0.807 and 0.839, but are still high correlations.

Table 39: Correlations between raters on the rank order of writings

| Criterion | N=15 | | Rater 1 | Rater 2 | Rater 3 | Rater 4 |
|---|---|---|---|---|---|---|
| Overall rank | Spearman's rho correlation coefficient | Rater 2 | .928(**) | | | |
| | | Rater 3 | .803(**) | .918(**) | | |
| | | Rater 4 | .813(**) | .850(**) | .807(**) | |
| | | Rater 5 | .874(**) | .914(**) | .904(**) | .839(**) |

** Correlation is significant at the 0.01 level (2-tailed).

#### 4.8.4.5 Correlation between criterion-based judgments on students' language levels using *CEFR* scales

In this section the relationship between assessors' ratings of scripts will be described along four different criteria of writing (GLR, OWP, GRA, & VOC).

Table 40 illustrates the relationship between raters' opinions on students' GLR. Five out of ten correlation coefficients are significant at the p<.01 level; one association is significant at the p<.05 level, whereas four connections do not reach the 95 percent confidence level; these relationships may be due to chance. These discrepant correlation data belong to R4. The possible reasons of disagreements will be examined in the section discussing assessors' feedback on their task.

Scrutinizing relationships between raters' opinions on GLR one by one, we find strong, significant correlations between Raters 1 and 2 (Spearman's rho=.965), Raters 1 and 3 (0.831), and also between Raters 1 and 5 (Spearman's rho=.693). R2 agreed basically on ranks of scripts not only with R1, but also with R3 (0.838), and 5 (0.731). Decisions of R3, besides having strong significant relationship with Raters 1 and 2, are less strongly interrelated with R5 (0.624) at p<.05 level of significance. None of R4's data show significant correlations with other assessors.

Table 40: Correlations between raters on the linguistic range of writings

| Knowledge area | N=15 | | Rater 1 | Rater 2 | Rater 3 | Rater 4 |
|---|---|---|---|---|---|---|
| General linguistic range | Spearman's rho correlation coefficient | Rater 2 | .965(**)<br>.000 | | | |
| | | Rater 3 | .831(**)<br>.000 | .838(**)<br>.000 | | |
| | | Rater 4 | .125<br>.657 | .218<br>.435 | .251<br>.367 | |
| | | Rater 5 | .693(**)<br>.004 | .731(**)<br>.002 | .624(*)<br>.013 | .268<br>.334 |

** Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

Although OWP consisted of merely three separate linguistic levels (Table 41), it notably diversified opinions. The table illustrates that four of the ten correlation coefficients are significant at p<.01 level; one relationship is significant at p<.05 level, whereas half of the associations between raters' decisions do not reach the 95 percent confidence level. This is mainly caused by R2's assessment of students' overall performance differently from the rest of the assessors. She had no significant correlations with any of her peers. This discrepancy may be due to her misapprehension of the descriptors in the scales, and the interchange of two

linguistic levels. By contrast, in the rank ordering task this rater had the highest correlation values with other raters, which shows that her holistic assessment concerning students' scripts is in accordance with her peers.

In this knowledge area R5's ratings have the strongest relationships with those of her colleagues; strong, significant correlations occurred between her and R1 (Spearman's rho=.823), R4 (0.732), and slightly weaker, but significant associations with R3 (0.708). Raters 1 and 3 also agreed (0.836) on students' overall written performance to a great extent, whereas Raters 3 and 4 have no significant relationships regarding this area.

Table 41: Correlations between raters on the overall written production of writings

| Knowledge area | N=15 | | Rater 1 | Rater 2 | Rater 3 | Rater 4 |
|---|---|---|---|---|---|---|
| Overall written production | Spearman's rho correlation coefficient | Rater 2 | .441<br>.100 | | | |
| | | Rater 3 | .836(**)<br>.000 | .298.<br>.281 | | |
| | | Rater 4 | .579(*)<br>.024 | .206<br>.460 | .457<br>.087 | |
| | | Rater 5 | .823(**)<br>.000 | .170<br>.544 | .708(**)<br>.003 | .732(**)<br>.002 |

** Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

Spearman' rho (see Table 42) reveals the second best level of associations between raters' rankings regarding GRA compared to the other investigated knowledge areas. Two out of ten correlation coefficients proved to be significant at p<.01 level, five at p<.05 level, and three interrelations are not significant. The latter correlation data belong to R4, and these discrepancies will be examined when discussing assessors' feedback on their task.

Investigating correlations regarding GRA between assessors one by one, we find strong, significant (p<.01) correlations between Raters 1 and 2 (Spearman's rho=.807), and also between Raters 1 and 5 (Spearman's rho=.756). Correlation coefficients are lower, but still significant (p<.05) between Raters 2 and 3 (0.612); similar, moderate relationships characterise Raters 2 and 5 (0.588). R3's decisions on GRA are less strongly interrelated with all other raters, although, they are significant at p<0.05 level. R4 has one significant correlation coefficient (0.588) showing moderate connection with R3.

Table 42: Correlation between raters on the grammatical accuracy of writings

| Knowledge area | N=15 | | Rater 1 | Rater 2 | Rater 3 | Rater 4 |
|---|---|---|---|---|---|---|
| Grammatical accuracy | Spearman's rho correlation coefficient | Rater 2 | .807(**)<br>.000 | | | |
| | | Rater 3 | .532(*)<br>.041 | .612(*)<br>.015 | | |
| | | Rater 4 | .353<br>.197 | .335<br>.222 | .588(*)<br>.021 | |
| | | Rater 5 | .756(**)<br>.001 | .588(*)<br>.021 | .633(*)<br>.011 | .428<br>.112 |

** Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

As shown in Table 43, assessors agreed on pupils' VOC, this is the second best correlating knowledge area by their opinions after the rank orders of students' scripts. Nine correlation coefficients are significant at p<.01 level, whereas one is significant at p<.05 level. Raters 5 and 2 expressed the closest opinions reaching 0.914 agreement. A similarly high correlation coefficient relates Raters 1 and 2 (0.778), and Raters 2 and 3 (0.771). In this area R2's agreement is the strongest with her colleagues.

Table 43: Correlation between raters on the vocabulary range of writings

| Knowledge area | N=15 | | Rater 1 | Rater 2 | Rater 3 | Rater 4 |
|---|---|---|---|---|---|---|
| Vocabulary range | Spearman's rho correlation coefficient | Rater 2 | .778(**)<br>.001 | | | |
| | | Rater 3 | .707(**)<br>.003 | .771(**)<br>.001 | | |
| | | Rater 4 | .646(**)<br>.009 | .722(**)<br>.002 | .725(**)<br>.002 | |
| | | Rater 5 | .672(**)<br>.006 | .914(**)<br>.000 | .571(*)<br>.026 | .737(**)<br>.002 |

** Correlation is significant at the 0.01 level (2-tailed).
*Correlation is significant at the 0.05 level (2-tailed).

Summarising findings in answer to RQ 8 regarding the correlation of the five raters' opinions, according to Cronbach's alpha, interrater reliability was high in all assessed areas, whereas Krippendorff's alpha measured high agreement in raters' judgments merely in the overall ranking of scripts. Both statistics produced the highest values of agreement in overall ranking of writings and, secondly, in relation to pupils' VOC. Thus, holistic judgments proved more reliable than criterion-based ones; at least this is what the results suggest at first sight.

As for OWP, R2 disagreed the most with the rest of the assessors, while concerning GRA and GLR, R4 shared common opinions the least. The reasons for these disagreements will be examined in the description of the rating process.

Figure 15 illustrates all five raters' criterion-based rankings regarding the four linguistic areas. The most outlying high values are R5's ratio of A2 level regarding GRA, R4's ratio of A2.1 level and R5's ratio of A1 in GLR, and R1's ratio of B1 level concerning VOC. R2's A2 bar in OWP also exceeds the rest of ratings in this category, which causes the absence of significant relationships with other raters' ranks. There are also outstanding low ratios compared to other raters' decisions, for example, in R1's B1.1 ranking concerning GRA, and also in her A2.2 ratio related to GLR. These data confirm the results computed with the help of Krippendorff's alpha, which measured the actual agreement between assessors and showed high correlation only in overall ranking of students' scripts.

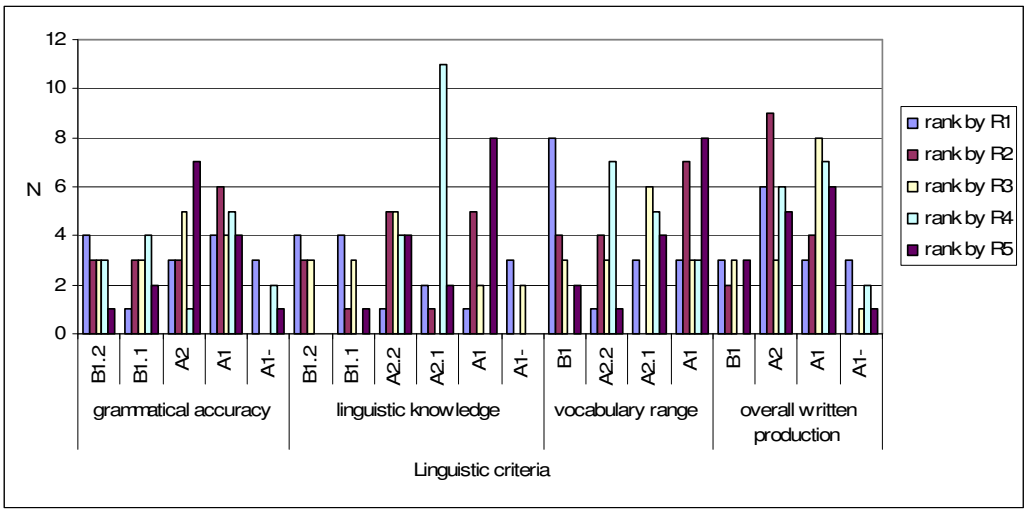



Figure 15: Placement of students' scripts in different *CEFR* levels along four criteria by five raters

The fact that all raters agreed on overall ranking of scripts during holistic measurement reveals that they relied on common underlying criteria while qualifying scripts. The discrepancies between their judgments in certain knowledge areas, on the other hand, indicate different interpretations of some of the descriptors of *CEFR* (Council of Europe, 2001). This supports Alderson et al.'s (2004) argument that the limitations of *CEFR* may cause uncertainty regarding ranking learners' proficiency levels.

#### 4.8.4.6 Intra-rater correlation coefficients: The relationship between criterion-based and holistic assessment

Besides inter-rater correlation coefficients it is also interesting to observe intra-rater correlation measures, that is, the extent to which raters' opinions regarding students' proficiency in certain knowledge areas coincide with their judgments on pupils' performances

in other knowledge areas. Being aware that language learners do not develop at the same rate in all linguistic areas, which places them in different bands along the vertical dimension of the scale, I did not expect total agreement in the following investigation. Nevertheless, the rank order of scripts and the rankings in *CEFR* levels along different criteria were expected to correlate with each other. The low correlation coefficients between raters in some of the knowledge areas, which suggested misinterpretations of certain *CEFR* descriptors by three of the raters, though, foretells the possible reoccurrence of low correlations regarding intra-rater coefficients. In order to obtain these statistics, Spearman rank-order correlation coefficients were computed.

R1 made consistent decisions concerning the five areas (Table 44), the relationships between all of which are significant ($p<.01$). Most of the correlation coefficients are above .906; they range between .887 and .959. The strongest relationship ties the GLR of writings and the overall ranks (.959). The weakest connection is computed between her assessments regarding GRA and OWP (.887).

Table 44: Intra-rater correlation coefficients of Rater 1

| Rater 1 | N=15 | Grammatical accuracy | General linguistic range | Overall rank | Vocabulary range |
|---|---|---|---|---|---|
| Spearman's rho Correlation Coefficient Sig. (2-tailed) | General linguistic range | .948(**) | | | |
| | Overall rank | .942(**) | .959(**) | | |
| | Vocabulary range | .926(**) | .935(**) | .906(**) | |
| | Overall written production | .887(**) | .953(**) | .958(**) | .909(**) |

** Correlation is significant at the 0.01 level (2-tailed).

Figure 16 demonstrates that R1 had low opinion of nearly half of the pupils' GRA presented in their scripts (N=7; 46.7%), and also of their OWP (N=6; 40%), as she assigned these scripts in levels A1- and A1. The most outstanding bar illustrates her appreciation of the VOC of eight scripts (level B1). GLR was also high quality in her opinion, as eight scripts were placed in B1 level (B1.1 & B1.2).

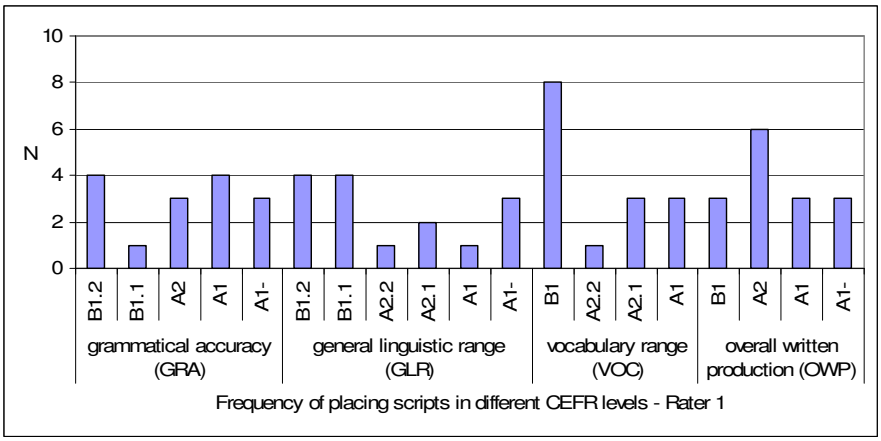



Figure 16: Placement of students' scripts in different *CEFR* levels along four criteria - Rater 1

R2 achieved lower intra-rater correlation coefficients (Table 45). The strongest associations characterise her rating of the GLR of writings and the overall ranks (.96), and, also her opinions on VOC and the overall ranks (.93). What reduces the strength of her correlation data is that none of her decisions on OWP are in significant relationship with her assessments on other linguistic areas.

Table 45: Intra-rater correlation coefficients of Rater 2

| Rater 2 | N=15 | Grammatical accuracy | General linguistic range | Overall rank | Vocabulary range |
|---|---|---|---|---|---|
| Spearman's rho Correlation Coefficient Sig. (2-tailed) | General linguistic range | ,880(**) | | | |
| | Overall rank | .834(**) | .960(**) | | |
| | Vocabulary range | .819(**) | .924(**) | .930(**) | |
| | Overall written production | .288<br>.298 | .389<br>.152 | .471<br>.076 | .263<br>.343 |

** Correlation is significant at the 0.01 level (2-tailed).

Concerning OWP (Figure 17), R2 ranked nine scripts in A2 level, whereas in other areas she distributed performances along the scale in a more balanced way. This data also supports the assumption that she misinterpreted the descriptors of two levels regarding this criterion. She placed two to six writings on level B along different criteria; the lowest ratio belongs to OWP.

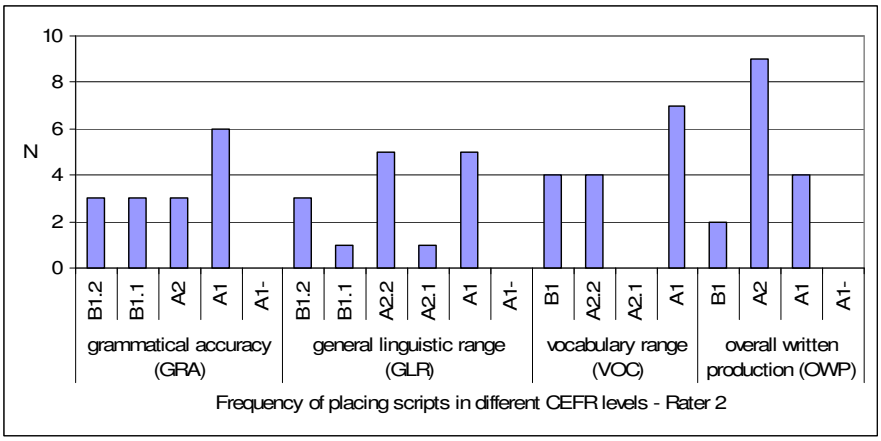



Figure 17: Placement of students' scripts in different *CEFR* levels along four criteria - Rater 2

All correlations are significant at $p<0.01$ level for R3 (Table 46), most of Spearman's coefficients are high, ranging between 0.81 and 0.973. The strongest relationship occurred between opinions with regard to the overall rank and GLR (.973), and also, between GRA and VOC (.965) of scripts. OWP and GRA showed the weakest association (.81) with each other, similarly to R1.

Table 46: Intra-rater correlation coefficients of Rater 3

| Rater 3 | N=15 | Grammatical accuracy | General linguistic range | Overall rank | Vocabulary range |
|---|---|---|---|---|---|
| Spearman's rho Correlation Coefficient Sig. (2-tailed) | General linguistic range | .890(**) | | | |
| | Overall rank | .866(**) | .973(**) | | |
| | Vocabulary range | .965(**) | .872(**) | .856(**) | |
| | Overall written production | .810(**) | .895(**) | .879(**) | .827(**) |

** Correlation is significant at the 0.01 level (2-tailed).

As shown in Figure 18, she ranked 11 texts above A1 level (6 scripts B level among them) regarding GRA, whereas only six of the scripts concerning OWP exceeded A1 according to her, three of them was ranked A2, three B1 level. R3 seems to have the highest opinion of students' GRA and GLR, as she placed three texts in both B1.1 and B.1.2 levels.

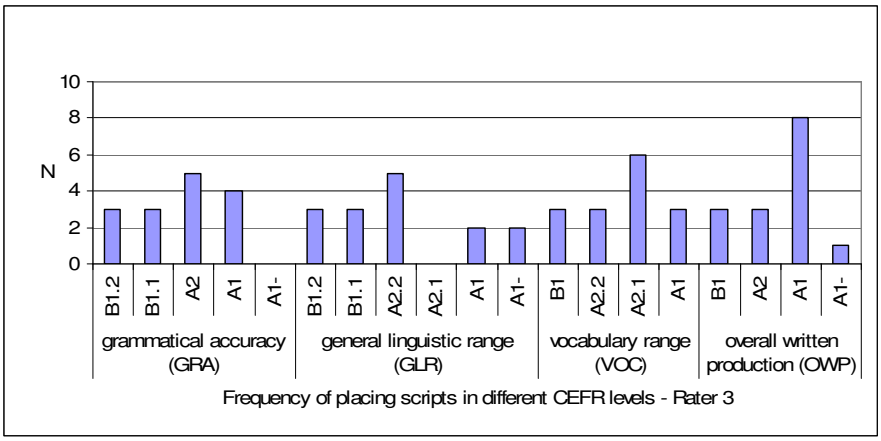



Figure 18: Placement of students' scripts in different *CEFR* levels along four criteria - Rater 3

R4's assessments fall into three categories regarding significance levels: two of his decisions on VOC and two concerning the overall ranks are significant at p<.01 level, three of his opinions reach p<0.05 level, and three rankings do not reach the 95% confidence level (Table 47). The most strongly interrelated areas of his measurement are overall ranks and VOC (Spearman's rho=.787) of writings, which is lower than R1's (.887), R2's (.819) and R3's (.81) lowest correlation coefficients.

Table 47: Intra-rater correlation coefficients of Rater 4

| Rater 4 | N=15 | Grammatical accuracy | General linguistic range | Overall rank | Vocabulary range |
|---|---|---|---|---|---|
| Spearman's rho Correlation Coefficient Sig. (2-tailed) | General linguistic range | .613(*)<br>.015 | | | |
| | Overall rank | .522(*)<br>.046 | .209<br>.454 | | |
| | Vocabulary range | .709(**)<br>.003 | .603(*)<br>.017 | .787(**)<br>.000 | |
| | Overall written production | .369<br>.176 | .439<br>.102 | .676(**)<br>.006 | .722(**)<br>.002 |

* Correlation is significant at the 0.05 level (2-tailed).
** Correlation is significant at the 0.01 level (2-tailed).

As seen in Figure 19, most writings (N=11=73.3%) were ranked as A2.1 level regarding their GLR, whereas none of them is considered to belong to a lower level, like A1 or A1-, which reveals the misinterpretation of the *CEFR* scale related to this knowledge area considering that in other areas three to nine of the scripts were classified in lower bands. In general, R4 had higher expectation towards scripts than Raters 1, 2, and 3, as he placed no scripts in level B concerning three criteria, although he judged the GRA of seven writings to belong to B1 level. This may be due to the lack of his teaching experience, during which the originally high

expectations concerning students' linguistic proficiency are gradually altered, and the variety of proficiency levels can be studied and developed.

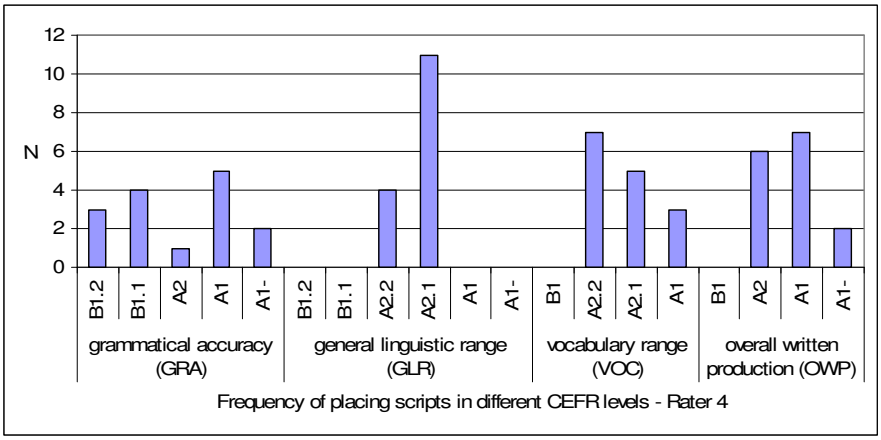



Figure 19: Placement of students' scripts in different *CEFR* levels along four criteria - Rater 4

Nine out of ten correlations are significant at p<.01 level in R5's assessments, whereas the relationship concerning GRA and GLR is not significant (Table 48). The other interrelations range between 0.68 and 0.907, two of them are moderate (.68 and .74), the rest are strong. Connections between VOC and OWP (.907) proved to be the strongest, and VOC is also strongly related to the overall rank (.905).

Table 48: Intra-rater correlation coefficients of Rater 5

| Rater 5 | N=15 | Grammatical accuracy | General linguistic range | Overall rank | Vocabulary range |
|---|---|---|---|---|---|
| Spearman's rho Correlation Coefficient Sig. (2-tailed) | General linguistic range | .487<br>.066 | | | |
| | Overall rank | .782(**)<br>.001 | .774(**)<br>.001 | | |
| | Vocabulary range | .680(**)<br>.005 | .885(**)<br>.000 | .905(**)<br>.000 | |
| | Overall written production | .740(**)<br>.002 | .775(**)<br>001 | .884(**)<br>.000 | .907(**)<br>.000 |

** Correlation is significant at the 0.01 level (2-tailed).

R5 seems to be stricter in her decisions concerning linguistic levels, as she assigned fewer scripts to level B (N=1-3) than her peers (Figure 20). The extremely high percentage (N=8=53.3%) of scripts classified as A1 regarding GLR and VOC may point to interpretation problems of the scale. I will return to this issue in the qualitative analysis of five scripts.

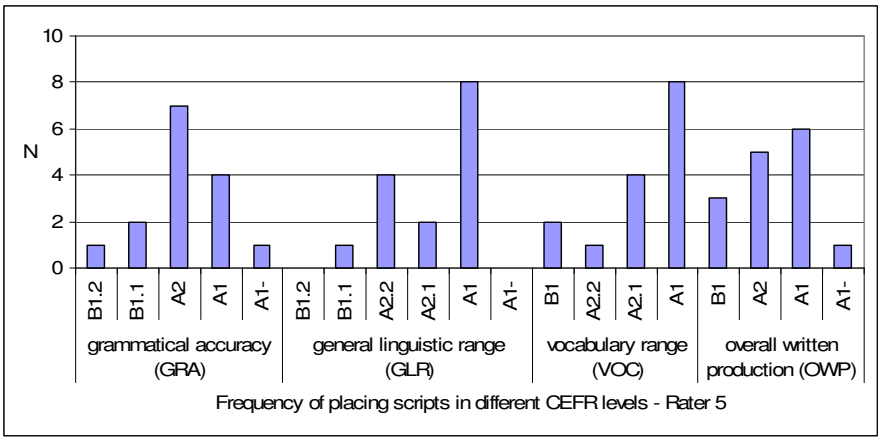



Figure 20: Placement of students' scripts in different CEFR levels along four criteria -Rater 5

To sum up the findings with regard to intra-rater correlation coefficients, raters demonstrated various combinations of consistency. Raters 1 and 3 made the most consistent decisions concerning students' linguistic performance, which was manifested in the high correlation coefficients between overall ranking and the four linguistic criteria of *CEFR* (Council of Europe, 2001). R4's opinions, on the other hand, revealed contradicting perceptions of students' linguistic levels in two areas.

The strongest relationships occurred between overall ranks and GLR in three of the assessors' judgments (Rs 1, 2, & 3), whereas R4's assessments correlated most strongly in relation to overall ranks and VOC. Rater 5's decisions were similar to R4, showing the strongest relationships between OWP and VOC, and also, between overall rank and VOC. Assessor 2 presumably had interpretation problems with OWP, whereas R4 misinterpreted ranks in GLR. R5's ratings produced no significant relationship between grammatical accuracy and linguistic range, which suggests the misunderstanding of the descriptors in at least one of these criteria.

RQ 9 inquiring into the relationship of criterion-based and holistic assessment can be answered partially after the comparison of raters' judgments in both types of measurement. According to statistical analyses, holistic assessment shows the strongest relationship with the measures of at least one of the linguistic areas, mostly with GLR and VOC. The areas of weak relationship, on the other hand, indicate misinterpretations of certain descriptors. I did not exclude any of the suspicious decisions first, as one of my aims was to investigate raters' interpretation of the *CEFR* scale. In a second exploration, however, outlying opinions were excluded from data to reach more reliable conclusions.

Thus, let us see statistics illustrating Cronbach's alpha if the item differing from the others most is deleted (Table 49). Relationships between raters' decisions became stronger concerning all criteria, except for VOC. Agreement is still the highest in overall ranks (Cronbach's alpha=.97), but the second place of VOC is handed over to GLR (.915). The reliability coefficient of GRA is still the lowest of the five values (0.855); nevertheless, it shows stronger associations of opinions than before. Concerning VOC, the exclusion of R5 was not necessary, as it reduced the reliability value.

Table 49: Reliability statistics of raters' assessment after the deletion of the most outlying opinion in each area

| Knowledge area (N=5) | Cronbach's α | Deleted item | Cronbach's α if item deleted |
|---|---|---|---|
| Grammatical accuracy | .835 | Rater 4 | .855 |
| General linguistic range | .867 | Rater 4 | .915 |
| Overall rank | .969 | Rater 4 | .970 |
| Vocabulary range | .913 | Rater 5 | .910 |
| Overall written production | .861 | Rater 2 | .902 |

#### 4.8.4.7 Raters' comments on the process of assessment

The last task on the raters' task sheet was to write how well they knew the *CEFR* and comment on the process of assessment. In this section I will offer an answer to RQ 10: How do raters' interpretations of *CEFR* scales compare?

Regarding their familiarity with the document, all raters selected the answer "partly", rather than "very well" or "not at all". R4 commented that he was familiar with the contents of the framework, but he did not know all the descriptors of the scales by heart, which was not at all expected.

In regard to the holistic task, R1 considered it difficult to rank certain scripts, so she decided to rank two pairs of texts on the same place (ranks 5-6; 8-9). R2 also mentioned that she usually assesses compositions relying on analytical scales, thus holistic measurement "was not easy" for her. R3 also thought that the fifteen scripts were too many to rank, and she preferred placing groups of scripts along the scale, as proved to be easier. These feedback reveal that raters were unable to identify fifteen separate language levels within the range of A1-B1; examples of the process will be given in relation to R4 and R5.

The most detailed description of the process of ranking was provided by R4. After reading the first text, he used it as a point of reference, and ranked the next script before or after this one, depending on its quality. He continued this way, always trying to place the latest script before a poorer one. He considered rank ordering to be a 'subjective' task, as he felt that it

was impossible to perceive differences between certain scripts. Instead of rank ordering he felt that placing writings in small groups, and then ranking them would be more reliable. During holistic assessment he tried not to apply special criteria for assessment known from analytical scales, but he still considered 'certain' linguistic features, which he did not name.

R5 was the only assessor who measured the scripts in three rounds: first by using criteria, later by ranking them holistically, and then by considering new (*CEFR*) criteria. This provides us with the opportunity to compare the three ratings and check intra-rater reliability.

As for rank orders of scripts, they slightly changed in the second, holistic assessment (Table 50). Some writings kept their positions (scripts 1, 2, 3, & 14) a few changed neighbouring places (scripts 9-10, & 11-12), others were placed two ranks away from their first positions (scripts 13 & 15). The biggest alterations occurred among rankings 4-8; in this group the rise of Script H from place 8 to 5 is the most outstanding change. Scrutinizing the table, we can observe that there are five groups of different language proficiency formed after the two rankings, as perceived by R5. In the following introduction of these groups, the range in total points awarded to each script in the criterion-based rating will be shown in brackets.

Table 50: Rank orders of scripts in criterion-based and holistic rating - Rater 5

| Rank \ script | **F** | **V** | **S** | **G** | **R** | **M** | **P** | **H** | **K** | **E** | **L** | **O** | **N** | **T** | **Z** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Criterion-based rating | 1 | 2-3 | 2-3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Holistic rating | 1 | 3 | 2 | 6 | 7 | 4 | 8 | 5 | 10 | 9 | 12 | 11 | 15 | 14 | 13 |
| Totals in criterion-based rating | 32 | 30 | 30 | 25 | 24 | 22 | 19 | 18 | 17 | 14 | 12 | 10 | 3 | 2 | 0 |

The first batch includes scripts F, V and S (30-32 points), the second, biggest group includes five writings (G, R, M, P, and H; 18-25 points), the third one comprises K and E (14-17), the fourth has L and O (10-12), whereas the last group includes three scripts (N, T, and Z; 0-3). According to her feedback, R5 also faced difficulties while ranking some of the writings, as they seemed to be on the same linguistic level, though one of them excelled in vocabulary range, while the other showed better cohesion or accuracy. This remark illustrates that during holistic assessment specific areas of linguistic knowledge are also considered.

In sum, concerning holistic assessment, four raters (R1, R3, and R4 & R5) arrived at similar conclusions: slight differences were hard to perceive, and small groups of writings fell in the same category. R1 demonstrated it by using joint placements (5=6, 8=9); whereas R5 expressed her opinion concerning students' unbalanced performance in separate linguistic areas. In the light of these opinions, it is not surprising that R5's first and second ratings

which relied on different methods resulted in a slightly different rank order of the scripts. It also reinforces Hamp-Lyons's (1990) observation that sometimes we make different judgments about the same writing on different occasions.

In spite of these changes in ranks, correlations between R5's two decisions are strong. Let us investigate correlation statistics on the total points awarded to scripts in the first criterion-based assessment, and rank orders established by the five raters during the holistic measurement. As Table 51 illustrates, totals have strong, significant ($p<.01$), negative relationships with each of the raters' rank orders; the strongest association joins total points with R2 (-.954) and R5 (-.938). The negative relationship is caused by the reverse connection of points and ranks. Best scripts received high points and low ranks, that is, the better the writing was, the lower the rank it was rated. In conclusion, R5's first and second ratings show very high correlation, and her assessments are also in line with the rest of the raters.

Table 51: Correlations between totals of criterion-based assessment and raters' rank orders

| | N=15 | Rank order 1 | Rank order 2 | Rank order 3 | Rank order 4 | Rank order 5 |
|---|---|---|---|---|---|---|
| Spearman's rho | Total (criterion-based) | -,887(**) | -,954(**) | -,915(**) | -,828(**) | -,938(**) |

** Correlation is significant at the 0.01 level (2-tailed).

With regard to criterion-based assessment, an important point of inquiry was whether raters would interpret the descriptors of the *CEFR* scales similarly to each other without the standardisation of criteria. As it is a European reference document, one of its main aims is to be comprehensive even for users with no specific training. I am aware, however, that professionals' international feedback (e.g., Weir, 2005, p. 281) has pointed out that in its present form the *CEFR* is not sufficiently comprehensive, coherent or transparent for uncritical use in language testing. An example for the variety of the numbers of bands for separate criteria was shown earlier (2.2.1), which weakens the comprehensiveness and transparency of the reference scales.

In this part, I return to the discrepancies observed in raters' judgments discussed in 4.8.4.4 and 4.8.4.5, focusing on the three raters (R2, R4 & R5) who showed no significant relationships with others in assessing certain areas.

The most conspicuous differences were related to R4. Fortunately, he provided a detailed description of his process of assessment, thus, we will see the reasons for disagreements in specific areas, first in assessing GLR.

The criterion of GLR consisted of five bands ranging from A1 to B1.2 (Table 52), concerning which R4 observed that there were far too many levels described for him to handle. Rather creatively, while neglecting the separating lines between the levels of the grid, he himself sub-divided the descriptor of A2.1 level into three separate bands. Thus, he added two more levels to the existing five, then, he rank ordered the seven linguistic levels. During this process he ranked the self-made new bands within A2.1 (level 4 out of 5) as levels 1, 2, and 5, where number 1 stood for the lowest level of proficiency. Thus, he created a mixed level, which, due to the discrepancy in levels, could not be considered at the registration of data, so all his questionable rankings in A2.1 were registered as A2.1. This, obviously, was not consistent with other raters' measurements, and resulted in low correlation coefficients in GLR.

Table 52: Ranking of descriptors of general linguistic range in *CEFR* by Rater 4

| CEFR level | R4's estimated linguistic level | **General linguistic range** |
|---|---|---|
| A2.2 | A2.2 ? | Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words. (4) |
| A1 | A2 | Has a very basic range of simple expressions about personal details and needs of a concrete type. (3) |
| B1.1 | B1.1 | Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interest, work, travel, and current events, but lexical limitations cause repetitions and even problems with formulation at times. (6) |
| A2.1 | *A2.2/ B1.1 ?*<br>*A2.1*<br>*A1* | *Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information.(5)*<br>*Can use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulae about themselves and other people, what they do, possessions etc.(2)*<br>*Has a limited repertoire of short memorised phrases covering predictable survival situations. (1)* |
| B1.2 | B1.2 | Has sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films. (7) |

Let us investigate how R4 rearranged descriptors: He identified "Has a limited repertoire of short memorised phrases covering predictable survival situations" as the lowest level of linguistic proficiency. He classified using "basic sentence patterns and memorised phrases", as part of A2.1 description, as the second lowest level. He categorised "Producing brief everyday expressions to satisfy simple needs of a concrete type" as level five out of the seven. He allocated the descriptor "Has a very basic range of simple expressions about personal details and needs of a concrete type" level three, that is, the third lowest language level. The difference between 'simple expressions' in the descriptor of A1 and 'basic sentence patterns' in A2 escaped the rater's attention. Otherwise, it is not surprising that he created a new rank order of linguistic levels, as descriptors do not seem to be comprehensive and transparent enough for identifying them. Concrete examples of typical vocabulary or structures would help identifying the different levels of proficiency (Huhta et al., 2002, p.131).

Although GLR is described in *CEFR* (Council of Europe, 2001) both in relation to vocabulary and sentence patterns, the majority of descriptors concern abilities of vocabulary use. R4's demonstration of the rating process proves that, as discussed under 2.3.3, the *CEFR*

provides little help in identifying the breadth of vocabulary needed to communicate at the various levels.

Regarding OWP this rater identified the levels correctly and suggested that a fourth level should be included, which would be placed under A1, worded ”Can use a list of fragments or words”. He allocated two writings (N & T) in this ’A1-’ category. His advice is in line with the Hungarian *NCC* (2003), that describes ’A1-’ level as approximately half of level A1, which means the ability to produce short, mainly word-level utterances that are not imbedded in syntactic structure.

In relation to vocabulary assessment, where A1, A2.1, A2.2 and B1 levels were described in the bands, he admitted that he could not establish the order of difficulty of descriptors. Accordingly, he interchanged the upper two levels. As for judging GRA, this rater claimed again that he was not able to use the grid, which was obvious from the rank ordering: he identified A2 as the lowest level of GRA, B1.2 as the second lowest, A1 as the third in the line, and B1.1 as the highest proficiency descriptor, thus his order compared to the one established in the *CEFR* was 3-1-4-2. R4’s decision making process confirms Jones’ (2002) claim that different people tend to understand ‘can-do’ statements differently, which suggests the need for greater precision and explicitness than is currently provided in *CEFR*.

In her comments R2 claimed that she sometimes had difficulties in identifying proficiency levels introduced by descriptors, but, as she added, she often faced this problem in using the complete version of *CEFR*. The evidence for her misinterpretation of the scale concerning OWP was demonstrated in both interrater and intra-rater correlation coefficients. She had no significant correlations with any of her peers in this linguistic criterion; moreover, none of her decisions on OWP were in significant relationship with her assessments regarding other linguistic areas. R2 ranked 60 percent of the scripts in A2 level, whereas in other areas she distributed performances along the scale in a more balanced way. For instance, 40 percent of writings were linked to level A1 regarding GRA, 33.3 percent in GLR, and 46.7 percent in VOC, whereas only 26.7 percent were ranked there in OWP. She placed 13.7 percent-40 percent of writings on level B along different criteria (GRA: 40%, GLR/VOC: 26.7%); the lowest ratio belongs to OWP.

Considering the fact that she classified many scripts both in A1 and B1 levels in other linguistic areas, I could not decide which level was meant by R2 when placing such a high number of scripts in A2. Therefore, I studied her rank ordering of scripts established during the holistic assessment, and found the following: Scripts K and L took places 10 and 11, whereas in OWP they were assigned to the highest (B1) level. As opposed to this, scripts

taking the first nine places in her rank order (scripts S, F, V, R, P, G, M, H, & E) were assigned to A2 level regarding OWP. The last four texts in her rank order (T, O, Z, & N) were placed in A1 level. From these data I deduced that R2 meant to place the best nine scripts in B1 and the ones ranked after them (scripts K & L) in A2. Thus, she must have exchanged A2 and B1 levels.

Rater 5's correlation concerning GRA and GLR was not significant, representing a possible misconception with regard to descriptors in these areas. The extremely high percent (53.3) of scripts placed in A1 level regarding GLR and VOC also pointed to interpretation problems of the scale.

Let us investigate R5's description of the rating process to clarify these disparities. In terms of GLR, it demonstrates that R5 interchanged three of the scale levels. She identified the lowest (A1) and highest (B.1.2) proficiency levels correctly, and misinterpreted the other three bands as follows: she mistook A.2.1 for B1.1, which is two levels higher, then, identified A2.2 as A.2.1, and B1.1 as A2.2; the latter two were perceived one band lower than the described level. Although "everyday expressions to satisfy simple needs of a concrete type" were part of the A.2.1 descriptor, it proved misleading for her in comparison with the expressions "lexical limitations/problems with formulation" in the descriptor of B.1.1, which made her perceive the latter a lower linguistic level. The descriptor of A2.2 requires the use of basic language to deal with everyday situations, which is very similar to the descriptor of A2.1. Fortunately, most of the scripts were allocated in level A1 by her, which level was identified correctly.

In terms of GRA, Rater 5 mistook B1.2 for A2.2, as in B.1.2 band reasonable accuracy was to be used in familiar contexts, with noticeable mother tongue influence and errors permitted. She perceived B1.1 as the highest linguistic level, in which a repertoire of frequently used routines was to be applied with reasonable accuracy in more predicable situations. As L1 influence and errors were involved in the description of B1.2, it seemed to represent a lower level compared to the other descriptor in which no errors were mentioned.

Summarising raters' assessments, in spite of the high reliability coefficients in general, there were a few areas where no significant relationships occurred between their decisions. These disagreements of opinions indicated misinterpretations of some descriptors of the *CEFR* scale, the reasons of which were scrutinised in this section, based on raters' written feedback on the rating process, on the one hand, and investigating their rating grids, on the other hand.

All differences demonstrated in correlation statistics proved to be misunderstandings of descriptors. This reinforces the stance that the descriptors of *CEFR* can be interpreted with great difficulties without concrete examples of language proficiency, thus, they may cause uncertainty regarding ranking learners' proficiency levels. This means that the comprehensibility, transparency and coherence of the *CEFR* are not sufficient, if the scales are used without standardisation.

While making decisions concerning students' linguistic proficiency, that is, in the interpretation process of descriptors, raters may have relied on their teaching and testing practice. It may be due to the their long years of teaching experience that most of their opinions showed strong relationships with each other, regardless of slight differences in the interpretations of the scale. The only exception is R4, who has only theoretical expertise, so his expectations must have been different from that of other raters'.

As a conclusion, it seems advisable to rely on *The Manual* (Takala, 2004), which describes four inter-related sets of procedures (familiarisation, specification, standardisation, and empirical validation) in order to design a linking scheme in terms of manageable activities. The activities carried out in the four sets of procedures contribute to the validation of the scales. It is a time-consuming process, but it can enhance the validity and reliability of assessments based on *CEFR* scales. I did not use this procedure as the aim of the investigation was to find out how raters' judgments are related to one another if they interpret the scales by themselves.

### 4.8.5 A comparative analysis of five students' texts: Corpus-based measurement of accuracy and vocabulary

In this section the following research questions will be answered:

RQ 9 How does criterion-based assessment compare with holistic assessment?

RQ 10 How are raters' opinions of sample tests related to one another?

RQ 11 What are typical performances like? What developmental levels do they reflect?

RQ 12 What range of vocabulary characterizes 8$^{th}$ graders? What is the relationship between criterion-based and corpus-based measurement of accuracy and vocabulary?

I aim to scrutinize five students' writings from several viewpoints, using both quantitative and qualitative methods. I will examine each script by

(1) discussing quantitative measures related to vocabulary,

(2) performing a qualitative analysis of the same criterion,

(3) introducing accuracy statistics of the coded texts,

(4) describing grammatical features qualitatively,

(5) scrutinizing task achievement and cohesion qualitatively.

After a thorough investigation of the scripts, raters' judgments will be examined in order to answer RQs 9 and 10, and to shed light on the relationship of assessors' opinions with quantitative measures on scripts, the interconnection of the five raters' decisions, and the consistency of the first and second rating of Rater 5. The issues of applicability of the Range Test (Nation, 2005) will also be discussed.

#### 4.8.5.1 Selection of texts

The five chosen scripts belong to the selected 15 that were re-evaluated by raters. The criterion of the selection was the overall score attained for the script during the analytic assessment, so that these typical L2 performances should represent a wide range of linguistic levels identified by various bands of the analytic scale. The marks awarded on the four criteria (task achievement, vocabulary, accuracy, and cohesion) were also supposed to reflect various levels of L2 performance. Thus, in order to introduce different levels of L2 proficiency, I selected scripts with overall points (0-32) decreasing by similar intervals. Accordingly, Script F (32 points) represents high-achievers, Script R illustrates an above average performance (24 points); Script K is slightly above the pass-level with 17 points, whereas Scripts O (10 points) and T (2 points) belong to the under-achiever category.

#### 4.8.5.2 Application of Range test

Let me start my comparative analysis by presenting statistics on vocabulary range and frequency displayed in students' texts run on the Range test (Nation, 2005). The texts were run on four base lists, which are also called word lists (WL), so they will be used interchangeably in the analyses.

Out of the four components of lexical richness (Read, 2000) introduced under 1.4.3 of this study, two types of TTR - lexical variation and the ratio of lexical words (TTRL) - lexical density (LD), and lexical sophistication (LS) were calculated. In this picture-comparison task students did not exhibit major differences in LS at this proficiency level (A1-B1); nevertheless, calculations were performed.

The instructions of the Range programme warn users to correct spelling mistakes and rewrite contracted verb forms in full before running the software, as otherwise the programme would place misspelt words in the 'not in the lists' category, and words separated by hyphen / apostrophe would be counted as unique items, and thus the results would be distorted. To illustrate this with two examples:

(1) the contracted form *can't* is listed as two words when run on this programme: *can* and *t*: the contracted form of *not* appears as a separate word in the list.

(2) Prior to the correction of mistakes, misspelt words, like *anithing, newspapier,* and *plaiing* were also listed in the 'not in the lists' group of Range test, which meant that these vocabulary items were not counted in the relevant WL, thus, the TTR and LS data changed.

In most cases it was obvious what the students meant to write: most of the spelling mistakes included one misspelt letter, as in *anithing* and *barbie;* transposing letters (*kicthen, colc'k, folwer);* writing a word as two separate vocabulary items (*new papers),* or altering more letters so that a word reminded me of various other words. For example, *tham* could easily be *then* or *them.* Fortunately, the context always provided sufficient clues for the correction. So, in the sentence (Script O) *He used to play the … Tham play the cat,* it was certainly intended to be *then*.

Thus, in order to prevent the emergence of unreliable data, I corrected all spelling mistakes in the five texts and changed contracted verb forms to full forms. The original scripts were saved though, for further analysis.

The next step was the creation of a Stop list, in order to exclude grammatical words from the scripts, and obtain the number of lexical words in them. The Stop list I compiled for this text contained the following 17 words that were needed for the description of pictures: *A, a, an, are, at, B, be, can, in, into, is, on, out of, the, there, to,* and *with.*

The other feature that can be surveyed by Range is the order and measure of frequency of each word in the script, so the Frequency test was also applied to these writings.

Table 53 summarizes the data gained on lexical richness of the five scripts and of the model text (M) in order to be able to compare them easily; nevertheless, each data will be attended to in the detailed analysis of scripts. The table contains data on types and tokens used in each script, and the ratio of these (TTR). Then, after the exclusion of function words, lexical types and tokens are counted, followed by their ratios. Lexical density is calculated by dividing the total number of lexical words by the tokens of the text. Lexical sophistication shows the rate of types belonging to the most frequently occurring one thousand words (WL 1) to those of less frequent WLs in the text (WLs 2, 3, & 4).

Table 53: Data describing lexical richness of the five scripts and the model text

| Data on lexical richness /script | F | R | K | O | T | M |
|---|---|---|---|---|---|---|
| Types | 58 | 49 | 43 | 47 | 32 | 55 |
| Tokens | 190 | 190 | 104 | 117 | 125 | 166 |
| Type -token ratio (TTR) | 0.31 | 0.26 | 0.41 | 0.40 | 0.26 | 0.33 |
| Lexical types | 49 | 42 | 36 | 38 | 27 | 44 |
| Lexical tokens | 99 | 83 | 57 | 72 | 69 | 77 |
| Type -token ratio of lexical words (TTRL) | 0.50 | 0.51 | 0.60 | 0.53 | 0.39 | 0.57 |
| Lexical tokens /tokens (LD) | 0.52 | 0.44 | 0.55 | 0.62 | 0.55 | 0.46 |
| Lexical sophistication (LS) | 0.14 (7/51) | 0.14 (6/43) | 0.13 (5/38) | 0.12 (5/41) | 0.14 (4/28) | 0.15 7/48 |

#### 4.8.5.3 Script F

1 In picture A the boy is playing with his toy but in picture B he is playing with his cat.
2 In picture A the girl is playing with a doll but in picture B she is playing with a toy bear.
3 In picture A the man is reading a book but in the other picture he is reading a newspaper.
4 In the first picture the woman is going in the room but in the other she is leaving it.
5 In the first picture there are two pictures on the wall, but in the other picture there's only one.
6 The difference between the two TV-s that in picture A it's switched on but on the other picture it's switched off.
7 In picture A there isn't a clock near the TV but in picture B there is.
8 In picture A there is only one plant but in picture B there are two.
9 In picture A the window is closed but in picture B it's open.
10 In the first picture the weather isn't so good because it's raining but in picture B the sun is shining.

This text was ranked best among the 15 scripts during the first analytic assessment: the maximum score (8) was assigned on each criterion. This result suggests that the script met the criteria of the top band descriptors of the scale: it contained relevant information on all things shown in the pictures, and used a wide range of vocabulary. Concerning accuracy, the text contained only few mistakes not disturbing comprehension. As for cohesion, the text was well structured: parts on different things were separated and the complex sentences were logically linked. Let us investigate how statistical data support this evaluation.

Results on the Range test regarding vocabulary are shown in Table 54. Script F contains 58 different words in a 190-word-long text; therefore, the type-token ratio representing LV of

the script is 0.31. Regarding lexical sophistication, 51 of the running words belong to the first WL making up 88 percent of the whole script, whereas about 12 percent of lexical words come from the second to fourth lists; the less frequently used vocabulary groups, thus producing the LS ratio 0.14. The text contains five words from WL two: *newspaper, plant, raining, toy, weather*, two of which (plant, weather) were included in the prompt; one word from WL three: *shining*; and one belonging to WL four: *doll*. The model text contains almost the same words belonging to the respective WLs, except for *weather*, which is not used in it, and the additional application of *teddy*, which belongs to WL four.

In order to obtain the second selected measure of lexical richness, TTRL, I excluded grammatical words by running the programme with the Stop list switched on (Appendix G). Thus, altogether 99 lexical words representing 49 different words remained, the ratio of which (.50) expresses the TTRL. The model text has a similar TTR, but a slightly higher TTRL, which points to fewer repetitions of the same lexical word.

The third measure, the ratio of lexical words to the total number of words used in the script, shows LD; in this script it is 0.52. The most frequently used words in Text F are *in* (20 times), *is, picture,* and *the* (19), *A/a* (12), *but* (10) (Appendix H). Altogether, these six types provide 52 percent of all the 190 tokens of the script. This frequency pattern is different from the general patterns in large corpus analyses, in which function words make up the most frequent words (e.g., The Brown Corpus; The LOB Corpus, cited in Horváth, 2001, pp. 115-116). The discrepancy was obviously induced by the task: students were required to compare two pictures, so the word *picture* and the letters *A* and *B* were naturally emerging ingredients of the comparative sentences in a number of scripts; the length of scripts was an additional condition. The following items are used three to six times: *B,* and *there* (6-6 times), *it* (5), *other, playing, with* (4), *first,* and *on* (3), the cumulative percent of these 14 types is 70.5 percent, that is, they constitute more than two thirds of the text, whereas the remaining 44 words cover nearly 30 percent. As shown in Table 54, Script F displays four word families fewer than types, which difference implies the forms of *be* (is, are) and the plurals of *picture* and *TV*. Comparing the word frequency of the model text with that of this text, striking similarities occur; 13 types cover 70 percent of the model script, out of which *in, is, picture, the, A,* and *B,* provide 51 percent, whereas *but, there, playing with, are, first,* and *it* give another 19 percent. Lexical density is lower (0.46), whereas LS (0.15) is slightly higher in the model text than in Script F.

Table 54: Script F: Range of vocabulary based on Range test, Tokens, types, and families in Word lists 1-4; lexical words

| WORD LIST Script F | TOKENS / % | TYPES / % | FAMILIES |
|---|---|---|---|
| One | 182 /95.79 | 51 /87.93 | 47 |
| Two | 6 / 3.16 | 5 / 8.62 | 5 |
| three | 1 / 0.53 | 1 / 1.72 | 1 |
| four | 1 / 0.53 | 1 / 1.72 | 1 |
| Not in the lists | 0/ 0.00 | 0/ 0.00 | |
| total | 190 | 58 | 54 |
| Lexical words | 99 | 49 | 46 |

After reviewing data on lexical richness, let us have a closer look at the quality of vocabulary displayed in the text. The number of words (190) used in Script F exceeds expectations: all required items and additional words are used, such as *leave, difference, and switch on / off*. The items belonging to WLs 2-4 have been shown among the results of the Range test. In sentence 10 (s10) the expression describing the weather *(isn't so good*) seems to reflect transfer from the mother tongue for stating that it *isn't very good*. This is an example for the permeable and transitional features of interlanguage, in which new linguistic forms are constructed internally through processes such as over-generalization, omission or transfer (Ellis, 1977, p. 33).

Statistics on accuracy show that Script F contains no mistakes in the investigated language areas; it includes ten correct PRC forms, two expletives in singular (*there is*), two in plural form (*there are*), and one plural noun. There are two correct negations in the text. Thus, the correctness ratio is 1, which demonstrates that all four analyzed language forms have been acquired.

The qualitative analysis of accuracy in Text F shows that the use of PRC, EXP, the copula *be* and NEG is excellent. One exception to this is demonstrated in sentence 6 (s6) where the student employs a lead-in sentence to call the reader's attention to the differences between the two TVs. Here the copula *be* is missing from *the difference is that.* This may reflect mother tongue influence, as in Hungarian the equivalent of copula *be* is not used in 3rd person singular. His interlanguage *be* is still developing, as (in s10) he correctly provides *be* in *the weather isn't so good*. In s6 a more complicated structure, passive voice, is correctly implemented. The use of the prepositional expression *in the picture* is incorrect only once (s6) where it is replaced by *on the other picture*. It may be another exhibition of L1 transfer, illustrating that interlanguage is variable; at any stage of development different forms for the

same grammatical structure are used. This variability in the use of *be* and the preposition *in* reveals the existence of competing rules of interlanguage.

Task achievement of Script F is excellent, as it reports on all items appropriately. The relevance is manifested mostly through one of the other criteria; appropriately used vocabulary, which has been examined.

As regards cohesion, the script demonstrates that its writer not only planned the structure of the text well, but their meaning-making activity is also socially purposeful. This student can be classified as a good writer who has tacit knowledge of formal features of reader-based prose (Flower & Hayes, 1980), thus they can rely on more options to organize their thoughts and ideas.

The organization of thoughts is manifested in not following the item order of the prompt automatically. Instead, the learner logically groups people (sentences [ss] 1-4) and objects (ss 5-8) in separate clusters (paragraph-like units), then writes about the *window* and the *weather* in a new cluster (ss 9-10), demonstrating the connection between them. This grouping represents background knowledge of families, their home activities, and the conformation techniques related to weather conditions (opening/closing windows).

Apart from logical linking of items, cohesion is ensured by exploiting numerous cohesive devices. Here I give short definitions of a few terms that will be used in the analyses of scripts. (1) Repetitions occur when a lexical item that has already been used in a text is repeated; they form two tokens of a type. There are other types of textual connection that serve the same function, for example, anaphors. (2) Anaphora is a process where a word or phrase refers back to another word which was used earlier in a text (Richards et al., 1992); we will see personal pronouns taking this role in the following scripts. (3) Ellipsis is the absence of some required stretch of language that has to be supplied by the reader to make sense of the sentence encountered (Hoey, 1996, p. 74).

In script F complex sentences are created by the coordinating conjunction *but* connecting all ten pairs. Lexical references, among them repetitions, are employed in various ways. Here are five pairs of variations for expressing the differences in the two drawings:

- *in picture A – in picture B* (in ss 1, 2, 7, 8, & 9),
- *in picture A – in the other picture* (in ss 3, & 6),
- *in the first picture – in the other* (s4),
- *in the first picture – in the other picture* (s5),
- *in the first picture – in picture B* (s10).

Ellipsis is applied to express the presence or absence of the clock, and to report on the discrepancy in the number of plants:

- *there isn't a clock – there is* (s7),
- *there is only one plant - there are* two (s8).

Anaphoric references are employed throughout the script; they either refer back to the person in the previous clause, or to the previously used noun, for example:

- *the boy is playing – he is playing* (s1),
- *the boy is playing with his toy* (s1),
- *the woman is going in - she is leaving* (s4),

or substitute the names of objects, thus the anaphor *it* is applied five times, for example:

- *going in the room – leaving it* (s4),
- *two TVs – … it's switched on, …it's switched off* (s6),
- *the window is closed – it's open* (s9).

The application of these devices makes the text easy to follow, and thus, reader-friendly.

Summarising quantitative and qualitative data, I can claim that this script represents linguistic knowledge higher than required (A1) in the final grade of primary education. This also certifies that a task estimated to measure language proficiency at A1-A2 level can elicit higher level learner performances (*EALTA* discussion, 2005). Irrespective of the deliberate simplicity of the task, the text is constructed in a thoughtful way characterizing good writers aware of their readers. Ratings concerning *CEFR* levels (Council of Europe, 2001) reinforce this stance.

As Table 55 illustrates, three raters ranked this script second, one fourth, and one the best text of all. Out of the 21 ratings presented in the table only seven put this writing on level A2, the majority (14) place it on B.1. Raters' decisions varied most concerning linguistic knowledge, as two of them matched this text with A2.1, whereas three ranked it on B1.2 level. Surprisingly, none ranked this script as level A2.2, which is situated between these linguistic levels. Considering that in other areas raters agreed to a high degree, this discrepancy suggests that some of them may have misinterpreted the descriptors.

Concerning OWP, which criterion comprised descriptors regarding cohesion, the script was classified as level A2 by two raters, and level B1 by three raters.

Two raters assigned the VOC of this text to level A2.2; two linked it to level B1 of *CEFR* (Council of Europe, 2001), with one opinion divided between the two, meaning that the rater placed the script in both bands. Raters agreed most in accuracy ranks; all of them assigned it

to B level: four of them considered that the writing achieved B1.1 level of *CEFR*, and Rater 2 placed it even higher, on B1.2. Rater 1 had the highest opinion of this text; she ranked it in level B1 along all criteria, whereas Rater 4 linked it to A2 in three.

Table 55: Raters' decision on script F: rank order and *CEFR* levels

| Criteria / Raters | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|
| Overall rank | 2 | 2 | 2 | 4 | 1 |
| General linguistic range (GLR) | B1.2 | B1.2 | B1.2 | A2.1 | A2.1 |
| Overall written production (OWP)/cohesion | B1 | A2 | B1 | A2 | B1 |
| Vocabulary range (VOC) | B1 | A2.2-B1 | A2.2 | A2.2 | B1 |
| Grammatical accuracy (GRA) | B1.1 | B1.2 | B1.1 | B1.1 | B1.1 |

Rater 5 placed this text first among 15 again, as in the first criterion-based assessment, and ranked it on B1 along three criteria, while concerning the linguistic range, she categorized the script as belonging to A2.1 level of the European reference scales for languages. The maximum points given in the first, criterion-based assessment are in line with this.

#### 4.8.5.4 Script R

1 In the picture A, the man is reading a book, but in the picture B, he is reading newspapier.
2 In the picture A the woman is coming in the room, but in the picture B she is going out from the room.
3 In the picture A the boy is plaiing with a little castle, but in the picture B he is playing with a cat.
4 In the picture A is two picture on the wall, but in picture B is one picture on the wall.
5 In the picture A isn't a clock on the table, but in the picture B is a clock on the table.
6 In the picture A is three plants, but in the picture B is four plants.
7 In the picture A is we can see something in the TV, but in the picture B we can't see anithing.
8 In the picture A the girl is playing with a pupet, but in B she is playing with a bear.
9 In the picture A is the window close, in the B is close.
10 In the picture A the weather is rainy, in the B is sunny.

Script R was ranked fifth in the first assessment with a total of 24 points, which placed the writing in the 'above average' category, as the mean score of 231 scripts was 17.26. The script showed uneven language proficiency levels across the four criteria; accordingly, it received different points --seven on task achievement, five on vocabulary range, five on accuracy, and seven on cohesion.

On scrutinising the scale, we can see that regarding task achievement the script falls in the top band, as it describes 9-10 things relevant to the pictures. The five points for vocabulary place this writing in the lower end of Band 4, which is characterized by the use of a wide

scale of relevant vocabulary. As for accuracy, the five points indicate that the whole text is comprehensible in spite of the few mistakes committed. Top-band cohesion score is achieved by logically linked sentences and structural variety, which is attained by employing more than three sentence-types.

As Table 56 illustrates, Script R is the same length (190 tokens) as Script F. Nevertheless, the number of types (49) is smaller in this script, which results in a lower TTR (.26). The ratio of lexical words, though, is slightly higher in this script (.51) indicating that lexical words are repeated fewer times.

The distribution of lexemes in different base lists is similar to that of Script F, so the LS index is the same in both of them (.14). One exception concerns Base list 4: this writing does not contain words belonging to that list. Nevertheless, it contains four words from WL 2: *plants, rainy, weather, newspaper*; one lexeme from WL 3: castle; and the word *puppet* were not found in the lists. When I ran the programme again on a higher number of base lists, it was found belonging to a list of lower frequency items (WL 6). Prior to my correction of the spelling mistakes, the words *anithing, newspapier,* and *plaiing* were also enumerated in the 'not found' category of Range test.

LD in this script it is 0.44 percent, lower than that of Script F, which points to the presence of a higher number of function words. The five most frequently used words are *the* (31 times), *in* (22), *picture* (19), *is* (18), and *A/a* (17), which total 56 percent of all tokens of the script. The extraordinarily high frequency of the definite article (31) marks its overuse. This hypothesis will be confirmed during the qualitative analysis. In the table we can see that the numbers of word families and types are equal, which indicates that only one form of each lexeme is applied, for example, only one conjugated form of the verb *be* appears in the script, and, also, that each word is used either in singular or plural form.

Table 56: Script R: range of vocabulary based on Range test:
Tokens, types, and families in Word lists 1-4; lexical words

| WORD LIST<br>Script R | TOKENS / % | TYPES / % | FAMILIES |
|---|---|---|---|
| One | 183 / 96.32 | 43 / 87.76 | 43 |
| Two | 5 / 2.63 | 4/ 8.16 | 4 |
| Three | 1 / 0.53 | 1/ 2.04 | 1 |
| Four | 0 | 0 | 0 |
| Not in the lists | 1/ 0.53 | 1/ 2.04 | |
| Total | 190 | 49 | 48 |
| Lexical words | 83 | 42 | 41 |

The text contains most of the necessary words to describe the pictures, though some of them are misspelt, as it was discussed in the computer-based analysis; some others are used inappropriately; others are overused.

The second part of the erroneous form *newspapier* (s 1) recalls the Hungarian word for *paper,* so it may indicate L1 transfer, but it might be *papier* from German as well. *Going out from* (s 2) also seems to be a translation of the mother tongue expression. The words *table* (s 5) and the misspelt *pupet* (s 8) are not relevant to the objects shown in the picture; *TV stand* and *doll* would be more suitable solutions. In sentence 7 the statement *we can see something in the TV* might be the application of an avoidance strategy compensating for not knowing *be on / be switched on.*

In sentence 9 the ungrammatical form of *close* (instead of *closed)* is used twice. For the second time, surprisingly, it is related to the open window. There are several possible explanations for this: the student either may have failed to remember the adverb *open* and wanted to use negation (not closed) as a compensation strategy, but absentmindedly missed the word *not*, or they might have been in a hurry, and might not have paid attention, and thus, they repeated the expression used in the first clause.

Returning to the issue of the overuse of the definite article, we can see that it is used 18 times inappropriately in the expression *in the picture A/B* throughout the script, nevertheless, twice (ss 4, 8) it is employed correctly. By this observation it is proved that the lower value of LD (.44) can be explained by the high frequency of the erroneously used definite article. Apparently, this structure is at a lower stage of development in the student's IL continuum.

This close analysis of Script R proves that similar lexical ratios cover different linguistic levels, as ratios fail to reveal inappropriacy in usage or spelling mistakes. Thus, the differences between texts cannot be shown by merely counting lexical ratios; criterion-based assessment or detailed analysis reveals more.

Error-tagging shows that Script R includes 12 correct forms of the four examined language forms, and altogether 23 mistakes (*other mistakes* included), thus the accuracy ratio is 0.34, relatively low compared to points awarded for the script. The high number of mistakes comes from the overuse of the article *the,* and the presence of misspelt words. According to the instruction of the analytic scale, repeatedly occurring mistakes were not to be considered more than once, accordingly, the points awarded to the text do not reflect the number of mistakes. Nevertheless, the accuracy ratio does, hence the disparity. This disparity of the accuracy ratio and the assessment raises the issue of the reliability of unbalanced error

count in tagged texts. By unbalanced I mean the count of errors regarding certain linguistic areas without the count of correct solutions of the same features.

As for the scrutinised language forms, the text contains eight correct PRCs, and six incorrect EXPs, all of which lack the expletive *there* (e.g., *in the picture A is two picture on the wall* in s4, *in the picture A isn't a clock on the table* in s5). The complete absence of the EXP *there* may reflect mother tongue influence, as in Hungarian it is sufficient to use the verb *be* for stating the existence or presence of an object. The use of plural nouns indicates that this form is still developing, as it appears twice correctly *(three/four plants* in s6) and once erroneously *(two picture,* s4). The negative construction is used twice correctly *(isn't a clock* in s5, *we can't see* in s7). Error tagging points to three verb agreement errors *(is two picture* in s4, *is three/four plants* in s6) and 13 other mistakes, a few of which were discussed in connection with vocabulary (e.g., *close* in s9, *going out from* (s2).

An additional example for this type is the omission of the indefinite article *(he is reading newspapier* in s1). On all other obligatory occasions it is applied appropriately (e.g., *reading a book* in s1, *with a little castle/with a cat* in s3). The omission may reflect transfer from the L1, as in Hungarian the absence of definite objects is indicated by the use of intransitive verb forms that make the application of indefinite articles unnecessary. In sentence 7 the first clause *(In the picture A is we can see …)* includes a dispensable form of *be* (*is)*, which might point to the lack of revision in the writing process. Another omission can be spotted in sentence 10 *(in the B is sunny),* the EXP *it* is missing.

Summing up the grammatical features of this text, we can claim that NEG and PRC are acquired, whereas PLU is well on the way with more correct than incorrect forms (67%). EXP, on the other hand, seems to be at the beginning phase of development, as no correct applications are demonstrated in the script. This developmental pattern is in line with the one characteristic of low-achieving schools (4.8.2.2) in this corpus, according to which present continuous tense is used correctly earlier than expletives.

Task achievement deserves good scores, as the script covers all required items appropriately. The text is mostly coherent; it relies on a few cohesive devices, although the arrangement of statements is not as logical as in Script F, and it does not follow the order suggested by the prompt either, which is not reader-friendly.

The description is started by reporting on people (man, woman, boy), but the girl is not included in this group; she is introduced later, in sentence eight. The observations about the wall and the clock are sequential, but the TV is described later, after the plants, which order

seems not to follow any logical directions; it reveals an accidental sequence. At last, the window and the weather are logically grouped, as it is implicitly proposed by the prompt.

In the first eight sentences the co-ordinating conjunction *but* connects clauses, whereas in the last two statements the connection is established by repetition. Lexical references are used in two forms:

- *In the picture A – in the picture B*
- *In the picture A – in B/in the B*

Anaphoric references concerning people are employed four times in the script, but the anaphor *it* is not applied to substitute objects, for example:

- *the man is reading a book – he is reading,*
- *the woman is coming in – she is going out.*

In sum, a few appropriately used cohesive devices support the coherence of the text; on the other hand, missing anaphoric references and the partly accidental sequence of picture description make the reading process of this text less fluent than that of Script F.

Reviewing the performance along the four criteria again, we can see that the writing reports about all items using mostly appropriate but sometimes misspelt vocabulary. The numbers of both lexical types (42) and tokens (83) are fewer than in Script F (49; 99); nevertheless, their respective ratios are almost identical (Script R .51, Script F .50). This result reveals that the mere count of ratios concerning lexical variation is not sufficient for judging the quality of lexis of texts. As for grammar, the investigated language forms are mostly acquired, with the exception of EXP. Cohesive devices employed in the text help in understanding it, but at places the absence of these markers slightly hinders comprehension. In the light of this analysis I think that 7 points provided for cohesion in the first rating are too high for this performance. IL development can be traced in the application of competing forms in parallel (*in B / in the B; plaiing /playing*) and the complete absence of the EXP *there /it.*

As shown in Table 57, this script distributed raters' opinions the most, as they ranked it between the third and seventh places with each rater assigning it to a different place. It received a total of eight ratings (40%) in B1 category, 11 in A2, and one in A1, as compared to 14 B1 and seven A2 levels of Script F. The difference between raters' severity can be observed in a few occasions. Although both Rater 4 and Rater 5 classified this script mostly in A2 level, Rater 4 ranked it third, while Rater 5 placed it seventh among the 15 writings. The majority of other assessors' rankings linked this writing with B1 level. The only example for

total agreement is in terms of OWP, as each rater assigned Script R to A2 level. In sum, Rater 5 seems to have judged the script the most severely, whereas Rater 1 placed it in the highest proficiency category (B1 three times out of four).

Rater 5 was more critical the second time; she placed the script two places lower than in the previous rating. Concerning GLR this script is ranked in A2.2, which is higher than that of Script F (A2.1), which points to the misinterpretation of the scale, as the rank order (F: 1st, R: 7th) and points of Script R (F: 32, R: 24) are far below those of F. According to retrospection and the description of the assessment process (in 4.8.4), Rater 5 did misinterpret A2.2 level for A2.1.

Table 57: Raters' decision on script R: rank order and *CEFR* levels

| Criteria \ Raters | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|
| Overall rank | 5 | 4 | 6 | 3 | 7 |
| General linguistic range (GLR) | B1.2 | B1.1 | B1.1 | A2.1 | A2.2 |
| Overall written production (OWP)/cohesion | A2 | A2 | A2 | A2 | A2 |
| Vocabulary range (VOC) | B1 | A2.2-B1 | B1 | A2.2 | A2.1 |
| Grammatical accuracy (GRA) | B1.2 | B1.2 | B1.2 | A1 | A2 |

#### 4.8.5.5 Script K

Picture A:

1 The TV is on.
2 The father is read book.
3 The boy is play with cubes.
4 The mother is go in the room.
5 The room have two glasses.
6 The girl is with a doll.
7 The weather outside is rainy.
8 No clock near by the TV.
9 The window is closed.

Picture B:

10 The weather outside is sunny.
11 TV is off.
12 The clock is near by the TV.
13 The mother is go out in the.
14 The room have one glasses.
15 The father is reading a news paper.
16 The boy play with the cat.
17 The girl is play with a Teddy Bear.
18 The window is open.

This writing was ranked ninth, based on its merits by the analytic scale, and it presented an almost even profile across the investigated language performance criteria. Task achievement was provided five points, whereas the other three criteria were awarded by four points each, adding up to 17, which is near the mean score for the whole cohort (17.26). This script represents a typical performance slightly exceeding pass-level (12-16 points). It describes nine things relevant to the prompts with mostly appropriate lexis, and contains several mistakes that do not disturb comprehension. As regards cohesion, the script displays two sequences of sentences with one or two sentence types repeated.

Table 58 shows that script K contains 43 types in a 104-word-long text, the TTR of which is 0.41. It is interesting to note that this ratio is higher than those of script F (.31) and R (.26). It is due to the extremely short text, which is the shortest script among the five. The distribution of types in WLs 1-4 is similar to that of the previously presented higher quality writings, whereas the ratio of lexical words is much higher (.60) than in those scripts (F: 0.50, R: 0.51). This high TTRL can be explained by the brevity of the text again, in which one of the lexical words, *picture,* is used only twice, as opposed to its application 20-22 times in the previously investigated writings. LD, the ratio of content lexemes and all tokens, in this script is 55%, higher than in the two previously examined writings and the model text, which points to the absence of function words. As Hyltenstam (1988) warns us, LD may be a misleading measure at low levels of proficiency, where students tend to omit function words. In these cases the ratio of content words becomes high, suggesting high lexical richness, although it actually reflects the writer's inability to construct a coherent text.

Table 58: Script K: range of vocabulary based on Range test:
Tokens, types, and families in Word lists 1-4; lexical words

| WORD LIST Script K | TOKENS / % | TYPES / % | FAMILIES |
|---|---|---|---|
| One | 98/94.23 | 38/88.37 | 36 |
| Two | 3/ 2.88 | 2/ 4.65 | 2 |
| Three | 1/ 0.96 | 1/ 2.33 | 1 |
| Four | 2/ 1.92 | 2/ 4.65 | 2 |
| Not in the lists | 0 | 0 | |
| Total | 104 | 43 | 41 |
| Lexical words | 57 | 36 | 34 |

Regarding low-frequency words in Script K, two of them (*rainy, weather*) belong to WL two; one (*cube*) to the third thousand most frequent lexemes; and two names of toys, *doll* and *teddy*, belong to the next list (WL 4). As for frequency features of the text, seven types: *the* (21), *is* (14), a/*A*, *TV* and *with* (4), *play* and *room* (3) constitute 51 percent of the 104 tokens. It stands out that lexical words (*TV, play*, and *room*) occur within the most frequently used lexemes, which points to a discrepancy from the general tendency of function words taking first positions in frequency. The next 19 percent of the script involves *boy, by, clock, father, girl, glasses, in,* and *mother*, each of which types is applied twice, and thus, 15 types cover 70 percent of the writing.

The text contains most of the necessary words to describe the pictures, though one of them is misspelt (*news paper* in s15), and some others are used inappropriately. In sentences 5 and

14 the pictures on the wall are described as *glasses.* The student may have meant *mirrors* (*looking-glasses),* as the paintings on the wall do not illustrate anything, so they may have been perceived as mirrors. In sentences 8 and 12 redundant prepositions (*by, near*) are applied to express the spatial relationship between the clock and the TV. In sentence 13 there are two lexical inadequacies: one is the incorrect phrase *go out in,* the other is the missing noun at the end of the sentence denoting the place where the mother is going. Nevertheless, this text also contains rarely occurring items (c*ube,* 7 times; *teddy bear,* 45) in the corpus. *Teddy bear* belongs to WL 4, and it is applied in about one fifth of the scripts.

Error-tagging shows that script K includes four correct forms of the four examined language forms, and altogether 16 mistakes (*other mistakes* included), thus the accuracy ratio is 0.2, which is relatively low compared to points awarded for the script (17). This ratio indicates the inconsistency of counting errors with no attention paid to correct forms.

In terms of accuracy, PRC is employed seven times, including only one correct application (s15), whereas the combination of the auxiliary *be* and the simple verb form (e.g., *is read* in s2, *is go* in s4) occurs five times, and the simple verb form is used once (*boy play* in s16). This script is exceptional in this respect: there are few writings in the corpus, in which three different forms of PRC occur. This variability of forms illustrates that in the child's interlanguage system three competing forms of expressing PRC exist, and unfortunately, the correct form is still underrepresented among these.

The application of the verb *have* is unexpected, especially in relation to the pictures on the wall (*The room have two/one glasses*, ss 5, 14), and it reveals another still developing area of interlanguage; the third person singular form of the verb is incorrectly replaced by the simple verb form. This is not surprising, though, as it is a late-acquired morpheme (Brown, 1973; Dulay & Burt, 1973), and this student cannot even use the present progressive form, which, according to the morpheme studies, is an earlier acquired morpheme. As for EXP, it also needs to be improved, as there is one attempt to apply it in the script, which is incorrect: *No clock near by the TV* (s8). This form not only lacks the EXP *there,* but also the verb *be*. Plurals are represented with 50 percent accuracy: there is one correct (two glasses, s5); and one erroneous form (one glasses, s15) in the script. This type of incorrect usage (overuse) is exceptional in the corpus, as few students used the PLU when it represented one object after the numeral *one*. NEG was performed once (*no clock*, s8), and it was coded as correct in spite of the mistake. The error was counted as the absence of the EXP *there* and the verb *be*. *Other mistakes* in this writing included inappropriate vocabulary use and the missing verb *be.*

In sum, the acquisition of the four forms is at the lower end of the interlanguage continuum, especially in the case of EXP, whose correctness ratio is zero. PRC is not acquired either (0.14). NEG is used correctly, PLU both correctly and erroneously. Regardless of the low results on the investigated forms, the text included several completely correct sentences (e.g., *The window is closed.,* in s9; *The weather outside is sunny.,* in s10).

Scrutinising task achievement, we can observe that except for the plants, all items have been covered in the text, and the relatively low point awarded was due to misused or missing vocabulary.

Text cohesion is not high partly because of the two strings of unconnected sentences related to the two pictures. The sentences concerning pictures A and B are placed beside one another, which could be a logical arrangement for comparison, and also, a good excuse for not repeating the expressions *in picture A/B* ten times in a row if the sentences beside one another were describing items in parallel. Unfortunately, this is not the case: items from the prompt are selected mostly in random, so the reader cannot establish coherence easily. In this separated introduction of the pictures cohesive devices are not applied, the student apparently did not have the reader in mind, just accomplished the task by showing differences between the two pictures.

Four of the raters ranked Script K (Table 59) 10$^{th}$ and 11th, whereas Rater 3 ranked it 7$^{th}$. Out of 21 placements along the four criteria it has 19 A (9 A1 and 10 A2) and two B1 levels, thus 57 percent of the ratings classify it higher than A1. Outlying decisions are as follows: (1) in terms of OWP, the categorisation of the text into B1 on part of Rater 2 against four A1 levels, and (2) the B1.2 accuracy rank given by Rater 4 among two A1 and two A2 ranks. These decisions again point to misinterpretations of the scale.

Table 59: Raters' decision on script K: rank order and *CEFR* levels

| Criteria \ Raters | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|
| Overall rank | 11 | 10 | 7 | 11 | 10 |
| General linguistic range (GLR) | A2.1 | A2.1 | A2.1-A2.2 | A2.2 | A1 |
| Overall written production (OWP)/cohesion | A1 | B1 | A1 | A1 | A1 |
| Vocabulary range (VOC) | A2.1 | A1 | A2.1 | A2.2 | A1 |
| Grammatical accuracy (GRA) | A1 | A1 | A2 | B1.2 | A2 |

Rater 5 ranked the script tenth this time, whereas in the first assessment she ranked it ninth, which is a small difference. The 17 points awarded to Script K, which exceed the estimated pass-level (12-16 points) according to the validated rating scale, are in line with the three A1 and one A2 levels assigned to this writing in the second assessment.

Let us investigate if the ratings concerning OWP are in line with points on cohesion. It was identified as A1 level by Rater 5 (‘can write simple isolated sentences’). Cohesion received four points in the first rating, as the middle band, which is worth 3 or 4 points, requires a sequence of sentences with one or two sentence types; therefore, the two decisions seem consistent.

#### 4.8.5.6 Script O

1 He used to play the … Tham play the cat.
2 She used to play the barbie tham play the bear.
3 The woman used to come on the door tham went to the „kicthen”.
4 The man in the A picture is read a book and in the B picture read a new papers.
5 I can see on the A picture the weather is raining any B is sunny.
6 The first pictures is two plants and second pictures is one plants.
7 I can’t see on first pictures the colc’k tham I can see on second pictures the colc’k.
8 It’s three o’clock.
9 The first pictures the boy behind can see folwers than second pictures the boy behind can’ see flowers.

Text O was 12th in the criterion-based assessment, as it presented a similarly low profile across the investigated criteria. It was ranked in the lower part of the third band in terms of task achievement and cohesion (3 points), and even lower regarding vocabulary use and grammatical accuracy (2 points). With a total of ten points this student belongs to the under-achiever category (below 12 points). According to the scale descriptors, the text is about seven-eight partly relevant things; it relies on a limited scale and often inappropriately used vocabulary; many mistakes occur, so only part of the text is comprehensible; the same sentence type is repeated.

As Table 60 illustrates, Script O is longer than Script K, as it contains 47 types in a 117-word-long text; it exceeds Script K by four types. The TTR (0.40), similarly to that of Script K, is higher than those of script F (.31) and R (.26). It can be explained by the difference in the lengths of the writings, as TTR changes when the number of tokens increases or decreases, making the comparison of data less reliable (Read, 2000).

Table 60: Script O: range of vocabulary based on Range test: Tokens, types, and families in Word lists 1-4, lexical words

| WORD LIST Script O | TOKENS / % | TYPES / % | FAMILIES |
|---|---|---|---|
| One | 109/93.16 | 41/87.23 | 39 |
| Two | 7/5.98 | 5/ 10.64 | 5 |
| Three | 0 | 0 | 0 |
| Four | 0 | 0 | 0 |
| Not in the lists | 1/ 0.85 | 1/ 2.13 | |
| Total | 117 | 47 | 44 |
| Lexical words | 72 | 38 | 36 |

The distribution of types in WLs 1-4 is similar to the higher quality writings, that is, 87 percent of the running words come from WL 1, whereas the rest mostly belong to WL 2 (11%). The ratio of lexical words (.53) is slightly higher than that of the better quality compositions (F: .50, R: .51). It can easily be explained by doing calculations again: fewer types (38) divided by fewer tokens (72) give a similar result to the calculation performed with higher numbers (e.g., Script F: 49/99, TTRL: .50). Thus, it is obvious that by merely counting TTR and TTRL we cannot say much about the scripts, as a very short text applying few lexical words will achieve the same ratios as longer, high-standard texts. Consequently, when analysing students' writings regarding lexis, a minimum level of task achievement concerning length needs to be established and scripts not reaching that length have to be excluded from the calculation of lexical ratios.

The distribution of lexemes in the base lists is somewhat different from the previously discussed scripts in that this writing contains five words from WL 2: *flowers, plants, raining, weather,* and *newspapers*, and does not contain words belonging to less frequent groups in WLs 3 and 4. Nevertheless, it contains *Barbie,* which was not found in the lists. As it is a name, I did not count it in the list of less frequent words, thus, the LS index of script O is inconsiderably lower (.12) than those of the previously analyzed scripts. Prior to the correction of the spelling mistakes, the words *colc'k, folwers, kicthen,* and *tham* were also enumerated in the 'not found' category of Range. LD is extremely high (.62) in Script O due to the fact that the expression *in picture A/B* is not used throughout the text.

The ten most frequently used items are as follows: *the* (18), *is* and *pictures* (6), *see* and *then* (5), *A/a*, *can*, *on, play, to* (4). They make up 51.28 percent of the script. Here again, we can observe that lexical words (*see, then,* and *play*) are applied within the most frequently used items. The following eight frequently occurring types include: c*lock, first, I, picture,*

*second,* and *used* (3-3 times), *and, B* (2), thus 18 types constitute 70 percent of the script, and 29 other words cover the remaining 30 percent.

As there was a small difference between the number of types in Scripts K (43) and O (47) in favour of Script O, as opposed to the points awarded for the texts (K: 4 points; O: 2 points), I have compared the lexical items in them to identify these discrepancies. Surprisingly, the differences were numerous, revealing a lot more about the two writings than the simple word count.

This script contained the following 24 words that were not used in K: *and, any, Barbie, behind, can, can't, come, door, first, flower, I, it, kitchen, plant, man, second, see, she, then, three, to, used, went, woman.* On the other hand, Script K included 20 extra words as compared to the present script: *by, closed, cubes, doll, father, girl, glasses, go, have, mother, near, no, off, open, out, outside, room, teddy, window, with.*

Some of the lexical differences reveal dissimilarities in the perception of the pictures, namely, this script describes the scene from a less social viewpoint; the people are introduced as a *woman* and a *man*, a *boy* and *she;* the personal pronoun is applied instead of the noun even for the first description of the girl. Script K, on the other hand, describes the four people as a family: the adults are seen as a *mother* and *a father*.

Now, let us investigate the appropriacy of the lexical items starting by erroneous applications. The following words are used inappropriately in this text (Script O): *flower* for describing a *plant; any* instead of the conjunct *and; used to* for describing two pictures in parallel, and *went* for something happening now. Two of the extra words in Script O occur in the expression *to come on the door;* the word *behind* is situated in the incomprehensible context *the boy behind can see folwers,* whereas the ordinal numbers *first,* and *second* appear in erroneous combinations *on first pictures / on second pictures.* Five of the words are misspelt, four of which were highlighted in connection with frequency lists, while *barbie* is not spelled by an initial capital. The definite article (*in/on the A picture)* is unnecessary in sentences 4 and 5. Thus, the majority of the 24 extra lexemes are used inappropriately; altogether the following few words were well employed: *I can see/ I can't see* and *three.*

A few words occur in two forms, (1) one correct and one incorrect application (e.g., *flower /folwer* in s9; *can't / can*' in ss 7 and 9), or (2) two erroneous forms (*them / than* instead of *then* in ss 1, 2, 3, and 9) in this writing, which points to the monitor not used efficiently by the student. *Newspaper* is divided into *new papers*. From this analysis it is clear that the number of lexemes displayed in writings does not supply sufficient information about the quality of

the text; as inappropriately used lexis, however rich it may be, can hinder comprehension, that is, the achievement of the communicative goal of the writing.

Language accuracy is limited, with two correct forms against 36 mistakes (0.05). PRC occurs once in a correct grammatical form, but in an erroneous lexical context in the expression *the weather is raining*. As this is the only correct PRC form in the script, I suppose that its second part is an unanalyzed chunk (from *It is raining)*, memorised as a whole (Krashen & Scarcella, 1978). Incorrect solutions are various, for example, in the second part of sentences 2 and 4 the simple verb forms (*play* and *read*) are used, whereas the complex verb form constructed from *be* and the simple verb (*is read* in s4) is applied later. Other verb forms which are not meant to refer to present time also occur: *used to play* (ss 1, 2), *used to come* (s3) and *went* (s3). The alteration of viewpoints can be detected not only in the abundant change of verb tenses in the script, but also in the structures of sentences; at the beginning it is a description of an outsider, whereas later (ss 5, 7) the writer is also involved by saying *I can('t) see.* These examples illustrate the transitional features of interlanguage -- in other words, learners pass through a series of stages before (hopefully) arriving at the target language form.

This writing provides several other examples for the variability of competing interlanguage structures. Regarding EXP, there are two incorrectly used forms (*is two/one plants* in s6). The other possibilities for applying EXP would have been in connection with the plants and the clock (in ss 7, & 9), but were avoided by the inclusion of the observer (*I can see*). As for PLU, it is overused after ordinals (*first/second pictures,* in ss 7, & 9); there is one totally flawless application in sentence 9: *flowers*. There is one correctly applied negative: *I can't see* (s7), but there is also an erroneous form in sentence 9: *can',* which lacks the letter *t,* and, more importantly, the identification of the person who cannot see the flowers. To sum up findings on grammatical accuracy, this script illustrates that its writer has not acquired any of the four examined language forms.

Concerning other mistakes, the most serious one, which hinders comprehension, is in word order and the absence of the subject (I) in the last sentence: *The first pictures the boy behind can see folwers than second pictures the boy behind can' see flowers.* The reader might think that it is the boy who can see *flowers* (plants), and the boy is behind something. The other feature of Script O, which makes it incomprehensible sometimes, is the combination of several errors in a row (*any B is sunny*, in s5). The previous example requires the reader to take at least five steps to disentangle the confusion and comprehend its meaning:

(1) it includes a misspelt conjunction (*any* instead of *and*), so, at first sight *any B* seems to be the subject of this clause. As it has no sense, these words need to be separated and *any* corrected to the conjunction *and* in the reader's mind.

(2) Then the *B is sunny* bunch needs to be clarified by inducing that *B* is the shortened form for *in picture B,* and

(3) it is not *B* that is sunny, but it is related to the *weather,* so

(4) the EXP *it* is missing from the *it is sunny* construct.

(5) Then, while constructing the meaning of the whole sentence, the reader realises that the conjunction *but* would be more appropriate to express the contrast in the weather conditions.

Task achievement is assigned low scores because of the inappropriately used vocabulary and structures that make the writing incomprehensible at places.

Script O contains a few cohesive devices, for example, repetitions are employed in various ways for expressing the differences between the two drawings:

- *In the A picture – in the B picture* (s4)
- *First pictures – second pictures* (ss 6, 7, 9)

Although the adverb of time *then* is used in erroneous forms (*tham,* in ss 1, 2, 3, 7; *than* in s9), it connects the two clauses by referring to something that happened before.

- *S/he used to play – tham* (*then*) (ss 1, 2)
- *Used to come – them went* (s3)

The conjunct *and* is applied in two sentences (ss 4, 6).

Summing up the linguistic features of Script O, the mostly inappropriately used vocabulary and various erroneous forms of grammatical structures prevent comprehension, thus sometimes the writer's attempts to apply cohesive devices fail.

Raters ranked Script O between 7$^{th}$ and 13$^{th}$ (Table 61); Rater 4 ranked it seventh and two assessors put it as 11$^{th}$. Regarding *CEFR* levels, it was assigned 11 A1 levels and nine A2 levels (45%), one of which is A2.2 concerning GLR. These ratings (A2) seem too high compared to the abundance of inappropriately used linguistic elements in Script O.

The other conclusion that may be drawn from the discrepancy between the high percentage of A2 ratings and the ten points received in the first rating is that level A1 of *CEFR* requires lower linguistic competence than the middle band of our rating scale, as it was estimated first. In other words, if 45 percent of ratings place this script in A2 of *CEFR*, and in the first assessment it gained 10 points, which renders it to the under-achiever category not

reaching A1 level, then there occurs a disparity between the two levels (A2 and A1-). It may point to the fact that the medium band of our scale (12-16 points) describes language proficiency higher than the initially estimated A1 level. Thus, an achievement of 10 points, falling between Bands 2 and 3 in our scale may cover level A1.

Table 61: Raters' decision on script O: rank order and *CEFR* levels

| Criteria \ Raters | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|
| Overall rank | 12 | 13 | 11 | 7 | 11 |
| General linguistic range (GLR) | A1 | A1 | A2.1-A2.2 | A2.1 | A1 |
| Overall written production (OWP)/cohesion | A1 | A1 | A1 | A2 | A2 |
| Vocabulary range (VOC) | A2.1 | A1 | A2.1 | A2.1 | A1 |
| Grammatical accuracy (GRA) | A1 | A1 | A2 | A1 | A2 |

Rater 5 placed this script 12th in the first assessment, whereas in the holistic rating process she ranked it 11th. It was ranked in band A2 concerning OWP and accuracy, although earlier it was awarded three points and two points, respectively, for relevant criteria (cohesion and accuracy). In terms of OWP the *CEFR* descriptor for A2 requests a series of simple phrases and sentences linked with simple connectors like *and, but,* and *because*. The script contained sentences connected with *and* and *then,* thus it met the requirement.

As far as accuracy is concerned, A2 descriptors prescribe the use of simple structures correctly, allowing for systematically occurring basic mistakes as long as it is usually clear what the learner is trying to say. A1 level, on the other hand, is characterized as a limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire. Both descriptions fit this writing partly, as A2 allows for basic mistakes, although the part requiring the usage of simple structures correctly may expect more than this writing can offer, thus, perhaps this text would fit A1 level more.

#### 4.8.5.7 Script T

1 It's a boy the play
2 It's a girl the ride book.
3 It's a man in the room.
4 It's a woman play the girl.
5 It's a no't clock.
6 It's a no't plants.
7 It's a TV watch TV.
8 It's a window close the window.
9 It's a weather the windi.
10 It's a wall the two wall.
11 It's a boy the
12 It's a girl the newspaper.
13 It's a man out the room.
14 It's a woman play the bear.
15 It's a clock.
16 It's a plants.
17 It's a TV don't wach TV.
18 It's a window open to the window.
19 It's a weather the suny.
20 It's a wall the one wall.

Script T was ranked 13th during the first analytic assessment with one point assigned to task achievement and one to vocabulary range, whereas accuracy and cohesion were given zero points. According to the rating scale, these scores suggest that in terms of task achievement the text is about more than three things, but those are only partly relevant. Vocabulary embraces a limited scale of words that are often inappropriate. As for accuracy, the text is incomprehensible because of grammatical or spelling mistakes, whereas cohesion is not established, as the structure of the text is limited.

As illustrated in Table 62, the 125-word script contains 32 different words, so its TTR is 0.26, the same as that of Script R, which was assigned five points for vocabulary richness. It is due to the difference in the length of texts, similarly to previous cases. The distribution of types in WLs 1-4 is similar to the higher quality writings, but a slightly higher percentage (91%) of the running words belong to WL 1, and, as opposed to those, here all the rest belong to WL 2 (12.5%). The ratio of lexical words (27/69= .39) is the lowest in this script.

Table 62: Script T: range of vocabulary based on Range test: Tokens, types, and families in Word lists 1-4; lexical words

| WORD LIST Script T | TOKENS / % | TYPES / % | FAMILIES |
|---|---|---|---|
| One | 1 1/4 | 29/90.63 | 29 |
| Two | 1 1/4 | 3/9.38 | 3 |
| Three | 0 | 0 | 0 |
| Four | 0 | 0 | 0 |
| Not in the lists | 0 | 0 | 0 |
| Total | 125 | 32 | 32 |
| Lexical words | 69 | 27 | 27 |

Regarding low-frequency words in Script T, all three of them *(plants, weather, and newspaper*) belong to WL two. The LS index (.14) equals the two best scripts investigated in this section, though we can not trace any signs of lexical sophistication in this writing. Before my correcting the spelling mistakes, the words *suny, wach,* and *windi,* were not found in any categories of the base lists, whereas the misspelt word *read (*spelt as *ride,* s2) was listed as a member of WL 2.

As for frequency features, three types: *a, is,* and *it* (20) constitute nearly half (48%) of the 125 tokens, if the article *the* (14*)* is included, 59 percent of the text is covered by these four words used 74 times. Three other words, *TV, wall,* and *window* (4) add up to 69 percent of the writing performed by using seven types.

As regards the appropriacy of vocabulary, a serious misconception of meanings concerning people can be traced, as three of them are described with words of wrong denotation. In sentences 2 and 12, a girl is claimed to be reading (*ride)* a book or a newspaper, whereas it is the man who is reading. It might be the misperception of the drawing, but, at the same time it shows that the writer does not have the English skills to capture the general schema of a family in which the children play; the father reads a newspaper, whereas the mother arranges something. The woman is referred to as a man in sentences 3 and 13, which can be deduced from the expressions *in/out the room.* The girl is called *woman,* which is understood from the actions linked to her (*play the girl/bear* in ss 4 & 14). As for verbs, altogether three are used, each partly appropriately (*ride* in s2; *play* in ss 1, 4 & 5; *watch* in ss 7 & 17). *Ride* is the misspelt form of *read*, *play* is used each time without the preposition *with*, whereas *watch* is misspelt in sentence 17. Other spelling mistakes include: *windi, suny* and *no't.* The lexeme *picture* is substituted by the word *wall*.

Accuracy was rated zero points; let us examine the reasons for this. The accuracy ratio for correct forms (2) against incorrect ones (29) is 0.06. As for PRC, no correct forms are shown, whereas four simple verb forms of two types are applied (*play* in ss 1, 4, 14; *ride* in s2). Due to the uniformity of the beginning of the sentences (*It's a …)* it is difficult to identify the sentences where the student may have meant to use the EXP, so they may be the following: sentences 5, 6, 10, 15, 16, and 20, in which the other EXP (*it)* is used instead of *there.* PLU is used twice, but in both cases there is a verb agreement error in the expression, the verb *be* is used in singular. On the single identifiable occasion when a numeral would require the PLU form of the noun (*two wall* in s10), it is not applied. One correctly used negative occurs in sentence 17: *don't wach TV,* but the verb is misspelt. In sum, none of the four examined structures is acquired.

Other mistakes reflect an early stage of interlanguage; the fact that all sentences are started by an unanalyzed chunk *it's a* (EXP *it* + verb *be* + indefinite article *a)* suggests that the pupil acquired this structure and relying on overgeneralization they use it to express all types of statements. For example, in sentences 5 (*It's a no't clock.*) and 15 *(It's a clock.)* this structure is applied to express the absence or presence of the clock in place of *there is a clock.* In sentences 9 and 19, on the other hand, the definite article is substituted by EXP *it*: *It's a weather the windi /suny.* Sentence 18 (*It's a window open to the window.)* shows that the student knows the key words (*window, open*), but cannot find the correct structure to describe the situation. The second part of the statement (*to the window)* suggests that the student might remember the expression *go to the window* as an unanalyzed chunk and substituted the verb

*go* with *open*. As subjects and predicates are often unidentifiable in the text, it is hard to comprehend.

This also makes task achievement minimal; due to the inappropriate use of vocabulary and grammatical structures most of the text is incomprehensible. The only sign of the student's effort to make the script coherent is the arrangement of the sentences; statements on persons and objects are placed in one row, beside each other. Nevertheless, they are not connected by using cohesive devices.

As shown in Table 63, this writing was ranked 14th by four assessors and 12th by one. Ratings along linguistic criteria are low; 14 of them assigned this text to A1 level, six placed it even lower, in A1-level and one rater ranked it A.2.1 in relation to linguistic knowledge. It was shown earlier that this rater misunderstood the scale concerning this language feature, and this level was identified the lowest of all on his task sheet ("has a limited repertoire of short memorised phrases covering predictable survival situations"), which suggests that he did not mean A2 level.

Table 63: Raters' decision on script T: rank order and *CEFR* levels

| Criteria \ Raters | R1 | R2 | R3 | R4 | R5 |
|---|---|---|---|---|---|
| Overall rank | 14 | 12 | 14 | 14 | 14 |
| General linguistic range (GLR) | A1- | A1 | A1- | A2.1 | A1 |
| Overall written production (OWP)/cohesion | A1- | A1 | A1 | A1- | A1 |
| Vocabulary range (VOC) | A1 | A1 | A1 | A1 | A1 |
| Grammatical accuracy (GRA) | A1- | A1 | A1 | A1 | A1 |

Rater 5 ranked this writing 14th this time, one rank lower than in the first assessment process. The script was classified in A1 rank along all criteria, the lowest level in the grid. This rater did not use the opportunity to place the script outside the grid (A1-) suggesting that it did not reach even the lowest described level.

To summarize the most important findings, in this section I have analyzed five typical performances of the Baranya corpus in terms of their vocabulary features, grammatical accuracy, text cohesion and task achievement. Corpus-based data on vocabulary and accuracy were compared with criterion-based assessment and text analysis, thus allowing triangulation of the data. Then, five raters' holistic and criterion-based assessments of these writings were related to each other. In addition, one of the raters' assessments in three phases were compared to check intra-rater reliability.

Analysing students' texts from multiple viewpoints with multiple methods provides an excellent opportunity to investigate the great variety of linguistic knowledge and the wide

range of language forms that are present in their interlanguage. Let us return to the RQs examined in this section and sum up the results.

RQ 12 inquired about 8th graders' range of vocabulary and aimed to explore the relationship between criterion-based and corpus-based measurement of accuracy and vocabulary. On a closer examination of vocabulary features, we have seen that better performances contained more words, but the essential disparity between high and low achievers was expressed not exclusively in text length, but rather in the appropriacy of usage.

The investigation of vocabulary richness confirmed Read's (2000) claim that the length of texts greatly influences results. TTRs and TTRLs of poorer scripts (Scripts K & O) were higher than those of excellent and good texts (Scripts F & R), thus, it is reinforced that by simply counting TTR and TTRL scripts can not be qualified. Lexical density is also sensitive to tokens, that is, text length. Script O, which belongs rather to the under-achiever category, has higher LD than the best text, and all others, for that matter. The result of frequency counts showed slight differences between scripts, too, LS ratios ranged between 0.12-0.14.

From the comparison of the analysis of scripts and lexical data gained by counting ratios I can confirm previous claims on the unreliability of comparing lexical ratios of texts of different length (Read, 2000), that is, the limitations of calculating lexical variation. From the investigation of scripts it also became obvious that texts elicited by this picture description task were not specifically suitable for counting lexical features of texts. To put it differently, lexical calculations gained from the data did not reveal much on the quality of the writings, as

1) texts were very short (104-190 words),

2) texts were not the same length,

3) the presence or absence of expressions of comparison (*in picture A*) changed the number of tokens and ratios excessively,

4) automatic counting does not take inappropriate word use and spelling mistakes into consideration.

Nevertheless, lexical calculations and frequency counts revealed useful data on the application of unique, rarely occurring lexemes, and, also, on the overuse of specific language elements.

RQ 11 aimed to examine typical performances and trace developmental levels of students' IL. The five selected scripts represented typical student performances at different levels of language development, and, at the same time, each of them illustrated the endless variability and uniqueness of task performance concerning linguistic criteria.

As regards accuracy, scripts showed several IL forms of the examined language features, demonstrating the transitional features of IL development. Due to the low number of the investigated forms in the five writings, I cannot draw extensive conclusions, but the tendency for correct usage of PLU and NEG can be traced even in the poorest scripts, whereas PRC is at the beginning stage of development in three of the texts. EXPs are correctly used only in Script F, whereas in the other four there are no appropriately applied forms. This reinforces the claim made earlier that EXP is the latest acquired language form out of the four scrutinised ones.

Attempts to create cohesion in their texts were traced in writings F, R, and O. Considering Bachman and Palmer's (1996) general taxonomy of components of language ability, out of the four components (grammatical/ textual/ functional/ sociolinguistic knowledge), these three scripts represented textual knowledge besides knowing grammatical building blocks of language. Applying communicative functions was not relevant when completing this task, whereas sociolinguistic competence as well as background knowledge of the world was shown in Script K in terms of naming members of the family (*mother, father*).

Strategic competence provides the link between one's language knowledge and the external situation, thus, it is not a language-specific ability, but rather a problem-solving construct. Its components (goal setting, assessment, and planning), and the strategies specifically characteristic of writing, that is, the process of reflection, can be spotted in Script F, which is a well-planned text in contrast with the other writings in which even spelling mistakes were not corrected.

Regarding the relationship between raters' opinions (RQ 10), the majority of their decisions were in harmony with each other, the outstanding differences always indicated misinterpretations of the rating scales.

RQ 9 inquired into the relationship between criterion-based and corpus-based measurement of accuracy and vocabulary. This was investigated by checking the consistency of R5's three ratings. Reliability measures indicated strong relationships. Points awarded in the first, criterion-based assessment were in close association with holistic decisions and the second, criterion-based assessment along a new scale (*CEFR,* Council of Europe, 2001). The few inconsistent ratings revealed misunderstanding of the scale descriptors. Table 64 illustrates the points and language levels of the five scripts. The fact that Script O received both A1 and A2 rankings, whereas it was awarded (3, 3, 2, 2) points in the initial assessment, raises the question of the identification of level A1 once again, as was indicated in an earlier discussion of interrelationships between raters' ranks.

Table 64: Rater 5's ranks concerning the five scripts compared with the first assessment

| Script | **F** | | **R** | | **K** | | **O** | | **T** | |
|---|---|---|---|---|---|---|---|---|---|---|
| Criterion | **Rating** | | | | | | | | | |
| | 1st | CEFR | 1st | CEFR | 1st | CEFR | 1st | CEFR | 1st | CEFR |
| Overall rank | 1 | 1 | 5 | 7 | 9 | 10 | 12 | 11 | 13 | 14 |
| General linguistic range (GLR) | - | A2.1 | - | A2.2 | - | A1 | - | A1 | - | A1 |
| Overall written production (OWP)/ cohesion | 8 | B1 | 7 | A2 | 4 | A1 | 3 | A2 | 0 | A1 |
| Vocabulary range (VOC) | 8 | B1 | 5 | A2.1 | 4 | A1 | 2 | A1 | 1 | A1 |
| Grammatical accuracy (GRA) | 8 | B1.1 | 5 | A2 | 4 | A1 | 2 | A2 | 0 | A1 |
| Total points | 32 | | 24 | | 17 | | 10 | | 2 | |

It is possible that scripts with points partly in the second band and partly in the third band, as in the case of Script O, can be ranked level A1. In other words, it indicates that the medium band in our scale covers slightly higher linguistic knowledge than A1. The linking of *CEFR* scales to our own scale is not an easy, one-step exercise, though. I am aware that in order to relate our scale to *CEFR* criteria, the whole process of 'familiarisation – specification- standardisation- empirical validation' (Takala, 2004) should be completed, which is beyond the scope of this study.

# Conclusion

### *Aims and personal motivation*

The study aimed to explore 231 Hungarian (Baranya) students' writing proficiency at the end of their primary education and compare it to the curricular achievement targets. It also scrutinized assessment issues: the relationship of criterion-based and corpus-based measurement, and the connection between analytic and holistic assessment.

The research project was motivated by my interest in average students' language competence as opposed to high-achieving pupils' whose proficiency I had been familiar with, as I had often assessed them at county competitions. The intriguing area of the reliability of writing assessment was my other point of inquiry.

### *Trends in writing assessment*

The first three chapters investigated theoretical and practical considerations connected to L2 writing; approaches to analyzing writing and international research findings on students' texts in L1 and L2 were critically overviewed. Morpheme studies were discussed (Dulay & Burt, 1973; Hakuta, 1974; Larsen-Freeman, 1976), as my study compared some of their findings with interlanguage development of Baranya county children. I also introduced the research methodology of writing studies and visited the key concepts of the assessment of writing in order to be able to use the most efficient methods. Then, being aware that L1 and L2 literacy are interrelated (Cummins, 1981; Cumming, 1989; Krapels, 1990; Kroll, 1990), I surveyed the socio-educational context of Hungarian primary-school learners' writing in L1 and L2 by studying both curriculum requirements and research findings regarding the level of meeting these requirements.

### *Research methodology of the study*

The triangulation of data was ensured by conducting both quantitative and qualitative analyses of scripts, which were weighed against the results of various kinds of measurement. The reliability of assessment was explored by comparing rankings of scripts based on holistic and analytic measures. Raters' opinions were also correlated on a sample of texts, thus reliability was monitored from another perspective. Another point of inquiry concerned the comprehensiveness of *CEFR* scales.

### *Hungarian students' proficiency in English*

**Writing versus other EFL skills**

First, through RQ1, I investigated how students' writing proficiency compared with their other EFL language skills. Several research studies found that students' writing skill is the least developed one among the four skills (Bors, Nikolov, Pércsich, & Szabó, 1999; Bukta & Nikolov, 2002; Nikolov, 2003; Sturman, 2001; Tagányiné, 2001 a, b; Várnai, 2000). This was also proved in the present study, nevertheless, surprisingly, the difference between writing (54%) and the other language skills was not as high as in previous studies; writing performances were only 6 percent behind the balanced results of reading, pragmatics and speaking.

**Baranya results versus national means**

In order to interpret these results in relation to the Hungarian national standard, RQ2 examined how Baranya primary-school learners' L2 proficiency compared with the L2 proficiency of the same age-group in a national sample. Listening and reading skills were on similar level in the two cohorts, although Baranya pupils slightly surpassed the national standard. In the national survey (Csapó & Nikolov, in press) I used for comparison, writing proficiency (33.7%) proved to be the poorest of the three skills measured. Although writing was also the least developed skill (53.93%) of students tested in this survey, it surpassed the national proficiency level by far.

Explanations do not offer themselves easily; the relatively good results might be caused by the slightly larger number of students in this study who learn in good socio-educational circumstances. The other reason for this great difference may lie in the compulsory participation of schools in the national survey, whereas in the Baranya research some pressure had to be used and more motivated teachers participated with their students. These teachers must have been somewhat more self-confident due to their previous results in teaching. Thus, teachers' motivation played a role in joining our project and this fact may partly explain the better results.

**Criterion-based assessment**

RQ3 aimed at identifying the level of Baranya primary-school leavers' writing proficiency. It was measured by a simple picture-comparison task, which was assessed by using an analytic scale of four criteria.

Although the average performance evidenced by the total scores on the writing task was 54 percent that exceeded the national mean, this result represented a wide scale of achievements among the eleven institutions. Four schools, providing 60 percent of participants, outperformed the rest with 64-82 percent achievement on the writing test. The two best schools (N=99) produced excellent results (76-82% in total scores), showing that the average performance of pupils is far beyond the expected language level, that is, A1 of *CEFR* (37.5% on this test). Three of the institutions performed poorly by achieving 29-36 percent of the best possible result, whereas four other schools did not even reach 20 percent of the maximum points, which means that they performed far below the required minimum achievement level. The results regarding the four assessment criteria (task achievement, vocabulary, accuracy, and cohesion) were in line with the total points.

**Curricular achievement targets**

Students' writing proficiency expressed in percentages of successful task performance provides useful information on their achievements, whereas the relationship between these numbers and the curricular achievement targets opens an additional perspective for language educators. To identify this interconnection, RQ4 examined how eighth-graders' writing proficiency meets the *NCC* targets, from two viewpoints: on institutional and individual levels. This time data are summed along the four linguistic criteria (task achievement, vocabulary, accuracy, and cohesion), thus the percentages cover a wider scale of performances than in the previous section accounting of total scores.

**Results at institutional and individual levels**

Viewed by institutions, five schools met *NCC* requirements; two (N=99) performed between 73 and 90 percent on average, two (N=39) between 59 and 76 percent, on the four linguistic criteria. The means of one institution were just about pass-level (N=10) with performance measures between 30 and 41 percent. Altogether these 148 pupils (64%) met the curricular requirements calculated on institutional level.

When investigated individually, the results are similar; as 149 students' (64%) scripts fall in the top three bands (12-32 scores). If the required minimum level (A1) is adjusted according to the findings of statistic and qualitative analyses, and, consequently, a few slightly weaker scripts (N=15) are accepted as reaching A1 level, it can be claimed that 71

percent of the participants performed at the nationally prescribed language level. The rest, altogether 63 students (29%) did not meet the *NCC* requirements.

**Results by students' SES**

Studies carried out in Hungary inform us that children's access to FL learning is greatly influenced by their SES (Vágó, 2007), and so is the FL proficiency level they attain (Bors, Lugossy, & Nikolov, 2001; Bukta & Nikolov, 2002; Csapó, 1998; 2002; Csapó & Nikolov, in press). RQ5 aimed to investigate if it was true for this cohort; therefore I studied the relationship between students' socio-educational background and their writing proficiency.

My findings also reveal significant differences between schools according to their SES and location. Large town institutions can be divided in two groups based on students' SES. Schools where students' SES is high performed extremely well, whereas two city schools with low SES pupils demonstrated poor language achievements. Students living in small towns displayed large variance in their performance. Village schools performed poorly showing a wide range of linguistic abilities even within this low achievement category.

**Corpus analysis regarding linguistic accuracy**

Besides the results gained from criterion-based assessment, I was curious about exact language performance data attainable from a corpus analysis. RQ6 inquired into how accurately 8th-graders use simple target language forms (PRC, EXP, NEG, & PLU), and what developmental patterns emerge in their compositions, thus implementing a purely linguistic perspective on texts.

The number of students attempting to use the four examined language forms and the ratio of success show large differences. First, let us see students' readiness for using these constructions. Present continuous tense (94%) and PLU forms (87%) were used by the great majority of participants, whereas EXPs occurred in 76 percent, and NEG in 48 percent of writings. Summing up all occurrences of the selected constructions: PRCs supply 52 percent of the four forms; EXP constructions contribute 24 percent, whereas PLUs add 16 percent, and NEG almost 8 percent to the total of the analyzed language forms.

As regards accuracy of usage, NEG is the most successfully employed construction (79%) in the whole corpus; PLUs come second with 65 percent, while EXPs reach only a 54 percent correctness rate and PRC is the least accurately managed area, although not far behind EXPs, with 48 percent successful attempts. Regarding developmental patterns, these data suggest that NEG, followed by PLU, is the first acquired language construction on this task in the

investigated schools. EXPs seem to be acquired third of the four examined forms, and PRC appears to be the last.

EXPs and PRC, though, show contradicting patterns of correctness ratios in the eleven examined institutions. In the four best achieving schools and one school performing in the middle band EXPs are produced more accurately than PRC forms. In generally underachieving schools, on the other hand, PRC shows higher correctness ratios than EXPs. The relatively correct usage of PRC by low-achieving students may represent the first stage of its development, where the compound construction is used as a memorised chunk of language. At a later stage of morpheme development more verb forms are salient for learners who use these forms interchangeably until the correct form is acquired at last (Bailey, Madden, & Krashen, 1974). The late development of EXP may be due to its lack of salience in the early stages of EFL learning. This study, however, has not examined the role of input, which ensures the frequency of morpheme occurrence claimed to be one of the determinant factors of morpheme ordering (Goldschneider & DeKeyser, 2001).

**Criterion- and corpus-based assessment**

Having scrutinised exact data on Baranya school-leavers' usage of the selected language forms, I will address assessment issues, and summarise findings on the relationship between criterion-based and corpus-based measurement of accuracy. RQ7 examined the relationship between scores assigned on accuracy and the ratio of correct application of grammatical forms. It was computed by Pearson product-moment correlation count, which revealed a strong (.797), significant ($p<0.01$), positive relationship between the number of correct language forms and accuracy scores, whereas the scores had a moderate, negative (-.318) relationship with the number of errors.

So as to explain the effects of the two independent variables (correct and erroneous forms) on the accuracy score as a dependent variable, a multiple regression model was built. According to this, 64 percent of the variance in accuracy scores is explained by the number of correct and erroneous language forms. The same regression model was built to explain the joint effect of the four investigated constructions (PRC, EXPs, PLU, & NEG) on the accuracy score. The model showed that correctness values of the four language forms explain a higher ratio (75%) of accuracy scores, than correct and incorrect forms in general. This contradiction, revealed by the regression model and, then, by the close analysis of students' scripts, led me to the realisation of my mistake/false hypothesis in error tagging. Fortunately,

it did not concern the main research questions investigating the selected language forms, moreover, it confirmed the results of the appropriate coding of the main language forms.

The ratio of contribution of each selected correctly used language form to the accuracy score was also computed. According to this, EXPs are the best predictors of accuracy scores (32%), whereas PRC explains 20 percent of the variance. The correctness level of NEG predicts merely 3 percent of the accuracy assessment, while the ratio of correct PLU forms does not significantly contribute to the accuracy score.

**Holistic and analytic assessment**

In order to investigate the reliability of assessment, five raters were involved in its second phase, where holistic and analytic decisions were made on students' writing proficiency. RQs 8-9 examined how raters' opinions of sample tests were related to one another, and, also, how the results on these types of assessment (holistic vs analytic) compared.

Two reliability measures, and then, Spearman rank-order correlations were computed to investigate these interrelations. According to Cronbach's alpha, interrater reliability was high (Cronbach's alpha >0.83) in all areas, whereas Krippendorff's alpha measured high agreement in raters' viewpoints concerning overall ranking (.87), and, lower agreement as regards VOC (.59). Spearman rank-order correlation coefficients ranged between 0.803 and 0.928 concerning the rank order of participants' written performance. This high agreement in overall ranking of scripts during holistic measurement indicates that assessors relied on common underlying criteria while rating scripts. These results reinforce Barkaoui's (2007) findings concerning the higher reliability of holistic assessment, but, as the criterion-based measures were based on *CEFR* criteria that were not standardised and validated with raters in this study, these results have to be interpreted with caution. Validated criteria would probably have supported higher agreement in opinions.

**Interrelations of raters' judgments**

Examining the ten interrelations of raters' judgments in each of the four assessment areas: VOC proved to elicit the highest agreement values (.571-.914) in ten significant connections, in GRA seven significant correlations values varied between 0.532 and 0.807, whereas GLR assessment resulted in six significant connections (.624-.965), and OWP reached merely five significant correlations ranging between 0.579-0.836. I found that the values not reaching 95 percent confidence were caused by differences in assessors' intepretation of *CEFR* scales.

To answer RQ10, the relationship of raters' interpretations of these scales was investigated by three methods; first, by examining reliability ratios of their ranking demonstrated in previous paragraphs, then, by analysing assessors' feedback on the rating process followed by a qualitative analysis of five students' scripts. By this close analysis, raters' judgments were compared with the quality of texts along four linguistic criteria.

According to raters' feedback and the few outlying rankings, the inconsistency in raters' opinions proved to be due to misinterpretations of some descriptors of the *CEFR* scales, which confirms the statement that the comprehensibility, transparency and coherence of the *CEFR* are not sufficient if the scales are used without standardisation. Due to the lack of precision and explicitness of the scales, it seems advisable to rely on *The Manual* (Takala, 2004), and carry out the four steps that contribute to the validation of the scales.

**Findings of the qualitative analyses**

I have revealed important findings on general features of Baranya learners' L2 competence, but I also aimed to investigate the achievements on the four language criteria of our scale (task achievement, vocabulary, accuracy, and text cohesion) on a sample of writings qualitatively. RQ11 explored what typical performances were like, and what developmental levels these scripts reflected. Thus, out of the fifteen reassessed scripts five were closely analyzed. This procedure served the triangulation of data, as quantitative measures were compared with the quality of writings.

The criterion of the selection of these texts was the overall score attained in the analytic assessment, so as to represent a wide range of linguistic levels. Thus, they represent high-achievers (32 points), above average performance (24 points); the level slightly above the pass-level (17 points), the level merely under the required one (10 points) and under-achievers (2 points). These scripts illustrate a great variety of language knowledge and the wide range of IL forms that are characteristic of the whole corpus.

The best script displays much higher language competence than required in the final grade of primary education. Accordingly, the majority of 21 ratings identified it as B.1, whereas seven considered it A2 level. This also exemplifies that a task estimated to measure language proficiency at A1-A2 level can elicit higher level learner performances (*EALTA* discussion, 2005), and this is an important contribution of this study to discussions on this theme.

The text representing above-average level distributed raters' opinions the most, as they assigned it five different ranks with eight ratings in B1, 11 in A2, and one in A1. The script classified as just beyond pass-level in the first phase was ranked 9 A1, 10 A2 and two B1

levels in the second phase of assessment, thus 57 percent of the ratings ranked it higher than A1. The below-average script was assigned 11 A1 levels and nine A2 levels (45%), which ranking seems too high compared to the number of inappropriately used linguistic elements in it. The high ratings based on *CEFR* criteria of the two latter scripts suggest a discrepancy between our initially estimated A1 level and the one described by *CEFR* criteria. There is still a lot to do for the standardised interpretation of *CEFR* levels, as discrepancies do occur in major Hungarian educational documents (NCC, 2003; FC, 2000).

Regarding developmental language patterns, the five samples showed similar acquisitional routes to those explored in the entire corpus. Several non-target-language-like forms of the examined four constructions emerged, well demonstrating the transitional features of IL development. The three poorest scripts demonstrated correct usage of PLU and NEG forms, whereas PRC represented an early stage of development in them. EXP is totally acquired only by the high-achieving student. This supports the claim that EXP is the latest acquired language form of the examined constructions in this study.

**Vocabulary features of texts**

RQ12 represents the complexity of this investigation by posing questions related to both language use and assessment: it aims to examine the range of vocabulary of 8$^{th}$ graders, and the relationship between criterion-based and corpus-based measurement of accuracy and vocabulary.

Vocabulary range and frequency characterising students' texts was investigated with the help of the Range test (Nation, 2005). Lexical richness was explored by two types of TTR; lexical density and lexical sophistication were also calculated. These calculations and frequency counts supplied useful data on individual students' specific vocabulary features, and, also, on the overuse of some language elements.

The close analysis of vocabulary richness confirmed Read's (2000) claim that the differences in the length of texts have a huge influence on results. During the analysis of scripts it also became transparent that texts elicited by the picture description task were not specifically suitable for quantifying lexical features of texts, as lexical ratios did not provide information on the quality of the writings. Surprisingly, TTRs and TTRLs of poorer scripts were higher than those of excellent and good texts, which confirm that scripts can not be qualified by merely counting TTR and TTRL. Lexical density is also sensitive to text length. Further research is needed on the influence of task types on the lexical ratios of scripts.

Comparing vocabulary features of texts, I have found that higher quality scripts were usually longer, but the essential difference between high and low achievers was not related to text length, but the appropriacy of use. Consequently, if lexical features of texts are to be scrutinised reliably, a minimum level of task achievement concerning length needs to be established and shorter scripts have to be excluded from the calculation of lexical ratios.

Originally, I intended to study vocabulary features characterising the whole corpus, but, as calculating lexical ratios in the analysis of five scripts did not provide sufficient information on the quality of students' lexis, I dismissed this aim. Torres, Naves, Celaya and Pérez-Dival (2006) also found that fluency measures, like the number of sentences is misleading evidence of the development of writing proficiency, as differences do lie in the internal complexity of sentences (coordination and subordination). Frequency measures seemed more promising in qualifying texts, but still not sufficient without close analyses of scripts. Thus, as the scope of the dissertation did not permit the examination of a high number of texts, I had to abandon this research area.

**Relationship between criterion- and corpus-based measurement**

RQ 12 also inquired into the relationship between criterion-based and corpus-based measurement of accuracy and vocabulary, which was accomplished by monitoring the consistency of R5's three ratings and comparing these with quantitative data on learners' language use. Reliability measures indicated strong relationships. Points in the first criterion-based assessment were in close association with holistic ranks and, also with the second, criterion-based measurement along *CEFR* scales (Council of Europe, 2001), which shows that R5's main underlying criteria are consistent and independent of the type of assessment. The few inconsistent ratings and R5's retrospection revealed the misunderstanding of some scale descriptors.

In sum, I can claim that Baranya eighth-graders' written L2 proficiency ranges along a wide scale, and it is slightly less developed than their other language skills measured in this study. Nevertheless, students in Baranya county displayed higher L2 proficiency than a national sample of their peers in Hungary.

**Limitations**

There are several limitations to this research. Although it has analyzed Hungarian primary-school-leavers' EFL writing competence using multiple methods, the analyses were based on one type of script: picture description. If the investigation was carried out on a different task type, results may have been also discrepant.

This research study focused solely on the product of the complex cognitive process of writing. The process itself was not investigated, neither were classroom interactions that had formed students' motivation and EFL writing skills. From among the abundance of background criteria influencing L2 acquisition merely SES of children was addressed, on an institutional basis. Thus, conclusions cannot be drawn in relation to the background factors of EFL writing competence. However, stages of this route and the order of acquisition of four language forms have been identified. These data did not yield information whether the absence of a particular language form in a script is due to lack of linguistic ability, or lack of exposure to the construction.

**Pedagogical implications and future research**

Young learners' FL writing is a scarcely researched area. This study is innovative in several ways. First, it created a corpus of 231 primary-school-leavers EFL texts, which is the biggest EFL corpus of this age group in Hungary. This is also the first corpus of young learners that has been error-tagged and analyzed for specific linguistic features. Additionally, this is a pioneer research study in Hungary to scrutinize interlanguage development of primary-school learners through writing. A few of the writings were also investigated qualitatively, no such close analysis has been attempted for this age group and this script type in Hungarian EFL writing assessment. This is the first attempt to compare students' L2 performances with the curricular achievement targets, and also with levels of the *CEFR*.

The findings of this research can be exploited in at least two areas. In everyday teaching practice they can raise EFL teachers' awareness of the acquisitional sequence of morphemes, thus, the order of introduction of these forms can be adjusted to it. In case of EXPs, for example, it could be considered to teach them later than in today's practice reflected in most course books, that is, the first year of L2 instruction.

The findings concerning assessment reinforce claims (Creswell, 2003) that quantitative and qualitative measurement complement each other, the first provides data on general tendencies, whereas the latter provides insight into fine details. Qualitative analysis of scripts showed that completely different levels of language proficiency were covered by similar

ratios of types and tokens and very different texts scored equal ratios, as text length influenced these ratios. The implication of this finding is that a minimal length of task achievement needs to be set, and shorter texts have to be excluded from the accounts.

The results gained from investigating raters' interpretation of *CEFR* scales suggest that in this area there is a lot to do. On a national level, it would be advisable to reconsider the matching of EFL requirements of core educational documents, namely, the *NCC* (2003) and *FC* (2000). Secondly, teachers should be offered refresher training courses in order to get acquainted with *CEFR,* and through that *NCC* requirement levels for EFL, and use this in their teaching.

The corpus collected and analyzed in this study can be further exploited. Texts can be tagged for other language forms influencing accuracy; vocabulary features can be investigated by frequency analysis. Students' strategies for creating texts can also be examined. The process of writing can be tapped into by repeating the test on a small sample while using think-aloud protocol to gain insight into the composing process. A longitudinal study of young learners' writing proficiency would supply more reliable data on the developmental phases of language acquisition. This study is hoped to have served a starting point for these inquiries.

# References


Alderson, J. C. (1991). Bands and scores. In J. C. Alderson, & B. North (Eds.), *Language testing in the 1990s: The communicative legacy* (pp. 71-86). London: Modern English Publications/British Council/Macmillan.

Alderson, J. C. (1996). Do corpora have a role in language assessment? In J. Thomas & M. Short (Eds.), *A new decade of language testing research* (pp. 65-78). Harlow: Longman.

Alderson, J. C. (2000). *Assessing reading.* Cambridge: Cambridge University Press.

Alderson, J. C. (Ed.). (2002). *Common European Framework of Reference for Languages: Learning, teaching, assessment: case studies*. Strasbourg: Council of Europe.

Alderson, J. C., Figueras, N., Kuijper, H., Nold, G., Takala, S., & Tardieu, C. (2004). The development of specifications for item development and classification within the Common European Framework of Reference for Languages: Learning, teaching, assessment. Reading and listening. Final report of the Dutch CEF construct project. Unpublished document.

Andor, M. (2000). A nyelvtudás szociális háttere. *Educatio, 9* (4), 717-728.

August, D., & Hakuta, K. (Eds.). (1997). *Improving schooling for language-minority children.* Washington, DC: National Academy Press.

Bachman, L. (1990). *Fundamental considerations in language testing.* Oxford & New York: Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language testing in practice.* Oxford: Oxford University Press.

Bailey, N., Madden, C., & Krashen, S. (1974). Is there a natural sequence in adult second language learning? *Language Learning, 21,* 235-243.

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing Writing, 12,* 86-107.

Bialystok, E. (1981). The role of linguistic knowledge in second language use. *Studies in Second Language Acquisition, 4,* 31-45.

Blades, M. (2007). Jim Cummins demolishes NCLB's ideology and practice. *Annual conference of the organization of California Teachers of Other Languages,* San Diego, U.S.A. 26 July, 2007. [Online]. Available http://www.dailykos.com/storyonly/2007/7/26/131722/394

Bors, L. (1999). An investigation of Hungarian primary school teachers' and pupils' views on the Baranya Reading Project and suggestions for improving the design of the project. Unpublished dissertation for the degree of MEd. in Primary Teacher Training (TESOL) at the University of Leeds.

Bors, L., Nikolov, M., Pércsich, R., & Szabó, G. (1999). A pécsi nyolcadik osztályosok idegen nyelvi tudásának értékelése. *Magyar Pedagógia,* 99(3) 289-306.

Bors, L., Lugossy, R., & Nikolov, M. (2001). Az angol nyelv oktatásának átfogó értékelése pécsi általános iskolákban. *Iskolakultúra, 4,* 73-88.

Bölcskei, M. (1996). Evaluating an achievement test administered to primary school leavers in Baranya county. In: M. Nikolov & Horváth J. (Eds.), *Learning lessons: Innovations in teacher education and assessment* (pp. 168-191). Pécs, Hungary: Lingua Franca Group.

Breen, M. P., & Candlin, C. N. (1980). The essentials of a communicative curriculum in language teaching. *Applied Linguistics, 1*(1), 89-112.

Brooks, N. (1960). *Language and language learning: theory and practice.* New York: Harcourt Brace and World.

Brown, R. (1973). *A first language: The early stages.* Cambridge, MA: Harvard University Press.

Brown, H. D. (1994). *Teaching by principles: An interactive approach to language pedagogy.* Englewood Cliffs, NJ: Prentice Hall Regents.

Brown, J. D., & Rodgers, T. S. (2002). *Doing second language research.* Oxford: Oxford University Press.

Bukta, K. (2001). Mit tanulnak a diákok angol órán? *Iskolakultúra, 11*(4), 36-47.

Bukta, K. (2007). Processes and outcomes in L2 English written performance assessment: Raters' decision-making processes and awarded scores in rating Hungarian EFL learners' compositions. Doktori értekezés. PTE, BTK, Nyelvtudományi Doktori Iskola.

Bukta, K., & Nikolov, M. (2002). Nyelvtanítás és hasznos nyelvtudás: az angol mint idegen nyelv. In B. Csapó (Ed.), *Az iskolai műveltség* (pp. 169-192). Budapest, Hungary: Osiris.

Bialystok, E. (1982). On the relationship between knowing and using forms. *Applied Linguistics, 3*, 181-206.

Bialystok, E. (2001). *Bilingualism in development.* Cambridge: Cambridge University Press.

Bialystok, E., & Hakuta, K. (1999). Confounded age: Linguistic and cognitive factors in age differences for second language acquisition. In D. Birdsong (Ed.), *Second language acquisition and the critical period hypothesis* (pp. 61-81). Mahwah, NJ: Lawrence Erlbaum.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language testing. *Applied Linguistics, 1,* 1-47.

Corder, S. P. (1967). The significance of learners' errors. *International Review of Applied Linguistics, 5,* 161-170.

Corder, S. P. (1971). Idiosyncratic dialects and error analysis. *International Review of Applied Linguistics,* 9, 149-159.

Council of Europe (2001). *Common European framework of reference for languages: learning, teaching, assessment*. Cambridge: Cambridge University Press.

Coxhead, A. (1998). The development and evaluation of an academic word list. MA Thesis, Victoria University of Wellington, New Zealand.

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

Chomsky, N. (1957). *Syntactic structures.* The Hague: Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Creswell, J. W. (2003). *Research design: Qualitative, quantitative, and mixed methods approaches* (2nd ed.). Thousand Oaks, CA: Sage.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning, 39,* 81-141.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7,* 31-51.

Cumming, A., Kantor, R., & Powers, D. E. (2002). Decision-making while rating ESL/EFL writing tasks: A descriptive framework. *Modern Language Journal*, *86* (1), 67–96.

Cummins, J. (1976). The influence of bilingualism on cognitive growth: A synthesis of research findings and explanatory hypotheses. *Working Papers on Bilingualism, 19,* 197-205.

Cummins, J. (1981). The role of primary language development in promoting educational success for language minority students. In *Schooling and Language Minority Students: A Theoretical Framework* (pp. 3-49). Los Angeles: Evaluation, Dissemination and Assessment Center, California State University.

Cummins, J. (2001). The influence of bilingualism on cognitive growth: a synthesis of research findings and explanatory hypotheses. In C. Baker & N. H. Hornberger (Eds.), *An introductory reader to the writings of Jim Cummins* (pp. 1-43). Clevedon, Avon: Multilingual Matters.

Csapó, B. (Ed.). (1998). *Az iskolai tudás.* Budapest: Osiris Kiadó.

Csapó, B. (2001). A nyelvtanulást és nyelvtudást befolyásoló tényezők. *Iskolakultúra, 11*(8), 25 -35.

Csapó, B. (Ed.). (2002). *Az iskolai műveltség.* Budapest: Osiris Kiadó.

Csapó, B., & Nikolov, M. (2002). The relationship between students' foreign language achievement and general thinking skills. *American Educational Research Association Annual Meeting,* New Orleans, U. S. A. 1-5 April, 2002.

Csapó, B., & Nikolov, M. (in press). The cognitive contribution to the development of proficiency in a foreign language. *Learning and Individual Differences.*

Csizér, K. (2007). A nyelvtanulási motiváció vizsgálata: Angolul és németül tanuló diákok motivációs beállítódása a nyelvválasztás tükrében. *Új Pedagógiai Szemle, 6,* 54-68.

Csíkos, Cs., & Steklács, J. (2006). Metakogníció és olvasás. In K. Józsa (Ed.), *Az olvasási képesség fejlődése és fejlesztése.* (pp. 75-88). Budapest: Dinasztia Tankönyvkiadó.

Dagneaux, E., Dennes, S. & Granger, S. (1998). Computer-aided error analysis. *System, 26*(2), 163-174.

Demuth, K. (1998). Collecting spontaneous production data. In D. McDaniel, C. McKee, & H. S. Cairns (Eds.), *Methods for assessing children's syntax* (pp. 3-22). Cambridge, MA: The MIT Press.

Doiz, A., & Lasagabaster, D. (2004). The effect of the early teaching of English on writing proficiency. *International Journal of Bilingualism, 4,* 525-540.

Douglas, D. (2000). *Assessing languages for specific purposes.* Cambridge: Cambridge University Press.

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

Dulay, H. & Burt, M. (1973). Should we teach children syntax? *Language Learning, 23,* 245-258.

Dulay, H., & Burt, M. (1974). Errors and strategies in child second language acquisition. *TESOL Quarterly, 8,* 129-136.

Dulay, H., Burt, M., & Krashen, S. (1982). *Language two.* New York: Oxford University Press.

Elley, W. (1998). *Raising literacy levels in third world countries: A method that works.* Culver City, California: Language Education Associates.

Ellis, R. (1994). *The study of second language acquisition.* Oxford: Oxford University Press.

Ellis, R. (2004). The definition and measurement of explicit knowledge. *Language Learning, 54*(2), 227-275.

Ellis, R. (1997). *Second language acquisition.* Oxford: Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language.* Oxford: Oxford University Press.

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

Eurobarometer 54 Special. February 2001. International Research Associates. [Online]. Available http://europa.eu.int/comm/dg10/epo

Eurobarometer 2005. in *Jelentés a magyar közoktatásról 2006.* http://oki.hu/oldal

Special Eurobarometer survey 64.3. December 2006. 'Europeans and Languages' [Online]. Available http://europa.eu.int

Favreau, M., & Segalowitz, N. S. (1982). Second language reading in fluent bilinguals. *Applied Psycholinguistics, 3,* 329-41.

Firbas, J. (1986). On the dynamics of written communication in the light of the theory of functional sentence perspective. In C. Cooper, & S. Greenbaum (Eds.), *Studying writing: linguistic approaches* (pp. 40-71). London: Sage.

Flower, L., & Hayes, J. (1981). A cognitive process theory of writing. *College Composition and Communication*, *40,* 282-311.

Freeman, D. (1996). 'E-mail interview with Andy Barfield' *Teacher Talking to Teacher, 4*(2), 3-5.

Gardner, R., & Lambert, W. (1972). *Attitudes and motivation in second-language learning.* Rowley, MA: Newbury House.

Gathercole, S. E., & Baddeley, A. (1993). *Working memory and language*. Hillsdale, NJ: Lawrence Erlbaum.

Gebhard, J. G., Fodor, M., & Lehmann, M. (2003). Teacher development through exploration: Principles, processes, and issues in Hungary. In J. Andor, J. Horváth, & M. Nikolov (Eds.), *Studies in English Theoretical and Applied Linguistics* (pp. 250-261). Pécs, Hungary: Lingua Franca Csoport.

Gee, J. P. (1989). What is literacy? *Journal of Education, 171,* 18-25.

Gee, J. P. (1996). *Social linguistics and literacies: Ideology in discourses.* London: Falmer Press.

Goldschneider, J., & DeKeyser, R. (2001). Explaining the 'natural order of L2 morpheme acquisition' in English: A meta-analysis of multiple determinants. *Language Learning, 51,* 1-50.

Goodman, K. S. (1970). Psycholinguistic universals in the reading process. *Journal of Typographic Research, 4,* 103-10.

Goodman, K. S. (1985). Unity in reading. In H. Singer, & R. B. Ruddel (Eds.), *Theoretical models and processes of reading.* (pp. 86-98). Newark, DE: International Reading Association.

Grabe, W. (2002). Reading in a second language. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 49-59). New York: Oxford University Press.

Grabe, W., & Kaplan, R. B. (1996). *Theory and practice of writing.* New York: Longman.

Granger, S. (Ed.). (1998). *Learner English on computer*. London: Longman.

Gregg, K. (1990). The variable competence model of second language acquisition and why it isn't. *Applied Linguistics, 11,* 364-83.

Grice, H. P. (1975). Logic and conversation. In P. Cole, & J. Morgan (Eds.), *Syntax and semantics* (pp. 41-58). New York: Academic Press.

Grotjahn, R. (1987). On the methodological basis of introspective methods. In C. Færch, & G. Kasper (Eds.), *Introspection in second language research* (pp. 58-66). Clevedon, U.K.: Multilingual Matters.

Gui, S., & Yang, H. (2001). Computer analysis of Chinese learner English. *Keynote lecture given at the Conference of Technology in Language Education,* Hong Kong, June 2001.

Hakuta, K. (1974). A preliminary report on the development of grammatical morphemes in a Japanese girl learning English as a second language. *Working Papers on Bilingualism,* 3, 18-43.

Halász, G., & Lannert, J. (1998). *Jelentés a magyar közoktatásról 1997.* Budapest: Országos Közoktatási Intézet.

Halász, G., & Lannert, J. (2000). *Jelentés a magyar közoktatásról 2000.* Budapest: Országos Közoktatási Intézet.

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

Halliday, M. A. K. (1994). *An introduction to functional grammar.* 2nd ed. London: Edward Arnold.

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

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll (Ed.), *Second language writing: Research insights for the classroom* (pp. 69-87). New York: Cambridge University Press.

Hamp-Lyons, L. (1991). Pre-text: Task-related influences on the writer. In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts.* Norwood, NJ: Ablex.

Hamp-Lyons, L. (2007) Worrying about rating. *Assessing Writing, 12,* 1-9.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-trait scoring instrument across ESL writing assessment contexts. *Language Learning, 41,* 337-373.

Harklau, L. (2002). The role of writing in second language acquisition. *Journal of Second Language Writing, 11,* 329-350.

Hasselgreen, A. (2005). Assessing the language of young learners. *Language Testing, 3,* 337-354.

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 9,* 122–159.

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures, 1*(1), 77-89.

Hayes, J. R. (1996). A new framework for understanding cognition and affect in writing. In C. M. Levy & S. Ransdell, (Eds.), *The science of writing* (pp. 1-27). NJ: Lawrence Erlbaum Associates.

Hayes, J. R., & Flower, L. S. (1980.) Identifying the organization of writing processes. In L. W. Gregg & E. R. Steinberg (Eds.), *Cognitive processes in writing* (pp. 31-50). Hillsdale, NJ: Lawrence Erlbaum Associates.

Heatley, A., Nation, I. S. P. & Coxhead, A. (2002). Range and frequency programs. [Online]. Available at http://www.vuw.ac.nz/lals/staff/Paul_Nation

Hedgcock, J. (2005). Taking stock of research and pedagogy in L2 writing. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 597-613). Mahvah, NJ: Lawrence Erlbaum Associates.

Hedgcock, J. S., & Lefkowitz, N. (1992). Collaborative oral/aural revision in foreign language writing instruction. *Journal of Second Language Writing,* 3, 255-276.

Hegedűs, I. (2001). Anyanyelv. In J. Noijons, & Zs. Várnai (Eds.), *Tanulói teljesítmény mérése az általános iskolában. A közoktatás-fejlődés mérése Magyarországon című projekt zárótanulmánya,* (pp. 19-38). Budapest, Hungary: Fővárosi Pedagógiai Intézet.

Hinkel, E. (2005). Analyses of second language text. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 615-628). Mahvah, NJ: Lawrence Erlbaum Associates.

Hoey, M. (1983). *On the surface of discourse.* London: Allen & Unwin.

Hoey, M. (1996). *Patterns of lexis in text.* Oxford: Oxford University Press.

Horváth, J. (2001). *Advanced writing in English as a foreign language: A corpus-based study of processes and products.* Pécs, Hungary: Lingua Franca Csoport.

Horváth, Zs. (1998). *Anyanyelvi tudástérkép: Mérés-értékelés-vizsga 4. Középiskolai tantárgyi feladatbankok.* Budapest: Országos Közoktatási Intézet.

Hudelson, S. (1988). *Write on: Children's writing in ESL.* Englewood Cliffs, NJ: Prentice Hall.

Hughes, G. (1997). Developing a computing infrastructure for corpus-based teaching. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 292-307). Applied linguistics and language study. London: Longman.

Huhta, A., Luoma, S., Oscarson, M., Sajavaara, K., Takala, S., & Teasdale, A. (2002). A diagnostic language assessment system for adult learners. In J. C. Alderson (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment: case studies* (pp. 130–146). Strasbourg: Council of Europe.

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

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual and Multicultural Development, 9,* 67-84.

Hymes, D. (1972). On communicative competence. In J. Pride, & A. Holmes (Eds.), *Sociolinguistics* (pp. 269-293). NY: Penguin.

Ionin, T., & Wexler, K. (2002). Why is 'is' easier than '-s'?: acquisition of tense/agreement morphology by child second language learners of English. *Second Language Research, 18*(2), 95-136.

Jarvis, S., Grant, L., Bikowski, D. & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing, 12,* 377–403.

Jones, N. (2002). Relating the ALTE framework to the Common European Framework of Reference. In J. C. Alderson, (Ed.), *Common European Framework of Reference for Languages: learning, teaching, assessment: case studies* (pp. 167–183). Strasbourg: Council of Europe.

Józsa, K. (2003). *Az idegen nyelvi készségek fejlettsége angol és német nyelvből a 6. és 10. évfolyamon a 2002/2003-as tanévben*. Függelék. Budapest: OKÉV.

Kádárné, F. J. (1990). *Hogyan írnak a tizenévesek? Az IEA fogalmazásvizsgálat Magyarországon.* Budapest: Akadémiai Kiadó.

Kaftandjieva, F., & Takala, S. (2002). Council of Europe scales of language proficiency: a validation study. In J. C. Alderson, (Ed.), *Common European Framework of Reference for Languages: Learning, teaching, assessment: case studies* (pp. 167–183). Strasbourg: Council of Europe.

Kaplan, R. B. (1988). Contrastive rhetoric and second language learning: Notes toward a theory of contrastive rhetoric. In A. Purves (Ed.), *Writing across languages and cultures* (pp. 275-304). London and Newbury Park, CA: Sage.

Kaplan, R. B., & Palmer, J. D. (1992). Literacy and applied linguistics. In W. Grabe, & R. B. Kaplan (Eds.), *Introduction to applied linguistics* (pp.191-212). New York: Addison-Wesley.

Kern, R. (2000). *Literacy and language teaching.* Oxford: Oxford University Press.

Kennedy, C. & Kennedy, J. (1996). 'Teacher attitudes and change implementation' *System, 24*(3), 351-60.

*Kerettanterv* (2000). Budapest: Oktatási Minisztérium.

Kiszely, Z. (2003). Students' writings in L1 Hungarian and L2 English: rhetorical patterns, writing processes and literacy backgrounds. Doktori értekezés. PTE, BTK, Nyelvtudományi Doktori Iskola.

Koda, K. (1993). Transferred L1 strategies and L2 syntactic structure in L2 sentence comprehension. *Modern Language Journal, 78* (4), 490-500.

Kobayashi, H. & Rinnert, C. (2002). High school student perceptions of first language literacy instruction: implications for second language writing. *Journal of Second Language writing*, *11,* 91-116.

*Közös európai referenciakeret: Nyelvtanulás, nyelvtanítás, értékelés* (2002). Pilisborosjenő: Pedagógus-továbbképzési Módszertani és Információs Központ Kht.

Kramsch, C. (1997). Rhetorical models of understanding. In T. Miller (Ed.), *Functional approaches to written text: Classroom applications* (pp. 50-63). Washington, DC: USIA.

Krapels, A. (1990). An overview of second language writing process research. In B. Kroll (Ed.), *Second language writing: research insights for the classroom* (pp. 37-56). New York: Cambridge University Press.

Krashen, D. S. (1981). *Second language acquisition and second language learning.* Oxford: Pergamon Press.

Krashen, D. S. (1984). *Writing: Research, theory and applications.* CA, USA: Laredo Publishing Co.

Krashen, D. S. (1993). *The power of reading.* Englewood, CO: Libraries Unlimited.

Krashen, D. S. (2003). Evidence for the Comprehension Hypothesis from method comparison studies. In J. Andor, J. Horváth, & M. Nikolov (Eds.), *Studies in English Theoretical and Applied Linguistics* (pp. 145-155). Pécs, Hungary: Lingua Franca Csoport.

Krashen, S., Butler, J., Birnbaum, R., & Robertson, J. (1978). Two studies in language acquisition and language learning. *ITL: Review of Applied Linguistics, 11,* 84-99.

Krashen, S. & Scarcella, R. (1978). On routines and patterns in second language acquisition and performance. *Language Learning*, *28,* 283-300.

Kroll, B. (1990). *Second language writing.* Cambridge: Cambridge University Press.

Kiss, M. (2001). Mérés a budapesti középiskolák 9. osztályaiban: Pedagógiai és háttérelemzés. *Budapesti Nevelő, 36*(2), 3-17.

Larsen-Freeman, D. (1976). An explanation for the morpheme acquisition order of second language learners. *Language Learning, 26,* 125-134.

Larsen-Freeman, D. & Long, M. (1991). *An introduction to second language acquisition research.* London: Longman.

Lasagabaster, D. & Doiz, A. (2003). Maturational constraints on foreign language written production. In M. P. G. Mayo, & M. L. G. Lecumberry (Eds.), *Age and the acquisition of English as a foreign language* (pp. 136-160). Clevedon: Multilingual Matters.

Lazaraton, A. (2005). Quantitative research methods. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 209-224). Mahvah, NJ: Lawrence Erlbaum Associates.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16,* 307-322.

Lee, N. & Huang, Y. Y. (2004). To be or not to be – The variable use of the verb BE in the interlanguage of Hong Kong Chinese children. *RELC Journal, 35*(2), 211-228.

Leech, G. (1992). Corpora and theories of linguistic performance. In J. Svartvik (Ed.) *Directions in corpus linguistics* (pp. 105/122. ). Berlin: Mouton de Gruyter.

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

Lehmann, M. (2006). Vocabulary as a filter with first-year English majors. In M. Nikolov, & J. Horváth (Eds.), *UPRT 2006: Empirical studies in English applied linguistics* (pp. 139-155). Pécs: Lingua Franca Csoport.

Leki, I. (1992). *Understanding ESL writers.* NH: Heinemann Educational Books.

Leki, I. (2002). Second language writing. In R. B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (pp. 60-69). New York: Oxford University Press.

Lewandowska-Tomaszczyk, B. (2003). The PELCRA project – state of art. In B. Lewandowska-Tomaszczyk (Ed.), *Practical applications in language and computers* (pp. 105-121). Frankfurt: Peter Lang.

Lewis, M. (1993). *The lexical approach: The state of ELT and the way forward.* Hove, England: Language Teaching Publications.

Li, Y. (2000). Linguistic characteristics of ESL writing in task-based e-mail activities. *System,* (28) 229-245.

Lindner, G. (2001). A német nyelv mérése – emelt szint. *Budapesti Nevelő, 36*(2), 66-75.

Lightbown, P. (1983). Exploring relationships between developmental and instructional sequences in L2 acquisition. In H. Seliger, & M. Long (Eds.), *Classroom oriented research in second language acquisition* (pp. 217-243). Rowley, MA: Newbury House.

Long, M., & Sato, C. (1984). Methodological issues in interlanguage studies: an interactionist perspective. In A. Davis, C. Criper, & A. Howatt (Eds.), *Interlanguage,* (pp. 253-279). Edinburgh: Edinburgh University Press.

Long, M. H. (2003). Stabilization and fossilization in interlanguage. In C. J. Doughty, & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 268-286). Oxford: Blackwell Publishing.

Lumley, T. (2002). Assessment criteria in a large-scale writing test: what do they really mean to the raters? *Language Testing, 19,* 246-276.

Magnuczné, G. Á. (2003). Hogyan válhat az írás a gondolkodás fejlesztésének eszközévé? *Iskolakultúra,13*(9), 93-98.

Maguire, M., & Graves, B. (2001). Speaking personalities in primary school children's writing. *TESOL Quarterly, 35*(4), 561-593.

Malvern, D., & Richards, B. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Clevedon: Multilingual Matters.

Malvern, D., & Richards, B. (2002). Investigating accomodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing, 19*, 85-104.

Matsuda, P. K. (1998). Situating ESL writing in a cross-disciplinary context. *Written Communication,* (15) 99-121.

McEnery, T., Xiaio, R., & Tono, Y. (2006). *Corpus-based language studies.* New York: Routledge.

McCarthy, M. (1990). *Vocabulary.* Oxford: Oxford University Press.

McKay, P. (2006). *Assessing young language learners.* Cambridge: Cambridge University Press.

McNamara, T. F. (1996). *Measuring second language performance.* London and New York: Longman.

Medgyes, P., & Nikolov, M. (2002). Curriculum development: The interface between political and professional decisions. In R. Kaplan (Ed.), *Handbook of applied linguistics.* (pp. 195-206.). Oxford: Oxford University Press.

Mihaljević-Djigunović, J., Nikolov, M., & Ottó, I. (2006). Horvát és magyar nyelvtanulók angol nyelvtudása nyolcadik osztályban. *Magyar Pedagógia, 106*(3), 171-186.

Mihaljević-Djigunović, J., Nikolov, M., & Ottó, I. (2008). A comparative study of Croatian and Hungarian EFL students. *Language Teaching Research, 12*(3), 433-452.

Milanovic, M., Saville, N., & Shen, S. (1996). A study of the decision-making behaviour of composition markers. In M. Milanovic, & N. Saville (Eds.), *Performance testing, cognition and assessment* (pp.92-114). Selected Papers from the 15th Language Testing Research Colloquium, Cambridge and Arnhem. Cambridge: Cambridge University Press and University of Cambridge Local Examinations Syndicate.

Miralpeix, I. (2006). Age and vocabulary acquisition in English as a foreign language. In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 89-106). Second Language Acquisition 19. Clevedon: Multilingual Matters.

Molnár, E. K. (2002). Az írásbeli szövegalkotás. In B. Csapó (Ed.), *Az iskolai műveltség.* (pp. 193-216). Budapest, Hungary: Osiris.

Molnár, E. K. (2006). Olvasási képesség és iskolai tanulás. In K. Józsa (Ed.), *Az olvasási képesség fejlődése és fejlesztése* (pp. 43-60). Budapest: Dinasztia Tankönyvkiadó.

Muñoz, C. (2006). Accuracy orders, rate of learning and age in morphological acquisition. In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 107-126). Clevedon: Multilingual Matters.

Nagy, J. (2004). A szóolvasó készség fejlődésének kritériumorientált diagnosztikus feltérképezése. *Magyar Pedagógia, 104*(2), 123-142.

Nagyné, S. É. (2001). A német nyelv mérése – normál szint. *Budapesti Nevelő, 36*(2), 57-65.

Nation, I. S. P. (2005). Teaching and learning vocabulary. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 581-595). Mahvah, NJ: Lawrence Erlbaum Associates.

*National Census 2001, Summary Data, Volume 1*. (2002). Budapest: Central Statistical Office.

*Nemzeti alaptanterv* (1995). Budapest, Hungary: Korona Kiadó.

*Nemzeti alaptanterv 2003* (2004). Budapest: Oktatási Minisztérium.

*Nemzeti alaptanterv 2007* [Online]. Available http://www.okm.gov.hu/doc/upload/200708/nat_070815.pdf

Nikolov, M. (1999). Classroom observation project. In H. Fekete, É. Major, & M. Nikolov (Eds.), *English language education in Hungary: a baseline study* (pp. 221-245). Budapest, Hungary: British Council.

Nikolov, M. (2003). Az idegen nyelvi készségek fejlettsége angol és német nyelvből a 6. és 10. évfolyamon a 2002/2003-as tanévben. Budapest: OKÉV.

Nikolov, M., & Csapó, B. (2002). Twelve-year-olds' attitudes towards classroom activities and their performances on tests of English and German as a foreign language. *American Association of Applied Linguists Annual Conference*, Salt Lake City, U. S. A. 6-9 April, 2002.

Nikolov, M., & Józsa, K. (2006). Relationships between language achievements in English and German and classroom-related variables. In M. Nikolov, & J. Horváth (Eds.), *UPRT 2006: Empirical studies in English applied linguistics* (pp.197-224). Pécs: Lingua Franca Csoport, PTE.

Nikolov, M., & Krashen, S. (1997). Need we sacrifice accuracy for fluency? *System, 25*(2), 197-201.

Nikolov, M., & Ottó, I (2005). *Jelentés a nyelvi előkészítő évfolyamos tanulók körében a 2004/2005. tanév őszi félévében elvégzett felmérésről angol és német nyelvből.* [Online]. Available http://www.om.hu/letolt/vilagnyelv/om_nyek_jelentes_2004_osz.pdf

Noijon, J., & Várnai, Zs. (Eds.). (2001). *Tanulói teljesítmény mérése az általános iskolában. A közoktatás-fejlődés mérése Magyarországon című projekt zárótanulmánya.* Budapest, Hungary: Fővárosi Pedagógiai Intézet.

Nystrand, M. (1989). A social interactive model of writing. *Written Communication, 6,* 66-85.

Oakhill, J. (1993). Developing Skilled Reading. in Beard, R. (ed.) *Teaching Literacy: Balancing Perspectives*. London: Hodder and Stoughton.

Orosz, S. (1972). *A fogalmazástechnika mérésmetodikai problémái és országos színvonala* Budapest: Tankönyvkiadó.

Papageorgiou, S. (12 10 2005). EALTA-discussion list.

Peirce, B. (1995). Social identity, investment and language learning. *TESOL Quarterly*, *29,* 9-31.

Pica, T. (1984). Methods of morpheme quantification: their effect on the interpretation of second language data. *Studies in Second Language Acquisition, 6,* 69-78.

Polio, C. (2003). Research on second language writing: An overview of what we investigate and how. In B. Kroll (Ed.), *Exploring the dynamics of second language writing* (pp. 35-69). Cambridge: Cambridge University Press.

Puckett, M. B. & Black, J. K. (2000). *Authentic assessment of the young child.* Upper Saddle River, NJ: Prentice Hall.

Raimes, A. (1987). Language proficiency, writing ability and composing strategies. *Language Learning, 37,* 439-68.

Reynolds, D. (1995). Repetition in nonnative speaker writing: More than quantity. *Studies in Second Language Acquisition, 17,* 185-210.

Richards, B. (1987). Type/token ratios: What do they really tell us? *Journal of Child Language, 14,* 201-209.

Richards, J. C., Platt, J., & Platt, H. (1992). *Longman Dictionary of Language Teaching & Applied Linguistics.* Harlow: Longman.

Sagasta, P. (2003). Acquiring writing skills in a third language: The positive effects of bilingualism. *International Journal of Bilingualism,* (7) 27-42.

Scarino, A., Vale, D., McKay, P., & Clark, J. L. (1988). *The Australian language levels guidelines.* Canberra: Curriculum Development Centre.

Schank, R., & Abelson, R. (1977). *Scripts, plans, goals and understanding.* Hillsdale, NJ: Lawrence Erlbaum.

Schmitt, N. (2000). *Vocabulary in language testing.* Cambridge: Cambridge University Press.

Schumann, J. H. (1978). *The pidginization process: A model for second language acquisition.* Rowley, MA: Newbury House Publishers.

Selinker, L. (1972). 'Interlanguage'. *International Review of Applied Linguistics, 10,* 209-31.

Shen, F. (1988). The classroom and the wider culture: Identity as a key to learning English composition (Staffroom interchange). *College Composition and Communication, 40*(4), 459-466.

Sigott, G. (12 10 2005). EALTA-discussion list.

Silva, T. (1993). Toward an understanding of the distinct nature of L2 writing: the ESL research and its implications. *TESOL Quarterly, 27,* 665-677.

Silva, T., & Brice, C. (2004). Research in Teaching Writing. *Annual Review of Applied Linguistics, 24,* 70-106.

Skinner, B. F. (1957). *Verbal behavior.* Englewood Cliffs: Prentice Hall.

Smith, F. (1981). Demonstrations, engagement, and sensitivity: A revised approach to language learning. *Language Arts, 58,* 103-22.

Smith, F. (1983). Reading like a writer. *Language Arts, 60,* 558-567.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics.* Oxford, UK: Blackwell.

Sturman, Á. (2001). Német nyelv. In J. Noijons, & Zs. Várnai (Eds.), *Tanulói teljesítmény mérése az általános iskolában. A közoktatás-fejlődés mérése Magyarországon című projekt zárótanulmánya* (pp. 67-72). Budapest: Fővárosi Pedagógiai Intézet.

Swain, M. (1985). Large-scale communicative language testing: a case study. In Y. P. Lee, A. C. Y. Fok, G. Lord, & G. Low (Eds.), *New directions in language testing* (pp. 35-46). Oxford: Pergamon Press.

Szalai, T. (2003). Gondolatok az idegen nyelvi írás-kompetencia folyamatorientált fejlesztéséről. *Iskolakultúra, 13*(3), 85-92.

Tagányiné, S. Á. (2001a). Tudásszintmérés angol nyelvből – normál szint. *Budapesti Nevelő, 36*(2), 35-45.

Tagányiné, S. Á. (2001b). Tudásszintmérés angol nyelvből – emelt szint. *Budapesti Nevelő, 36*(2), 46-56.

Taillefer, G. F. (1996). L2 reading ability: further insight into the short-circuit hypothesis. *Modern Language Journal, 80*(4), 461-77.

Takala, S. (2004). Manual for Relating Examinations to the CEFR. *EALTA Confererence,* Granjska Gora, Slovenia, May 14-16, 2004.

Tarone, E. (1988). *Variation in interlanguage.* London: Edward Arnold.

Taylor, L., & Saville, N. (2002). Developing English language tests for young learners. In University of Cambridge Local Examination Syndicate. *Research Notes 7* (pp. 2-5). Cambridge: Cambridge University Press.

Terestyéni, T. (1996). Vizsgálat az idegennyelv-tudásról. *Modern Nyelvoktatás, 2*(3), 3-16.

Tono, Y. (1999, May 1). Using learner corpora for L2 lexicography: Information of collocational errors for EFL learners. [Online]. Available http://www.lancs.ac.uk/postgrad/tono/userstudy/LEXICOS6.html

Tono, Y. (2003). Learner corpora: design, development and applications. In D. Archer, P. Rayson, A. Wilson, & A. McEnery (Eds.) *Proceedings of corpus linguistics* (pp. 800-809). Lancaster: Lancaster University.

Torras, M. R., & Celaya, M. L. (2001). Age-related differences in the development of written production. An empirical study of EFL school learners. Special issue edited by R. M. Manchón: Writing in the L2 classroom: Issues in research and in pedagogy. *International Journal of English Studies, 1,* 103-126.

Torras, M. R., Navés, T., Celaya, M. L. & Pérez-Vidal, C. (2006). Age and IL development in writing In C. Muñoz (Ed.), *Age and the rate of foreign language learning* (pp. 156-182). Second Language Acquisition 19. Clevedon: Multilingual Matters.

Ure, J. (1971). Lexical density and register differentiation. In G. E. Perren, & J. L. M. Trim (Eds.), *Applications of linguistics: Selected papers of the Second International Congress of Applied Linguistics* (pp. 443-452). Cambridge: Cambridge University Press.

Vágó, I. (2007). Nyelvtanulási utak Magyarországon. In I. Vágó (Ed.), *Fókuszban a nyelvtanulás* (pp. 137-144). Oktatáskutató és Fejlesztő Intézet, Budapest.

Vágó, I., & Vass, V. (2006). Az oktatás tartalma. In G. Halász, & J. Lannert, (Eds.), *Jelentés a magyar közoktatásról 2006.* Országos Közoktatási Intézet: Budapest.

Valdes, G., & Sanders, P. (1999). Latino ESL students and the development of writing abilities. In C. Cooper, & L. Odell (Eds.), *Evaluating writing: The role of teachers' knowledge about text, learning, and culture* (pp. 249-278). Urbana, IL: National Council for Teachers of English.

Várnai, Zs. (2001). Angol nyelv. In J. Noijons, & Zs. Várnai (Eds.), *Tanulói teljesítmény mérése az általános iskolában. A közoktatás-fejlődés mérése Magyarországon című projekt zárótanulmánya* (pp. 73-88). Budapest: Fővárosi Pedagógiai Intézet.

Vaughan, C. (1991). Holistic assessment: what goes on in the rater's mind? In L. Hamp-Lyons (Ed.), *Assessing second language writing in academic contexts* (pp.11-25). Norwood, NJ: Ablex.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*(1), 65-83.

Vidákovich, T. (1990). *Diagnosztikus pedagógiai értékelés.* Budapest: Akadémiai Kiadó.

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

Weir, C. J. (1990). *Communicative language testing.* NJ: Prentice Hall Regents.

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing, 22,* 281-300.

Widdowson, H. G. (1978). *Teaching language as communication.* Oxford: Oxford University Press.

Williams, E. (1999). *Extensive reading: Keeping heads above water during book floods.* Unpublished handout at the IATEFL 33rd Conference, Edinburgh.

Wilson, A. (1997). The automatic generation of CALL exercises from general corpora. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and language corpora* (pp. 116-130). Applied linguistics and language study. London: Longman.

White, E. M. (1984). Holisticism. C*ollege Composition and Communication, 35*(4), 400-409.

Winter, E. O. (1977). A clause relational approach to English texts: a study of some predictive lexical items in written discourse. *Instructional Science, 6*(1), 1-92.

Wolfe-Quintero, K., Inagaki, S., & Kim, H. (1998). *Second language development in writing: Measures of fluency, accuracy, and complexity*. Hawaii: University of Hawaii at Manoa.

Zamel, V. (1983). The composing processes of advanced ESL students: six case studies. *TESOL Quarterly, 17*, 165-187.

# 5. A disszertáció magyar nyelvű tézisei

## *5.1 A témaválasztás indoklása*

Nyelvtanárként mindig nagyon érdekelt a magyar diákok, elsősorban az általános iskolai tanulók idegennyelv-tudása. Vajon milyen lehet egy átlagos nyelvtanuló tudásszintje az általános iskolai tanulmányai befejeztével? Szaktanácsadóként ismertem a baranyai megyeszékhely legjobbjainak tudását; évről évre a megyei és országos versenyek zsűritagjaként volt alkalmam értékelni őket. Két pécsi idegen nyelvi tudásmérésben is részt vettem (Bors, Nikolov, Pércsich és Szabó, 1999; Bors, Lugossy és Nikolov, 2001), ahol a megyeszékhely nyolcadikos diákjainak idegennyelv-tudását térképeztük fel. A falvak és kisvárosok tanulóival ellenben csupán egy-egy óralátogatás alkalmával találkoztam.

Régóta szerettem volna objektív képet kapni a megye angolul tanuló diákjainak tudásáról, ezért kiváló lehetőségnek kínálkozott az a horvát-magyar összehasonlító kutatás (Mihaljević-Djigunović, Nikolov és Ottó, 2006, 2008), amely nyolcadikos tanulók anyanyelvi és idegen nyelvi tudásszintjét térképezte fel. Ennek baranyai szervezőjeként és a magyar tanulók írásainak értékelőjeként dolgoztam, és a kutatócsoport engedélyével ezeknek az írásoknak az elemzésére vállalkozom. A disszertáció egyik célja tehát a baranyai 8. osztályos tanulók angol mint idegen nyelvi írásainak elemzése, elsősorban az írásproduktumok vizsgálatával. Ennek a korosztálynak nemzetközileg is kevesen vizsgálták íráskészségét (Molnár, 2002; Muñoz, 2006; Silva és Brice, 2004), feltehetően a tanulók korlátozott nyelvi és háttértudása, valamint fejlődőfélben lévő stratégiai kompetenciájuk miatt.

A disszertáció másik kiindulópontja az értékelés megbízhatóságának kérdése, amely tanítási gyakorlatom során sokszor foglalkoztatott. A nyelvi tudásszint mérésében a nemzetközi nyelvvizsgák és a *Közös európai referenciakeret* (*KER,* 2002) megjelenésével jelentős változások következtek be. Már a *Nemzeti alaptanterv* (*NAT*, 2003) is a *KER* szintjeit használja irányadóként, bár az íráskészség A1 szintjének *NAT*-ban szereplő leírása nem azonos a *Kerettantervben* (2001) kiadott íráskövetelménnyel. Ezek a nem egyértelmű előírások és saját pályám során megtapasztalt értékelési dilemmák indítottak arra, hogy az értékelés megbízhatóságát vizsgáljam. Több szempontból kutattam e területet: elemeztem a *KER* értékelő skálák értelmezésének sikerességét, a holisztikus és analitikus értékelési eljárások megbízhatósága közti kapcsolatot és a különböző értékelők véleménye közti

összefüggést. A kutatási kérdéseket, adatgyűjtő eszközöket és a kutatásmódszertant az 1. sz. táblázat foglalja össze.

## *5.2 A kutatási kérdések*

1. sz. táblázat: Kutatási kérdések, adatgyűjtő eszközök és kutatásmódszertan

| **Kutatási kérdések** | **Adatgyűjtő eszközök** | **Elemzési módszer** |
|---|---|---|
| Milyen a baranyai tanulók íráskészségének fejlettségi szintje más idegen nyelvi képességeikhez viszonyítva? | A nemzetközi projekt adatai a négy nyelvi készségről;<br>Értékelési pontszámok | Analitikus értékelő skála;<br><br>Leíró statisztika |
| Hogyan viszonyul a baranyai általános iskolai tanulók idegen nyelv tudása az ugyanebbe a korosztályba tartozó országos átlaghoz? | Országos mérés eredménye | Leíró statisztika |
| Milyen fejlettségi szintet érnek el írásból angolul a baranyai diákok általános iskolai tanulmányaik végére? | Tanulók írástesztjei;<br>Értékelési pontszámok | Leíró statisztika |
| Mi a kapcsolat a tanulók írásának fejlettségi szintje és a tantervi előírások között? | Értékelési pontszámok,<br>Nemzeti alaptanterv,<br>*KER* | Leíró statisztika |
| Milyen összefüggés van a tanulók szociokulturális háttere és íráskészségük fejlettségi szintje között? | Értékelési pontszámok,<br>Az iskolák szociokulturális háttere | Leíró statisztika |
| Mit tudhatunk meg a tanulók szövegeiről korpusz-elemzéssel? Milyen helyesen használják a 8. osztályosok az egyszerű célnyelvi formákat (folyamatos jelen idő / létezés kifejezése / tagadás / főnevek többes száma)? Milyen nyelvi fejlődési szintek mutathatók ki? | Tanulók írástesztjei;<br>Morféma-kutatások | korpusz-elemzés |
| Milyen összefüggés van a kritérium-orientált és korpusz alapú nyelvhelyesség-mérés között? | Értékelési pontszámok,<br>Nyelvészeti elemzés adatai | Korreláció analízis<br>Tartalomelemzés |
| Milyen összefüggés van az értékelők minta tesztekről kialakult véleménye között? | Sorrendbe állított minta tesztek; *KER* kritériumai alapján értékelt írások | Korreláció analízis |
| Milyen kapcsolatot mutat a kritérium-orientált és holisztikus értékelés? | Értékelési pontszámok,<br>Rangsorolt minta tesztek; *KER* kritériumai alapján értékelt írások | Korreláció analízis |
| Milyen összefüggés van az értékelők *KER* skála értelmezései között? | Minta tesztek értékelése a *KER* kritériumai alapján | Korreláció analízis |
| Milyenek a tipikus teljesítmények? Milyen fejlődési szintet képviselnek? | Öt tanuló írása;<br>morféma-kutatások | Leíró statisztika,<br>Tartalomelemzés |
| Milyen szókincs jellemzi a 8. osztályosokat? Milyen összefüggés van a kritérium-orientált és korpusz alapú nyelvhelyesség- és szókincsmérés között? | Öt tanuló írása;<br>Értékelés pontszámai,<br>Nyelvészeti elemzés adatai | Szókincs gazdagsága és a gyakoriság elemzése |

### *5.3 Az értekezés felépítése*

A disszertáció négy fejezetet tartalmaz: az első háromban a kutatás elméleti hátterét mutatom be, a negyedikben az empirikus kutatást. Az első fejezet az olvasás- és íráskészség kapcsolatát mutatja be anyanyelvi és idegen nyelvi vonatkozásban, külön kitér az idegen nyelven történő írás sajátosságaira. A továbbiakban az írás vizsgálatának három fő megközelítési módját elemzi, melyek a (1) produktumra, (2) az írás létrehozásának szociokulturális körülményeire, illetve (3) az írás közben lejátszódó kognitív folyamatokra helyezik a hangsúlyt. Disszertációmban az idegen nyelvi írás produktumának elemzésére vállalkozom a tanulók intézményi szinten vizsgált szociokulturális hátterének tükrében.

Az első fejezet a tanulói nyelv vizsgálatának eredményeit is összefoglalja a morféma-kutatások áttekintésével, ezt követően tartalmazza a legfontosabb tudományos elemzéseket, melyeket az idegen nyelvi íráskészséggel kapcsolatban végeztek általános iskolai korosztályban.

A második fejezet az írás értékelésének kérdéskörét tekinti át kritikusan, kitérve a tesztelés céljára, az értékelő skálák fő típusaira, az értékelés megbízhatóságának biztosítására. Bemutatja a *Közös európai referenciakeret* (KER, 2002) célját és alsó három szintjét (A1-B1), melyekre kutatásomban támaszkodtam. Ez a rész elemzi azokat a korpusz-nyelvészeti eljárásokat, melyeket az írások értékelésében fel lehet használni.

A harmadik fejezet helyezi kontextusba a kutatást, bemutatja a magyar diákok szociokulturális hátterét, a tantervi követelményeket anyanyelvből és idegen nyelvből, kiegészítve ezt a területet a hazai nyolcadikos korosztályra vonatkozó idegen nyelvi mérések eredményeivel.

A negyedik fejezetben ismertetem empirikus kutatásomat, melyben Baranya megyei 8. osztályos tanulók angol íráskészségét vizsgálom egy egyszerű, képek összehasonlítását igénylő feladattal. Az írásokat több szempontból elemzem, mellyel egyúttal az értékelés megbízhatóságát is vizsgálom. A tanulók eredményeit először iskolánként vetem össze a szociokulturális háttér hatását vizsgálva.

Az értékelés megbízhatóságát öt vizsgáztató bevonásával kutatom, akik mind holisztikus, mind analitikus módon értékelték a kiválasztott mintát. Munkájukat a *KER* (2002) skála deskriptorainak használatával végezték, ezzel a *KER* átláthatóságát, érthetőségét is megpróbálom feltárni. Végül öt szöveg kvalitatív elemzésével ellenőrzöm, hogy az értékelések és az írások minősége milyen kapcsolatban állnak egymással. Az összegzésben az eredmények számbavétele, a kutatás korlátai és pedagógiai következtetések szerepelnek.

### *5.4 Elméleti háttér*

A disszertáció elméleti gyökerei az idegennyelv-elsajátítás kutatásának több területét átszövik. Egyik fontos terület az anyanyelv és a második nyelv összefonódása a kétnyelvűség kialakulása során, melyet az interdependencia hipotézis (Cummins, 1981) fogalmaz meg. Eszerint a nyelvek kölcsönösen hatással vannak egymásra, feltéve, hogy a tanuló elér egy küszöbszintet az adott idegen nyelven. Ezt nem egy statikus nyelvi szintként kell elképzelnünk, hanem a nyelvtanuló egyéni sajátosságaként, amely függ az idegen nyelv tanulásával eltöltött időtől, a nyelvtanulás minőségétől, és azoktól a kognitív folyamatoktól, amelyeket a diák eddigi tanulása során megélt. Ez azt jelenti, hogy az anyanyelven elsajátított ismeretek és nyelvi készségek dinamikusan hatnak az idegen nyelvi ismeretek és készségek fejlődésére és viszont. Vajon igaz-e ez a feltevés az idegen nyelvi íráskészségre is? Az anyanyelvi íráskészség hogyan ad alapot az idegen nyelvi írás folyamatában?

Számos kutatás mutat rá (Cumming, 1989; Krapels, 1990; Kroll, 1990), hogy az interdependencia érvényesül az írott szövegekben is egy bizonyos nyelvi szint fölött, mivel ugyanazok a kognitív folyamatok játszódnak le az anyanyelvi és idegen nyelvi szövegek alkotásakor. Ugyanakkor, az idegen nyelven történő írás folyamán a tanulók nyelvi nehézségekkel küzdenek, ezért Silva (1993) szerint kevesebb időt fordítanak a tervezésre, a tartalmat kevésbé ellenőrzik, és hibásabban fogalmaznak, mint anyanyelvükön.

A nyelvek közti általános transzfer mellett fontos szerepet játszik a különböző nyelvi készségek közti átjárás is, így az olvasás- és íráskészség egymásra hatása is. Krashen (1984), Elley (1998), és Bors (1999) beszámolnak az extenzív olvasás pozitív hatásáról a tanulók idegen nyelv tanulása iránti motivációjára és nyelvi teljesítményére, beleértve az íráskészséget is.

Hazánkban kevesen foglalkoztak az anyanyelvi és idegen nyelvi íráskészség összehasonlításával. Kiszely (2003) és Magnuczné (2003) a középiskolai és egyetemista diákok írásait elemezték. Kiszely rámutat, hogy a különböző szövegszerkezeti jellegzetességek idegen nyelvbe való átviteléhez különböző küszöbszintek léteznek. Magnuczné kiemeli, hogy a jó idegen nyelvi fogalmazás írásához nem elegendő a megfelelő nyelvtudás, szükség van a tartalom elemzésének és rendezésének képességére is. Ezeket a készségeket lehet fejleszteni, ha az írást folyamatként és nem produktumként kezeljük, melyre sajnos kevés dokumentált példa van Magyarországon (Horváth, 2001).

A másik fontos terület, melyet felhasználtam, a tanulói köztes nyelv sajátosságainak kutatása (Selinker, 1972), és a morfémák elsajátításának sorrendje, melyről sok egymásnak

ellentmondó adat felsorakoztatása után (pl. Dulay és Burt, 1973; Hakuta, 1974), bebizonyosodott, hogy lényegében többnyire azonos az anyanyelv- és idegennyelv-elsajátítás során, függetlenül a tanuló anyanyelvétől. Ez az eredmény szorította háttérbe a behaviorista tanulásszemléletet, mely a külső ingerek hatását tartotta a leglényegesebb elemnek a tanulás szempontjából, szemben a tanuló kognitív képességeinek hozzájárulásával a tanulási folyamathoz.

A morféma-kutatások adatainak újra elemzése során Goldschneider és DeKeyser (2001) arra a megállapításra jutottak, hogy az elsajátításban a legfontosabb szerepet az adott morféma esetében az játssza, hogy mennyire kerül a figyelem előterébe (*salience*), tehát a második nyelv elsajátítása az input hatására a tanuló belső mechanizmusainak működésével valósul meg. Disszertációmban a nemzetközi eredmények ismeretében keresztmetszeti mérésben vizsgáltam megyénk tanulónak angol nyelvi morféma-elsajátítási sorrendjét és összehasonlítottam azt egy korábbi hazai kutatás adataival (Nikolov és Krashen, 1997).

A fiatal nyelvtanulók csoportjában végzett kevés kutatás közül a spanyol BAF projekt (Muñoz, 2006; Torras és mtsai., 2006) és az ehhez kapcsolódó idegen nyelvi tudást feltérképező vizsgálatok (pl. Lasagabaster és Doiz, 2003) egyik fontos kérdése az életkor szerepe a nyelvelsajátításban. Egybecsengenek megállapításaik, miszerint a korai kezdés (8 éves kor) előnyt jelent a folyékony beszédkészség (*fluency*) kialakulásában, míg a 12 éves kor körüli kezdés az összetettebb nyelvi eszközök gyorsabb fejlődését eredményezi. Adataik szerint ez az életkor jelentett fordulópontot az íráskompetencia elsajátításában is, melyet több tényező befolyásolt (pl. az életkor, az oktatási órák száma, a nyelvtudás szintje). Kutatásom résztvevői a kezdés szempontjából valamennyien a korai kezdő csoportba tartoznak, de különbségek vannak köztük ezen belül is: egy részük 7, másik részük 10 éves korában kezdte az angol nyelv tanulását a közoktatásban.

Hazánkban az anyanyelvi íráskészséget feltérképező kutatások (pl. Kádárné, 1990) alacsonyabb szintet találtak az elvártnál, illetve kismértékű fejlődést mutattak (Molnár, 2002) a tanulók íráskészségében 13-17 éves koruk között. Az általános iskolások idegen nyelvi írástudását a helyi méréseken (pl. Bors, Lugossy és Nikolov, 2001; Bukta és Nikolov, 2002) kívül néhány országos kutatás is vizsgálta (pl. Csapó és Nikolov, 2002; Nikolov és Józsa, 2006). Ezek egybehangzóan az íráskészség többi nyelvi készségtől való elmaradását állapították meg, melyet részben az osztálytermi eljárások alacsony hatékonyságának tulajdonítottak (Nikolov, 2003).

A nyelvelsajátítás folyamatának tanulmányozása mellett fontos a nyelvi produktumok értékelésének módja is, mert ez lényeges visszajelzést adhat a tanulónak a további feladatokra

vonatkozóan, a nyelvtanárnak pedig a fejlesztési célok meghatározásában. Áttekintésemben a holisztikus és analitikus értékelő skálák előnyeit és hátrányait elemeztem (Barkaoui, 2007; Hamp-Lyons, 1990; White, 1984), valamint a vizsgáztatók skála-értelmezési eljárásait (pl. Bukta, 2008; Lumley, 2002).

Mivel a *NAT* (2003) idegen nyelvi követelményének értelmezéséhez és a jelen kutatás értékelési eljárásához is szükséges a *KER* (2002) szintjeinek ismerete, értekezésemben részletesen elemzem az európai dokumentum három alsó nyelvi szintjének deskriptorait, valamint egybevetettem a magyar oktatási előírásokat a *KER*-skála szintleírásával. Az ezekben a dokumentumokban tapasztalt pontatlanságot és félreérthetőséget az empirikus kutatás eredménye megerősítette.

## *5.5 A kutatás módszerei*

### Résztvevők

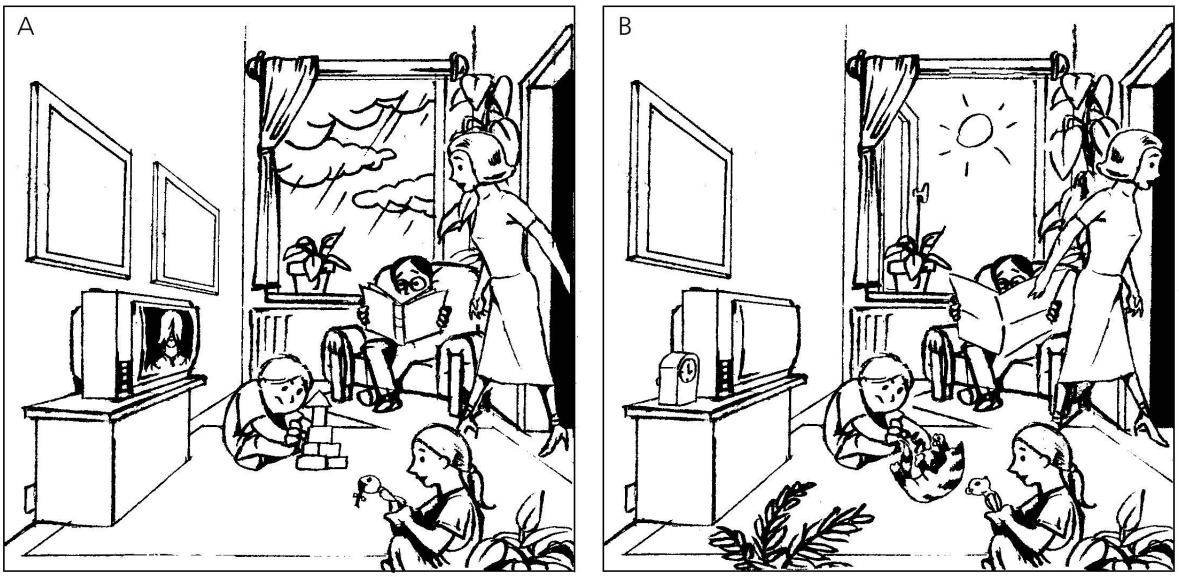
A kutatásban 231 baranyai nyolcadik osztályos vett részt, akik közel reprezentatív arányban tanultak falusi (18%), kisvárosi (23%) és nagyvárosi (Pécs, 59%) iskolákban. A pécsi, kiemelkedő szociokulturális körülmények között dolgozó iskolák 5-6%-kal magasabb arányban képviseltették magukat a kisvárosiakkal szemben az országos reprezentatív mintához képest (Csizér, 2007). A tanulók különböző életkorban kezdtek ismerkedni az angol nyelvvel és eltérő heti óraszámban tanulták azt, ezekkel a háttérváltozókkal azonban jelen kutatás nem foglalkozik, egyedül a szociokulturális háttér hatását próbálja kimutatni intézményi szinten. A háttérváltozók hatásai megismerhetőek a horvát-magyar kutatást összegző cikkekből (Mihaljević-Djigunović, Nikolov és Ottó, 2006, 2008).

Az értékelés megbízhatóságának és következetességének méréséhez öt vizsgáztatót kértem fel 15 minta-teszt több szempontú újraértékelésére.

### Mérőeszközök

#### a) Angol nyelvi írásfeladat

Az angol nyelvi szint méréséhez a teszt érvényességét 2002-ben kipróbálták (Nikolov, 2003). A teszt úgy készült, hogy lefedje a *KER* azon részét, amelyeket nyolcadik osztály végén a résztvevők várhatóan teljesíteni tudtak (A1 és az A2 szint alsó sávja). A tanulók feladata (1. ábra) két hasonló, de tíz apró különbséget tartalmazó kép összehasonlítása volt, melyhez a tíz kulcsszót megadtuk.




1. ábra: Az írásfeladat

Look at pictures A and B. Write about the differences between them. Write about the boy and the girl, the man, the woman, the TV, the clock, the wall, the plants, the window and the weather outside.

A témakört és a képleíráshoz szükséges szókincset, valamint nyelvi formákat ismernie kellett a tanulóknak, mivel ezek a *Nemzeti alaptanterv* (NAT, 1995) előírásaival összhangban vannak.

**b) A vizsgáztatók feladatlapja**

Az értékelők három feladatot kaptak:

(1) a 15 kiválasztott írást rangsorolniuk kellett,

(2) ugyanezeket az írásokat a *KER* A1-B1 szintjeit leíró deskriptorok alapján kellett besorolniuk,

(3) írásban visszajelzést kértem tőlük az értékelés folyamatáról.

**c) Értékelő skálák**

Az íráskészséget mérő feladatnál négy szempont alapján készült a skála: a feladat teljesítése, szókincs, nyelvhelyesség és szöveg kohézió. A készséget értékelő vizsgáztatók előzetes tréningen vettek részt.

A második, kisebb mintán (15) végzett értékeléshez a *KER* skálát használtuk. Az öt vizsgáztatónak a szövegeket a *KER* véletlenszerű sorrendben táblázatba tett A1-B1 deskriptorai alapján a következő négy, írásokat jellemző kritériumsorban kellett elhelyezniük:

(a) általános nyelvi szint
(b) általános írásprodukció
(c) szókincs gazdagsága
(d) nyelvhelyesség

A vizsgáztatóknak tehát először értelmezniük kellett a deskriptorokat és A1-B1 sorrendbe tenniük ahhoz, hogy elkezdhessék a tanulók szövegeinek értékelését.

**A kutatás menete**

2004 májusában írták meg a tanulók a felmérést osztálytermi környezetben, ahol 45 perc állt rendelkezésükre a hallás utáni értés és az írásfeladat teljesítésére. Az értékelés után minden iskola megkapta saját eredményeit és a teljes kódolt statisztikát 2004 novemberében. Az adatbevitelre 2005-ben, a kódolásra és a második értékelésre 2006-tól kezdődően került sor.

A tanulók szövegeit már ismertem az analitikus értékelés idejéről, a szövegek kódolását az azokban legjellemzőbben előforduló nyelvi formák figyelembe vételével végeztem el: a folyamatos jelen idő (PRC), létezés kifejezése (EXP), tagadás (NEG) és a többes szám kifejezése (PLU) volt a négy leggyakoribb nyelvi jelenség. Ezeknek minden helyesen s helytelenül használt alakját külön kóddal láttam el. A négy leggyakoribb nyelvi forma használatán kívül egyéb nyelvi jelenségek hibáit is kódoltam (pl. szórendi hibák, on/in elöljáró cseréje), végül ezeket a hibákat is beszámoltam a helytelenül használt alakok számításánál.

Előzőleg a tanulók szövegeit a korpusz-elemzés előkészítéseként felvittük számítógépre, majd tagmondatokra bontottam azokat és soronként jelöltem a kiválasztott nyelvi alakok jó és helytelen megoldásait.

A nyelvhasználat helyességi mutatóját a kötelező előfordulás-elemzés (Obligatory Occasion Analysis) módszerével számoltam ki: a helyes nyelvi alakok számát osztottam az összesen használt alak számával. Az arányszámításokat és a korrelációkat is SPSS szoftver segítségével végeztem, ennek köszönhetően sikerült felfedeznem a kódolás során kialakult tévedésemet a kiegészítő hibák kódolásának szükségességéről. Az értékelők által kialakított sorrendet és *KER*-szintbe sorolásokat ugyanezzel a programmal elemeztem.

## *5.6 A kutatás eredményei*

Az eredményeket a kutatási kérdések megválaszolásával mutatom be.

*1. Milyen a baranyai tanulók íráskészségének fejlettségi szintje más idegen nyelvi képességeikhez viszonyítva?*

Az utóbbi évtizedben hazánkban végzett kismintás és reprezentatív kutatások (Bors, Nikolov, Pércsich, és Szabó, 1999; Bukta és Nikolov, 2002; Nikolov, 2003; Tagányiné, 2001 a, b; Várnai, 2000) egyaránt azt mutatták, hogy a tanulók íráskészsége jócskán elmarad többi idegen nyelvi készségüktől. Kutatásom eredményei a gyengébb íráskészség tényét megerősítik, de az általam mért populáció írásteljesítménye csak 6%-kal alacsonyabb az egyébként kiegyensúlyozott 60%-os szövegértési és pragmatikai eredményénél. Ennek okait a következő kérdésnél igyekszem megmagyarázni.

*2. Hogyan viszonyul a baranyai általános iskolai tanulók idegen nyelvi tudása az ugyanebbe a korosztályba tartozó országos átlaghoz?*

Ahhoz, hogy a baranyai diákok eredményét értelmezhessük, összehasonlítási alapot kellett találnom. Egy korábbi országos reprezentatív idegen nyelvi mérés (Csapó és Nikolov, előkészületben) adatait használtam fel, amely azonos feladattal és értékelő skálával dolgozott. Összehasonlítva a két kutatást: a baranyai nyolcadikosok hallás utáni értése és olvasáskészsége kissé magasabb szinten áll. Az átlagos írásteljesítmény országos szinten 33.7%, míg Baranyában elérte az 53.93%-ot. Nehéz megmagyarázni a jelentős eltérést. Feltételezésem szerint okozhatja ezt a előnyösebb szociokulturális háttérrel rendelkező tanulók túlreprezentáltsága baranyai mintánkban. Ennek oka lehet az, hogy az országos mérésnél kijelölték iskolákat a reprezentativitás elérése céljából, míg Baranyában az iskolák nyelvtanárait személyesen kellett meggyőznöm a mérésben való részvétel előnyeiről, melyet sokan nem láttak be, ezért nem is vállalták azt. Akik viszont vállalták, feltehetően a magabiztosabb, nyitottabb tanárok közül valók, akiknek ez az eredményeiben is megmutatkozik.

3. *Milyen fejlettségi szintet érnek el írásból angolul a baranyai diákok általános iskolai tanulmányaik végére?*

Bár az 54%-os eredmény az írásfeladaton sokkal magasabb, mint az országos átlag, a részt vevő 11 iskola egyenkénti teljesítménye nagyon széles skálán helyezkedik el.

A négy legjobban teljesítő iskola az összpontszám alapján 64-82%-os eredménnyel messze túlszárnyalta a többit. Ezek közül is kiemelkedik két intézmény (N=99) átlagos írásteljesítménye (76% és 82%), amely magasan meghaladja az általános iskola végén elvárt tudást. Három iskola kissé alulteljesített 29-36%-os átlagokkal, további négy iskola 20% alatti eredményekkel erősen lemaradt a mezőnytől és a minimumkövetelménytől. Az egyes szempontok (a feladat teljesítése, szókincs, nyelvhelyesség és szöveg kohézió) teljesítési szintje alapján értékelve az iskolákat hasonló sorrend alakult ki.

*4. Mi a kapcsolat a tanulók íráskészségének fejlettségi szintje és a tantervi előírások között?*
Az értékelők négy szempont alapján kialakított véleménye megmutatta, melyik intézmény tanulói milyen sikeresen oldották meg az ő tudásszintjükre megalkotott feladatot. Az eredmények viszonyítása a tantervi követelményekhez úgy történt, hogy a mérőeszköz által leírt öt nyelvi szint középső szintjét jelöltük meg, mint a teljesítés minimumát, amely becslésünk szerint a *KER* A1-es szintjének felel meg. A középső szinten szerzett 12 pont 37,5%-os feladatteljesítésnek felel meg a 32 pontos feladaton. Az e szint alatti teljesítményeket tekintettük alulteljesítésnek.

Ezúttal a négy nyelvi szempont alapján külön vizsgáltam a tanulók íráskészségét, így a teljesítmények szóródása nagyobb, mint az összpontszámok esetében. Mind intézményi, mind egyéni alapon megnéztem, hányan felelnek meg a *Nemzeti alaptanterv* (2003) által előírt követelményeknek.

Intézményenkénti számítás alapján öt iskola teljesítette a tantervet és érte el az A1-es nyelvi szintet. Két iskola (N=99) tanulói 73-90%-ot értek el, további két iskola (N=39) diákjai 59-76%-ot nyújtottak a négy nyelvi szempont szerint. Egy intézmény tanulói (N=10) 30-41%-kal éppen a minimum szint körül teljesítettek. Tehát összesen 148-an (64%) érték el a tantervi minimum követelmény szintjét intézményi alapon számítva. Ezek között a diákok között nyilvánvalóan voltak, akik nem érték el a kívánt szintet, de a többiek magasabb teljesítménye felhúzta az átlagot. Ezért néztem meg a diákok egyéni eredményeit is.

Érdekes módon az egyéni teljesítmények áttekintése is hasonló végeredményt hozott: 149 tanuló pontszáma érte el a középső, 12-16 pontos sávot, amelyet az A1-es szinttel azonosítottunk. Ha lejjebb visszük az A1-es szintet skálánkon, amint ezt sugallták az öt vizsgáztató (11. kutatási kérdés) *KER* alapján történő besorolásai, akkor még 15 szöveget kapcsolhatunk az A1-es szintet teljesítők közé, ezzel a tanulók 71%-a éri el a tantervi követelményszintet.

*5. Milyen összefüggés van a tanulók szociokulturális háttere és írásuk fejlettségi szintje között?*

Számos kutatás (Andor, 2000; Vágó, 2007) bizonyította, hogy a diákok idegennyelv-tanulási lehetőségeit jelentősen befolyásolja szociokulturális hátterük, amelyek aztán későbbi eredményeiket erősen befolyásolják (Bors, Lugossy, és Nikolov, 2001; Bukta és Nikolov, 2002; Csapó, 1998, 2002).

Jelen kutatás megerősítette a szociokulturális háttér szerepét az intézmények szintjén vizsgálva. Településtípus szerint nagyvárosi, kisvárosi és falusi iskolák vettek részt a vizsgálatban, ezen belül a nagyvárosi intézményeket két csoportra oszthatjuk, a bel- és külvárosi iskolákra. A belvárosi, elit iskolák tanulói kimagasló íráskészséget mutattak, míg a külvárosi, nehéz körülmények között élő diákok a lemaradók csoportjába kerültek. A kisvárosi tanulók írásai széles tudásskálán helyezkedtek el, míg a falusi iskolák a gyenge eredményei ezen a teljesítményszinten belül is nagy különbségeket mutattak. Utóbbi két településtípusban nincs elegendő számú oktatási intézmény ahhoz, hogy külön intézményekbe kerüljenek az előnyös szociokulturális háttérrel rendelkező tanulók.

*6. Mit tudhatunk meg a tanulók szövegeiről korpusz-elemzéssel? Milyen helyesen használják a 8. osztályosok az egyszerű célnyelvi formákat (folyamatos jelen idő / létezés kifejezése / tagadás / főnevek többes száma)? Milyen nyelvi fejlődési szintek mutathatók ki?*

A kritérium-alapú értékelés eredményeinek ismeretében érdekes megvizsgálni, hogy a korpusz-elemzés adatai mit mutatnak, és a két mérőeszköz által nyert adatok milyen kapcsolatban vannak egymással. Először nézzük meg a négy vizsgált nyelvi forma előfordulásának és helyes használatának viszonyát! Ezek fordított arányosságot mutatnak egymással, ami a NEG nyelvi megformálásánál a legészrevehetőbb.

Legtöbben a PRC-t (94%) és a PLU-t (87%) alkalmazták, de a létezés kifejezésére (76%) is vállalkozott a diákok nagy többsége. A NEG angol nyelvi megformálását a tanulók közel fele kísérelte meg.

Ha nem a vizsgált formákat használó tanulók száma, hanem az összesen papírra vetett nyelvi forma száma alapján nézzük a gyakorisági rangsort, akkor a PRC vezet 52%-os aránnyal. Az alkalmazott nyelvi formák közel negyedét teszi ki a EXP (24%), míg a PLU 16%-ban szerepel a négy nyelvi elem között, a NEG pedig alacsony arányban (8%). Ezek az eloszlási viszonyok nem meglepőek, mivel a feladatból természetesen adódott a PRC és EXP gyakoribb használatának szükségessége.

A négy nyelvi forma alkalmazásának sikeressége ellenkező sorrendet jelez. A legsikeresebben elsajátított nyelvi formának a NEG tűnik 79%-os helyességi mutatóval, ezt követi a PLU alkalmazása 65%-os eredménnyel. A EXP (54%) és a PRC (48%) egymáshoz közeli sikerességi arányt mutat. Abból, hogy tanulóink a négy vizsgált forma közül először a tagadást, majd a többes számú főnévi alakokat képesek helyesen alkalmazni, arra következtethetünk, hogy ezeket a nyelvi formákat sajátítják el korábban. Ezután következik a létezés kifejezésének és a folyamatos jelen időnek a helyes megformálása a teljes minta adatai alapján. Ezek az eredmények összecsengenek egy korábbi tanulmány következtetéseivel (Nikolov és Krashen, 1997), melyben pécsi 7-8. évfolyamos tanulók beszélt nyelvi mérése alapján a PLU elsajátítása megelőzte a PRC helyes használatát.

Jelen kutatásban ugyanakkor a EXP és PRC helyes használatának sorrendje nem mutat azonos mintát minden iskolában. Öt helyen, a négy legmagasabb eredményt elérő intézményben és az egyik középmezőnybe tartozó falusi iskolában a EXP sikeresebb, mint a PRC megformálása. Az alulteljesítő iskolákban viszont a folyamatos jelen időt helyesebben használják a létezés jelölésénél, ugyan mindkettőt alacsony hatékonysággal.

Ennek magyarázata véleményem szerint az elsajátítás úgynevezett U-alakú fejlődési szakaszaiban rejlik (Bailey, Madden, és Krashen, 1974). Eszerint a PRC-t az elsajátítás korai szakaszában azért használják helyesen a nyelvtanulók, mert memorizált egységként, elemzés nélkül alkalmazzák. A morféma-elsajátítás későbbi szakaszában már több nyelvi formát ismernek a jelenidejűség kifejezésére; ezeknek egyes formai elemeiből túláltalánosítással szabályt alkotnak és a különböző morfémákat váltogatva használják addig, amíg el nem érik a teljes elsajátítás szintjét. A EXP-nek fejlődésbeli eltolódása talán azzal magyarázható, hogy később kerül a magyar tanulók figyelmének előterébe (salient) az összetett angol nyelvi forma *(there is / are)*, ugyanis anyanyelvünkben a létige önmagában kifejezi a létezés fogalmát.

Disszertációmban nem vizsgáltam az osztálytermi input szerepét, mely a morfémák megjelenési gyakoriságát biztosítja és valószínűleg egyik lényeges kialakítója a morféma-elsajátítás sorrendjének (Goldschneider és DeKeyser, 2001), ezért nem tudok erre vonatkozó megállapításokat megfogalmazni az elsajátítás sorrendjével kapcsolatban.

*7. Milyen összefüggés van a kritérium-orientált és korpusz alapú nyelvhelyesség-mérés között?*

Ez a kutatási kérdés az értékelés megbízhatóságára keres választ: a tanulók írásainak nyelvhelyességére adott vizsgáztatói pontszám és a korpuszban bekódolt nyelvi formák helyességi mutatói között keresi az összefüggést. Ennek kimutatására Pearson-féle korrelációs

együtthatókat számoltam, melyek szoros pozitív összefüggést jeleztek a helyesen alkalmazott nyelvi formák száma és a nyelvhelyességre adott pontszámok között, ezzel szemben mérsékelt negatív kapcsolatot mutattak a hibák száma és a pontszámok között. A hibák és pontszámok közötti negatív kapcsolat egyértelmű; a mérsékelt összefüggés azt jelzi, hogy nem egyenes arányban csökkent a pontszám a hibák számának növekedésével, mivel az értékelés a kommunikatív nyelvtanítási szemléleten alapult, tehát azonos hibák ismétlődéséért nem vont le többször az írás értékéből.

Regresszió-elemzéssel azt is megvizsgáltam, hogy az összes kódolt nyelvi forma helyesen, illetve hibásan alkalmazott száma milyen mértékben képes megmagyarázni a nyelvhelyességi pontszámok között kialakult különbségeket. Együttes hatásuk a varianciát 64%-ban magyarázza. Ugyanilyen modellel ellenőriztem a négy kiválasztott nyelvi forma (PRC, EXP, NEG, PLU) helyességének együttes hatását a nyelvhelyességi pontszámok között kialakult különbségekre, ezek közösen 75%-ban járultak hozzá a pontszámok varianciájához, tehát jelentősebb mértékben befolyásolták azt, mint a hibás és helyes nyelvi alakok általában véve. Az ellentmondás a két adat között segített felfedeznem a kódolás során kialakult téves eljárást a kiegészítő hibák kódolásával kapcsolatban. Ezt a felismerést később az írások kvalitatív elemzése is megerősítette.

A négy nyelvi forma hatását az értékelésre egyenként is megvizsgáltam, eszerint egyedül a PLU helyes használatának mértéke nem járult hozzá szignifikánsan a pontok alakulásához. A modell szerint legfontosabb változónak a EXP bizonyult, melynek helyességi aránya a pontszámok 32%-át magyarázza, ezt követi a PRC 20%-os magyarázó erővel, majd a NEG 3%-kal.

Az egyes nyelvi formák helyességi mutatóinak magyarázó ereje a nyelvhelyességet kifejező pontszámban azonban nem arányos ezeknek a formáknak a szövegbeli gyakoriságával, tehát nem az EXP szerepelt legtöbbször a tanulók írásaiban; a PRC több mint kétszer annyiszor volt jelen. A vizsgáztatót valószínűleg a használt formák kommunikatív értéke befolyásolta leginkább a pontok kialakításánál: a létezés kifejezésében elkövetett hibák nagyobb arányban vezethettek félreértéshez (pl.: *It's a clock by the TV*, a *There is a clock by the TV* helyett, vagy *It's a not plants* a *There are no plants* helyett), mint a PRC helytelen használata (pl.: *mother is go* az *is going* helyett). A PLU jelölésének elmulasztása nem rontja jelentősen a megértést, mivel a főnév előtt álló számnév már utal a *több* fogalmára, a NEG-t pedig a *no/ not* jelenléte sugallja akkor is, ha helytelen a nyelvi megformálása.

*8. Milyen összefüggés van az értékelők minta tesztekről kialakult véleménye között?*

*9. Milyen kapcsolatot mutat a kritérium-orientált és holisztikus értékelés?*

Az értékelés megbízhatóságát öt vizsgáztató bevonásával mértem, akik az értékelés második szakaszában 15 diák íráskészségéről alkottak véleményt kétféle értékelési eljárás alkalmazásával: holisztikus és analitikus módszerrel. Így a vizsgáztatók véleménye közötti kapcsolatot is kutattam, valamint a két értékelő eljárás eredménye közötti összefüggést.

A vizsgáztatók közötti megbízhatóság a Cronbach alpha mutatói szerint magas volt minden területen, míg a teljes egyetértést megkívánó Krippendorff alpha érték csak a rangsor-kialakítás tekintetében jelzett magas egyezést; a szókincs megítélésében mérsékeltet (.59). Ezután Spearman rangsor-korrelációs számítást végeztem, amely magas szintű egyetértést jelzett a vizsgáztatók között (.83-.928) a 15 szöveg rangsorolásában. Ezek az eredmények Barkaoui's (2007) megállapítását erősítik meg a holisztikus értékelés jobb megbízhatóságáról a kritérium-alapúval szemben. Jelen disszertáció valamivel alacsonyabb megbízhatósági adatai a kritérium-alapú értékelés területén magyarázhatóak azzal, hogy - a kutatási célnak megfelelően - nem standardizáltuk és validáltuk a használt *KER* (2002) skálákat.

Az öt szakember négy területen hasonlította össze a diákok írásteljesítményét, így minden nyelvi kritériummal kapcsolatban tíz korrelációs együtthatót tudtam vizsgálni. A szókincs megítélésében volt legnagyobb az összhang, mind a tíz kapcsolat szignifikánsnak bizonyult. A nyelvhelyesség értékelése hét szignifikáns összefüggést adott, melyek közepes és erős egyezést mutattak. Ezt követte az általános nyelvi szint hat szignifikáns kapcsolata az előzőhöz hasonló skálán, majd az írásteljesítmény besorolásai közti öt érvényes összefüggés. Az összes lényeges eltérést az értékelők között a *KER*-skála értelmezésének különbsége, esetenként a deskriptorok félreértése okozta. A következő kérdés ezt a területet járta körül.

*10. Milyen összefüggés van az értékelők KER skála értelmezése között?*

Erre a kérdésre három eszköz felhasználásával kerestem a választ:

(1) a megbízhatósági indexek kiszámításával (amelyet az előző részben tárgyaltam),

(2) a vizsgáztatók visszajelzéseinek értékelésével,

(3) öt tanulói írás kvalitatív elemzésével, melyek lehetőséget biztosítottak arra, hogy az írások minőségét az értékelők *KER*-kritériumok alapján tett besorolásával összevessem.

Mind a vizsgáztatók visszajelzéseinek elemzése, mind a többi értékeléstől nagyon eltérő írás-besorolásaik azt mutatták, hogy a vizsgáztatók félreértettek néhány deskriptort. Ez megerősíti az értékelési szakértők (Alderson, 2002) azon megállapításait, hogy a *KER*-skálák nem elég

világosan értelmezhetőek, tehát nem használhatóak megbízhatóan standardizálás nélkül. Ezért tanácsos a *Kézikönyv*-re (Takala, 2004) támaszkodni és végigvinni azt a folyamatot, mely a skála megismeréséből, specifikációjából, standardizálásából és empirikus validálásából áll, ha a skálát saját tesztjeink *KER*-hez illesztésére szeretnénk használni. Ezt a folyamatot már több országban kipróbálták saját vizsgáik kialakítása során.

*11. Milyenek a tipikus teljesítmények? Milyen fejlődési szintet képviselnek?*

A baranyai 8. osztályos diákok angol nyelvi tudásáról és nyelvi fejlődési szintjéről sok általános adatot ismertünk meg az iskolai teljesítmények elemzése során, de a tanulók írásainak kvalitatív vizsgálata a négy szempont (a feladat teljesítése, szókincs, nyelvhelyesség és szöveg kohézió) alapján újabb megvilágításba helyezheti mindezt. Ezért öt tipikus írást kiválasztottam a 15 újraértékelt közül, hogy mélységében megvizsgáljam, és az értékelésekkel összevethessem őket. Ezzel a kvantitatív és kvalitatív mérőeszközökkel mért eredmények összehasonlítása is megvalósult, valamint fény derült a saját mérőskálánk és a *KER*-skála szintezése közötti különbségekre is.

Az öt kiválasztott írás markánsan eltérő nyelvi szinteket képviselt: kiemelkedő; átlag feletti; a minimum követelményt kissé meghaladó; a minimum szinttől kissé elmaradó és erősen alulteljesítő színvonalat. Ezek a szövegek az egész kohortra jellemző tanulói köztes nyelv (interlanguage) számos megnyilvánulását példázzák.

A kiemelkedő írás nyelvi szintje magasan meghaladja a tantervi követelményszintet. Az értékelők a négy szempont szerint kétharmad részben a *KER* B.1-es szintjébe sorolták, egyharmadban pedig A2-esre helyezték. Ez az eredmény bizonyítja, hogy az A1-A2-es szintre tervezett feladaton lehetséges magasabb nyelvi szintet teljesíteni. Ez fontos eredménye a kutatásnak.

Az átlag feletti szintű szöveg rangsorolásában mutatkozott a legnagyobb véleménykülönbség a vizsgáztatók között; öten ötféle helyre sorolták be. Nyolc szempont szerint B1-es nyelvi szintre tették a pontozók, 11-szer A2-re és egyszer A1-re, melyek jelzik, hogy a *NAT* követelményszintje fölötti írásról van szó. A következő, a követelményeket kissé meghaladó tanulói szöveget még mindig 57%-ban az A1-es minimumszint fölé sorolták az értékelők.

A saját értékelő skálánk szerint minimum szint alatti szöveg 45%-ban kapott A1 szint feletti besorolást, bár ez magasnak tűnik az írásban előforduló számos nyelvi hiba tükrében. Ez az eredmény kétféle kétséget ébreszthet: egyik a *KER*-szintek vizsgáztatók általi félreértelmezése, mely okozhatja a magasabb kategóriába osztályozást. Másik magyarázata

annak, hogy az írás az egyik skála szerint A1 alatti nyelvi teljesítményt képvisel, a másik skála szerint annál magasabbat, az lehet, hogy a két skálán bemért A1 szint nem azonos egymással. Sok tennivaló akad még a *KER* szintjeinek értelmezésében, hiszen az iránymutató magyar oktatási dokumentumok szintleírásai sem egyeznek egymással az A1 és A2 szint bemutatásakor (*Kerettanterv*, 2000; *NAT*, 2003).

A köztes nyelv megnyilvánulási formái az öt szövegben jól képviselték az egész korpuszra jellemző nyelvelsajátítási sorrendet: a négy vizsgált nyelvi forma helytelenül használt alakjai jelezték a tanulói köztes nyelv átmeneti vonásait. Az EXP kizárólag a kiemelkedő írásban haladta meg az elsajátítás szintjét (75%). Ez alátámasztja a megállapítást, hogy a EXP az utolsóként elsajátított forma a négy elemzett nyelvi jelenség közül. A PRC a három leggyengébb szövegben az elsajátítás kezdeti szakaszát képviselte, míg a PLU-t és a NEG szerkezetet helyesen használták a diákok.

*12. Milyen szókincs jellemzi a 8. osztályosokat? Milyen összefüggés van a kritérium-orientált és korpusz alapú nyelvhelyesség- és szókincsmérés között?*

Ez a kérdés képviseli a kutatás összetettségét azáltal, hogy a nyelvhasználat és az értékelés közti összefüggéseket együtt vizsgálja.

A szókincs gazdagságát és gyakoriságát a Range teszt (Nation, 2005) alkalmazásával kutattam. Egy szöveg szógazdagságának a mérésére legáltalánosabban használt módszer az adott szöveg megalkotására használt lexémák (ragok és jelek nélküli, „szótári" szó = típus) számának és a szöveg szóelőfordulásainak (jel) az összevetése, a típus-jel (type-token) arány. Ezt *típus/jel viszonynak* nevezzük, melynek értéke maximálisan 1,00 — akkor ennyi, ha egy szövegben egyetlen szó sem ismétlődik. Egy szöveg hosszának növekedésével az ismétlődés esélye is növekszik, így a típus/jel érték a szöveg hosszával csökken. Az előbbiekből világosan következik, hogy a különböző hosszúságú szövegek típus-jel értékei nem mérhetők össze.

A szókincs gazdagságát kétféle típus-jel viszony kiszámításával, valamint a lexikai sűrűség és szofisztikáltság meghatározásával igyekeztem számszerűsíteni. A lexikai sűrűség a grammatikai szavak számának és a tartalmas szavak számának egybevetéséből áll, a lexikai szofisztikáltságot a leggyakrabban használt 1000 szót tartalmazó listából használt szavak és a kevésbé gyakori szavak szólistáiból használt szavak arányának kiszámításával kapjuk meg. Ezek a számítások és a gyakorisági adatok hasznos információt nyújtottak a tanulók szókincséről és bizonyos lexémák túl gyakori alkalmazásáról.

A szókincsgazdagság elemzése során bebizonyosodott Read (2000) állítása, miszerint a szövegek hossza nagy mértekben befolyásolja az eredményeket. Az elemzés azt is világossá tette, hogy a képek összehasonlítását igénylő feladat nem volt alkalmas a szövegek lexikai tulajdonságainak számszerűsítésére, mivel a lexikai mutatók nem adtak értelmezhető információt az írások minőségéről. A kiváló színvonalú és gyenge írások mutatói szinte azonosak voltak, néhány esetben a gyengébb szövegek szókincsmutatói voltak magasabbak, ami azt sugallja, hogy az írásokat nem minősíthetjük kizárólag a típus/jel értékekre támaszkodva. Az előzőekhez hasonlóan, a lexikai sűrűség mérőszámát is erősen befolyásolja a szöveg hossza.

Az írások szókincsének jellemzőit vizsgálva azt találtam, hogy a minőségileg jobb szövegek ugyan hosszabbak is voltak, de a lényeges különbség a jól és gyengén teljesítő diákok között nem szövegeik hosszában, hanem a szóhasználatuk megfelelőségi szintjében és helyességében mutatkozott meg.

Ezek a megállapítások arra figyelmeztetnek, hogy amennyiben a szövegek lexikai minőségét megbízhatóan kívánjuk elemezni a lexikai arányszámokkal, a feladatteljesítés minimum szöveghosszát meg kell határoznunk és az ennél rövidebb szövegeket ki kell zárnunk a vizsgálódás köréből.

Az értékelés következetességét, a kritérium-orientált döntések kapcsolatát a korpusz-alapú értékelésekkel, valamint a holisztikus és analitikus véleményalkotások közötti összefüggést az ötös vizsgáztató három alkalommal nyújtott besorolásainak segítségével ellenőriztem. A megbízhatósági indexek szoros kapcsolatot igazoltak a különböző értékelési eljárások eredményei között. Az első értékelés alkalmával adott pontszámok erősen korreláltak a 15 írás rangsorával és a második, *KER*-kritériumok szerinti véleményekkel. A három különböző értékelés magas korrelációja megnyugtatóan bizonyítja, hogy a vizsgáztató belső 'rejtett' kritériumrendszere összhangban áll önmagával, bármilyen skála alapján kell a döntéseket meghozni. A pontozó néhány következetlen döntése - saját írásbeli visszajelzése és a kvalitatív elemzés alapján - a *KER*-skála deskriptorainak félreértését jelzi.

## *5.7 A kutatás korlátai*

A magyar általános iskola utolsó évfolyamán tanuló diákok angol nyelvi íráskészségét többféle módszerrel vizsgáltam, és a kutatás fókuszában az írás produktuma állt. Az írás folyamatát, annak kognitív és affektív tényezőit nem kutattam, tehát az ok-okozati összefüggések feltárására csak közvetve és óvatosan vállalkozhattam. A háttértényezők közül

kizárólag a tanulók szociokulturális hátterét vizsgáltam intézményi megközelítésben. Ezért a diákok íráskészségének kialakulásáról, az azt befolyásoló tényezőkről nem tudok következtetéseket levonni, ez nem is volt célom.

Azonban az angol nyelv négy, a tanulói szövegekben gyakran előforduló morfémájának elsajátítási szintjét és azok elsajátításának sorrendjét bemutattam az egész korpuszra vonatkozóan, valamint néhány tanuló írásában kvalitatívan is elemeztem azt. Mivel osztálytermi megfigyeléseket nem végeztem, az adatok nem szolgáltattak információt arról, hogy egy adott nyelvi forma megjelenésének hiányát a szövegben mi okozza: az input hiányossága vagy a nyelvi képességek alacsony szintje.

## *5.8 A kutatás jövőbeli kiterjesztése*

Az eredmények egyik pedagógiai implikációja, hogy a nyelvelsajátítás folyamatának mélyebb ismeretében az idegennyelv-tanítás során ehhez alkalmazkodó tanmenetet, feladatokat és mérőeszközöket tervezünk. Például a nyelvi formákat a morfémák természetes elsajátítási rendje szerint, azzal összhangban érdemes bevezetni. Eredményeim azt sugallják, hogy a létezés kifejezését indokolt később tanítani annál, ami a tankönyvek többségében megszokott (az idegen nyelv tanulásának első félévében). Osztálytermi megfigyelésekkel pontos adatokat lehetne gyűjteni a tanítási módszerekről, és az egyes nyelvi jelenségek előfordulási arányáról.

Az értékeléssel kapcsolatban beigazolódott Creswell (2003) állítása, mely szerint a kvantitatív és kvalitatív mérések kiegészítik egymást, az egyik az általános jellemzőket, tendenciákat tárja fel, míg a másik a részletekbe enged betekintést. Vizsgálatomban a kvalitatív elemzés során tapasztaltam, hogy lényegesen különböző nyelvi szinteket ugyanazok a lexikai típus-jel arányok jellemeztek, tehát a számarányok önmagukban hamis képet festettek a tanulók íráskészségéről.

A *KER*-skálák értelmezésével kapcsolatban bizonyossá vált, hogy a tapasztalt értékelőknek is szükségük van ezek közös átgondolására, és standardizálására. A magyar oktatási alapdokumentumok (*NAT,* 2003; *Kerettanterv*, 2000) ellentmondó nyelvi szintleírásaira is fény derült, ezek követelményszintjét is szükséges lenne egyeztetni.

A nyelvtanárok számára továbbképzéseket kellene szervezni, hogy az értékelés alapelveit pontosabban megismerjék, tudatosuljanak bennük a *KER*-skála kommunikatív nyelvszemléleten alapuló követelményei, és ennek segítségével a mindennapokban elérendő nyelvi célok.

A kutatás során létrehozott korpuszt több módon lehet felhasználni további elemzésre:

A szövegeket más nyelvi formák megfigyelésére lehet kódolni, a szókincset meg lehet vizsgálni a gyakoriság szemszögéből, valamint a tanulók szövegalkotó stratégiáit is hasznos lenne elemezni.

Az írás folyamatát is érdemes nyomon követni hangos gondolkodás alkalmazásával. Egy hosszanti vizsgálat még megbízhatóbb adatokat biztosítana a fiatal nyelvtanulók nyelvfejlődési szakaszairól. Jelen értekezés jó kiindulópontként szolgálhat ezekhez a kutatásokhoz.

# Appendices

**Appendix A**

## Assessment Criteria for Writing Task: Year 8 English

| | **Task achievement** | **Vocabulary** | Grammar/accuracy | **Text** |
|---|---|---|---|---|
| **7-8** | Text is on 9 or 10 things relevant to pictures A & B. Text is about both A and B. | Rich scale & good choice of vocabulary, appropriate to task. | Whole text comprehensible; a few grammar or spelling mistakes do not interfere with comprehension. | Text is well structured: parts on different things are separated. Sentences are logically linked. There are some complex sentences. More than 3 sentences types vary. |
| **5-6** | Text is on 7 or 8 things relevant to pictures, or on 9 or 10 things, partly relevant to pictures. Text is about both A and B. | Wide scale & choice of vocabulary, mostly appropriate to task. | Some mistakes occur, but the whole text is comprehensible. | There are some links between sentences, but the text is less structured. Minimum three sentence types vary (e.g. stating existence, positive, negative statement, actions). |
| **3-4** | **Text is on 5 or 6 things relevant to pictures, or on 7 or 8 partly relevant things. Text is on both A and B.** | **Good scale or choice of vocabulary, mostly appropriate to task.** | **Several mistakes occur, but most of the text is comprehensible.** | **Text consists of sequence of sentences. One or two sentence types are repeated.** |
| **1-2** | Text is on 2 or 3 things relevant to pictures or on more but only partly relevant. Text is on either A or B. | Limited scale & choice of vocabulary or often inappropriate. | Many mistakes occur, only part of the text is comprehensible. | Same sentence type is repeated. |
| **0** | No text or a few words or sentences irrelevant to pictures. Handwriting illegible. Text not related to task. | Very limited scale & inappropriate vocabulary. | Text is incomprehensible because of grammatical mistakes and/or spelling mistakes. | Text unstructured, incomprehensible. |

- In each box two scores can be given to allow for differences, except for 0. If the criterion is fulfilled, the higher score is to be given. Maximum Score: 32, 8 for each criterion.
- **The four scores are to be put on the cover of the booklet, one below the other, to make data entry convenient. No need to add them up. In case there is no text, the only one score on the cover is: 9.**
- Length of text is not included in the boxes. A minimum of 20 simple or 10 complex sentences are necessary to achieve good scores.
- In case Task achievement is 0, no need to go on with assessment. This may be the case if text was rotelearnt and inappropriate to task.
- In case text is structured in two columns below pictures, the same criteria apply.
- Start assessment in middle boxes (printed in bold); if fulfilled go upwards, if not downwards.
- Repeated mistakes are considered just once.

**Appendix B: Raters' task sheet**

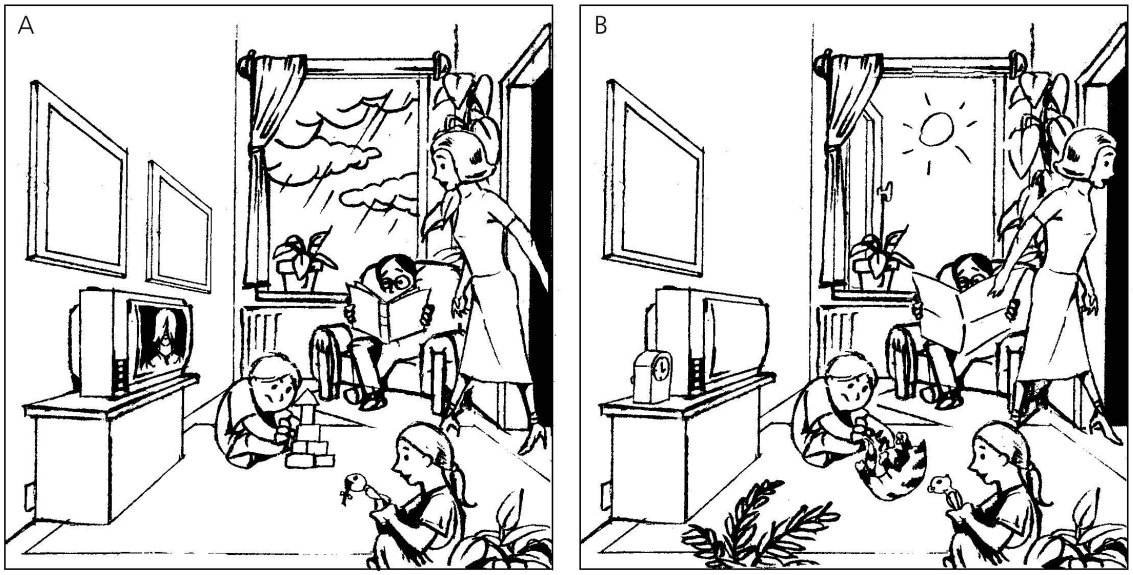
Dear Colleague,
You are kindly requested to spend a few hours of your precious time if you are ready to help me with my research. All data gained from your assistance will be handled anonymously and your name will be mentioned in the acknowledgement of the dissertation.

In this research my aim is to gain insights into the EFL writing proficiency of primary school-leavers. They wrote a writing task similar to ones published in widely used course books for this age group. These writings have been assessed and I chose 15 typical performances from a large number of papers.
Your task is to (1) rank order 15 students' writings, then (2) match the 15 texts with descriptors taken from the Common European Framework of Reference *(CEFR,* 2001).

The writing task was this: **Look at pictures A and B. Write about the differences between them. Write about the boy and the girl, the man, the woman, the TV, the clock, the wall, the plants, the window and the weather outside.**

---

You are provided with 15 writings of different proficiency levels produced by 8th-grade Hungarian students at




different schools and of various sociocultural backgrounds in Baranya. These 15 texts represent typical performances of the 231 students who participated in the research.

Task 1:
Please, rank order the 15 writings according to their level of English language proficiency using holistic assessment. You do not need to rely on an assessment scale. Simply put in the following boxes the code letters of the writings (E-Z) so that in the left box you put the best and in the right box you put the code of the poorest writing.

Best Poorest

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | | | | | |

Task 2:
You are given a grid with randomly placed descriptors representing levels A1-B1 in *CEFR*. The descriptors are related to following criteria of writing:

a) general linguistic range
b) overall written production
c) vocabulary range
d) grammatical accuracy

Write the code letter of the student's text in all boxes in which the descriptor (described language proficiency) characterises the student's writing most. Optimally, each test can be related to 4 descriptors.

Task 3:
3.1 Please answer the following question:
How well do you know the *CEFR*?
a) very well b) partly c) not at all

3.2 Please write your comments on the process of assessment:

Appendix C: *Task 2: This is the grid of randomly ordered descriptors in CEFR for describing levels of language proficiency. Please tick as many boxes for each writing (E-Z) as you think appropriate. Leave other boxes empty.*

| Descriptor / student's code | **E** | **F** | **G** | **H** | **K** | **L** | **M** | **N** | **O** | **P** | **R** | **S** | **T** | **V** | **Z** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **General linguistic range** | | | | | | | | | | | | | | | |
| Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words. | | | | | | | | | | | | | | | |
| Has a very basic range of simple expressions about personal details and needs of a concrete type. | | | | | | | | | | | | | | | |
| Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interest, work, travel, and current events, but lexical limitations cause repetitions and even problems with formulation at times. | | | | | | | | | | | | | | | |
| Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information.<br>Can use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulae about themselves and other people, what they do, possessions etc.<br>Has a limited repertoire of short memorised phrases covering predictable survival situations. | | | | | | | | | | | | | | | |
| Has sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films. | | | | | | | | | | | | | | | |
| **Overall written production** | | | | | | | | | | | | | | | |
| Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but', and 'because'. | | | | | | | | | | | | | | | |
| Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. | | | | | | | | | | | | | | | |
| Can write simple isolated phrases and sentences. | | | | | | | | | | | | | | | |

| **Appendix C: page 2** | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Vocabulary range** | **E** | **F** | **G** | **H** | **K** | **L** | **M** | **N** | **O** | **P** | **R** | **S** | **T** | **V** | **Z** |
| Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations. | | | | | | | | | | | | | | | |
| Has a sufficient vocabulary for<br>c) The expression of basic communicative needs.<br>d) Coping with simple survival needs. | | | | | | | | | | | | | | | |
| Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events. | | | | | | | | | | | | | | | |
| Has a sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics. | | | | | | | | | | | | | | | |
| **Grammatical accuracy** | | | | | | | | | | | | | | | |
| Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is clear what he/she is trying to express. | | | | | | | | | | | | | | | |
| Uses simple structures correctly, but still systematically makes basic mistakes – for example tends to mix up tenses and forget to mark agreement; nevertheless, it is usually clear what he/she is trying to say. | | | | | | | | | | | | | | | |
| Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire. | | | | | | | | | | | | | | | |
| Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations. | | | | | | | | | | | | | | | |

Appendix D: Descriptors of *CEFR* for describing levels of language proficiency

| CEFR level | **General linguistic range** |
|---|---|
| B1.2 | Has sufficient range of language to describe unpredictable situations, explain the main points in an idea or problem with reasonable precision and express thoughts on abstract or cultural topics such as music and films. |
| B1.1 | Has enough language to get by, with sufficient vocabulary to express him/herself with some circumlocutions on topics such as family, hobbies and interest, work, travel, and current events, but lexical limitations cause repetitions and even problems with formulation at times. |
| A2.2 | Has a repertoire of basic language which enables him/her to deal with everyday situations with predictable content, though he/she will generally have to compromise the message and search for words. |
| A2.1 | Can produce brief everyday expressions in order to satisfy simple needs of a concrete type: personal details, daily routines, wants and needs, requests for information.<br>Can use basic sentence patterns and communicate with memorised phrases, groups of a few words and formulae about themselves and other people, what they do, possessions etc.<br>Has a limited repertoire of short memorised phrases covering predictable survival situations. |
| A1 | Has a very basic range of simple expressions about personal details and needs of a concrete type. |
| CEFR level | **Overall written production** |
| B1 | Can write straightforward connected texts on a range of familiar subjects within his field of interest, by linking a series of shorter discrete elements into a linear sequence. |
| A2 | Can write a series of simple phrases and sentences linked with simple connectors like 'and', 'but', and 'because'. |
| A1 | Can write simple isolated phrases and sentences. |
| CEFR level | **Vocabulary range** |
| B1 | Has a sufficient vocabulary to express him/herself with some circumlocutions on most topics pertinent to his/her everyday life such as family, hobbies and interests, work, travel, and current events. |
| A2.2 | Has a sufficient vocabulary to conduct routine, everyday transactions involving familiar situations and topics. |
| A2.1 | Has a sufficient vocabulary for<br>a) The expression of basic communicative needs.<br>b) Coping with simple survival needs. |
| A1 | Has a basic vocabulary repertoire of isolated words and phrases related to particular concrete situations. |
| CEFR level | **Grammatical accuracy** |
| B1.2 | Communicates with reasonable accuracy in familiar contexts; generally good control though with noticeable mother tongue influence. Errors occur, but it is clear what he/she is trying to express. |
| B1.1 | Uses reasonably accurately a repertoire of frequently used 'routines' and patterns associated with more predictable situations. |
| A2 | Uses simple structures correctly, but still systematically makes basic mistakes – for example tends to mix up tenses and forget to mark agreement; nevertheless, it is usually clear what he/she is trying to say. |
| A1 | Shows only limited control of a few simple grammatical structures and sentence patterns in a learnt repertoire. |

**Appendix E: Tagged corpus in Excel table** (on attached CD)

**Appendix F: Sample for a low achievement**

Script Z; 0 points; the length of the composition is 60 words

It's a sunny. Window is a flower. A man is a read newspaper. Man between the flower. Man is beehain woman. Woman on the boy on the girls. Boy is play cat. Girls is the play bear. Girls between flower. Girls beehain the flower. Flower is left desk. Desk is TV. TV right is clock. TV between is the photo.

**Appendix G: Script F corrected and run on Range test with the stop list on**

**Script F**

In picture A the boy is playing with his toy but in picture B he is playing with his cat.

In picture A the girl is playing with a doll but in picture B she is playing with a toy bear.

In picture A the man is reading a book but in the other picture he is reading a newspaper. In the first picture the woman is going in the room but in the other she is leaving it. In the first picture there are two pictures on the wall but in the other picture there is only one. The difference between the two TVs that in picture A it is switched on but on the other picture it is switched off. In picture A there is not a clock near the TV but in picture B there is. In picture A there is only one plant but in picture B there are two.

In picture A the window is closed but in picture B it is open. In the first picture the weather is not so good because it is raining but in picture B the sun is shining.

The stop list contained: *A, a, an, are, at, B, be, can, in, into, is, on, out of, the, there, to,* and *with.*

## Results:

Number of lines: 5
Number of words: 190

| WORD LIST | TOKENS/% | TYPES/% | FAMILIES |
|---|---|---|---|
| one | 91/91.92 | 42/85.71 | 39 |
| two | 6/ 6.06 | 5/10.20 | 5 |
| three | 1/ 1.01 | 1/ 2.04 | 1 |
| four | 1/ 1.01 | 1/ 2.04 | 1 |
| not in the lists | 0/ 0.00 | 0/ 0.00 | ??? |
| Total | 99 | 49 | 46 |

**Types Found In Base List One**

| TYPE | RANGE | FREQ | F1 |
|---|---|---|---|
| BEAR | 1 | 1 | 1 |
| BECAUSE | 1 | 1 | 1 |
| BETWEEN | 1 | 1 | 1 |
| BOOK | 1 | 1 | 1 |
| BOY | 1 | 1 | 1 |
| BUT | 1 | 10 | 10 |
| CAT | 1 | 1 | 1 |
| CLOCK | 1 | 1 | 1 |
| CLOSED | 1 | 1 | 1 |
| DIFFERENCE | 1 | 1 | 1 |
| FIRST | 1 | 3 | 3 |
| GIRL | 1 | 1 | 1 |
| GOING | 1 | 1 | 1 |
| GOOD | 1 | 1 | 1 |
| HE | 1 | 2 | 2 |
| HIS | 1 | 2 | 2 |
| IT | 1 | 5 | 5 |
| LEAVING | 1 | 1 | 1 |
| MAN | 1 | 1 | 1 |
| NEAR | 1 | 1 | 1 |
| NOT | 1 | 2 | 2 |
| OFF | 1 | 1 | 1 |

| | | | |
|---|---|---|---|
| ONE | 1 | 2 | 2 |
| ONLY | 1 | 2 | 2 |
| OPEN | 1 | 1 | 1 |
| OTHER | 1 | 4 | 4 |
| PICTURE | 1 | 19 | 19 |
| PICTURES | 1 | 1 | 1 |
| PLAYING | 1 | 4 | 4 |
| READING | 1 | 2 | 2 |
| ROOM | 1 | 1 | 1 |
| SHE | 1 | 2 | 2 |
| SO | 1 | 1 | 1 |
| SUN | 1 | 1 | 1 |
| SWITCHED | 1 | 2 | 2 |
| THAT | 1 | 1 | 1 |
| TV | 1 | 1 | 1 |
| TVS | 1 | 1 | 1 |
| TWO | 1 | 3 | 3 |
| WALL | 1 | 1 | 1 |
| WINDOW | 1 | 1 | 1 |
| WOMAN | 1 | 1 | 1 |

## Types Found In Base List Two

| TYPE | RANGE | FREQ | F1 |
|---|---|---|---|
| NEWSPAPER | 1 | 1 | 1 |
| PLANT | 1 | 1 | 1 |
| RAINING | 1 | 1 | 1 |
| TOY | 1 | 2 | 2 |
| WEATHER | 1 | 1 | 1 |

## Types Found In Base List Three

| TYPE | RANGE | FREQ | F1 |
|---|---|---|---|
| SHINING | 1 | 1 | 1 |

## Types Found In Base List Four

| TYPE | RANGE | FREQ | F1 |
|---|---|---|---|
| DOLL | 1 | 1 | 1 |

**LIST OF FAMILY GROUPS**

| BASE ONE FAMILIES | RANGE | TYFRQ | FAFRQ | F1 |
|---|---|---|---|---|
| BEAR | 1 | 1 | 1 | 1 |
| BECAUSE | 1 | 1 | 1 | 1 |
| BETWEEN | 1 | 1 | 1 | 1 |
| BOOK | 1 | 1 | 1 | 1 |
| BOY | 1 | 1 | 1 | 1 |
| BUT | 1 | 10 | 10 | 10 |
| CAT | 1 | 1 | 1 | 1 |
| CLOCK | 1 | 1 | 1 | 1 |
| CLOSES | 1 | 0 | 1 | 1 |
| DIFFERENCE | 1 | 1 | 1 | 1 |
| FIRST | 1 | 3 | 3 | 3 |
| GIRL | 1 | 1 | 1 | 1 |
| GO | 1 | 0 | 1 | 1 |
| GOOD | 1 | 1 | 1 | 1 |
| HE | 1 | 2 | 4 | 4 |
| IT | 1 | 5 | 5 | 5 |
| LEAVE | 1 | 0 | 1 | 1 |
| MAN | 1 | 1 | 1 | 1 |
| NEAR | 1 | 1 | 1 | 1 |
| NOT | 1 | 2 | 2 | 2 |
| OFF | 1 | 1 | 1 | 1 |
| ONE | 1 | 2 | 2 | 2 |

| | | | | |
|---|---|---|---|---|
| ONLY | 1 | 2 | 2 | 2 |
| OPEN | 1 | 1 | 1 | 1 |
| OTHER | 1 | 4 | 4 | 4 |
| PICTURE | 1 | 19 | 20 | 20 |
| PLAY | 1 | 0 | 4 | 4 |
| READ | 1 | 0 | 2 | 2 |
| ROOM | 1 | 1 | 1 | 1 |
| SHE | 1 | 2 | 2 | 2 |
| SO | 1 | 1 | 1 | 1 |
| SUN | 1 | 1 | 1 | 1 |
| SWITCH | 1 | 0 | 2 | 2 |
| TELEVISION | 1 | 0 | 2 | 2 |
| THIS | 1 | 0 | 1 | 1 |
| TWO | 1 | 3 | 3 | 3 |
| WALL | 1 | 1 | 1 | 1 |
| WINDOW | 1 | 1 | 1 | 1 |
| WOMAN | 1 | 1 | 1 | 1 |

| **BASE TWO FAMILIES** | RANGE | TYFRQ | FAFRQ | F1 |
|---|---|---|---|---|
| NEWSPAPER | 1 | 1 | 1 | 1 |
| PLANT | 1 | 1 | 1 | 1 |
| RAIN | 1 | 0 | 1 | 1 |
| TOY | 1 | 2 | 2 | 2 |
| WEATHER | 1 | 1 | 1 | 1 |

**BASE THREE FAMILIES**

| | | | | |
|---|---|---|---|---|
| SHINE | 1 | 0 | 1 | 1 |

**BASE FOUR FAMILIES**

| | | | | |
|---|---|---|---|---|
| DOLL | 1 | 1 | 1 | 1 |

## Types Not Found In Any List

| TYPE | RANGE | FREQ | F1 |
|---|---|---|---|

**Appendix H: Frequency test results of Script F**

Words in frequency order**:**

| Word Type | Rank | Frequency | Cumulative % |
|---|---|---|---|
| IN | 1 | 20 | 10.53 |
| IS | 2 | 19 | 20.53 |
| PICTURE | 3 | 19 | 30.53 |
| THE | 4 | 19 | 40.53 |
| A | 5 | 12 | 46.84 |
| BUT | 6 | 10 | 52.11 |
| B | 7 | 6 | 55.26 |
| THERE | 8 | 6 | 58.42 |
| IT | 9 | 5 | 61.05 |
| OTHER | 10 | 4 | 63.16 |
| PLAYING | 11 | 4 | 65.26 |
| WITH | 12 | 4 | 67.37 |
| FIRST | 13 | 3 | 68.95 |
| ON | 14 | 3 | 70.53 |
| TWO | 15 | 3 | 72.11 |
| ARE | 16 | 2 | 73.16 |
| HE | 17 | 2 | 74.21 |
| HIS | 18 | 2 | 75.26 |
| NOT | 19 | 2 | 76.32 |
| ONE | 20 | 2 | 77.37 |
| ONLY | 21 | 2 | 78.42 |
| READING | 22 | 2 | 79.47 |
| SHE | 23 | 2 | 80.53 |
| SWITCHED | 24 | 2 | 81.58 |
| TOY | 25 | 2 | 82.63 |
| BEAR | 26 | 1 | 83.16 |
| BECAUSE | 27 | 1 | 83.68 |

| | | | |
|---|---|---|---|
| BETWEEN | 28 | 1 | 84.21 |
| BOOK | 29 | 1 | 84.74 |
| BOY | 30 | 1 | 85.26 |
| CAT | 31 | 1 | 85.79 |
| CLOCK | 32 | 1 | 86.32 |
| CLOSED | 33 | 1 | 86.84 |
| DIFFERENCE | 34 | 1 | 87.37 |
| DOLL | 35 | 1 | 87.89 |
| GIRL | 36 | 1 | 88.42 |
| GOING | 37 | 1 | 88.95 |
| GOOD | 38 | 1 | 89.47 |
| LEAVING | 39 | 1 | 90.00 |
| MAN | 40 | 1 | 90.53 |
| NEAR | 41 | 1 | 91.05 |
| NEWSPAPER | 42 | 1 | 91.58 |
| OFF | 43 | 1 | 92.11 |
| OPEN | 44 | 1 | 92.63 |
| PICTURES | 45 | 1 | 93.16 |
| PLANT | 46 | 1 | 93.68 |
| RAINING | 47 | 1 | 94.21 |
| ROOM | 48 | 1 | 94.74 |
| SHINING | 49 | 1 | 95.26 |
| SO | 50 | 1 | 95.79 |
| SUN | 51 | 1 | 96.32 |
| THAT | 53 | 1 | 97.37 |
| TVS | 54 | 1 | 97.89 |
| WALL | 55 | 1 | 98.42 |
| WEATHER | 56 | 1 | 98.95 |
| WINDOW | 57 | 1 | 99.47 |
| WOMAN | 52 | 1 | 96.84 |
| TV | 58 | 1 | 100.00 |

# Name index

## D



## E



## F



## G



## H

## I



## J



## K

## L



## M

## N



## O



## P



## R



## S

## SZ



## T



## U



## V

## W



## X



## Y



## Z