CONTENTS

INTRODUCTION

This thesis is focused on reading skills testing (in second language learning). It is divided in two main parts: theoretical and practical. The theoretical part is further lie out into three main chapters. They are: testing, reading and testing reading. The practical part is devoted to a research focused on techniques used for testing reading from the pupils´ point of view. The objective of the paper is to try to evaluate the difficulties with reading skills testing techniques and compare them with pupils´ test results. The issue was chosen on the basis of either little or bad experience with teaching and testing reading in English lessons at elementary schools. (It is discussed in more details at the very beginning of the practical part.)

In the first chapter of the theoretical part called *Testing* the relationship between teaching or learning and testing is emphasized. The importance of testing reading skills in teaching English as a second language is also brought up. Moreover, the issue of backwash, the effect of testing on teaching and learning, is discussed here. Test characteristics (such as validity, reliability and efficiency), types of tests, types of testing and stages of test construction are introduced here as well.

The following chapter is devoted to reading. Various definitions are used to get an insight into this skill and questions like “What do we read?, Why do we read? and “How do we read” are noted at this point. Different types of skills involved in reading are stated here together with an explanation of various reading techniques.

The final chapter of the theoretical part focuses on two main issues: the text as such (and the factors affecting the reading test items and test texts) and different testing techniques that can be used when checking reading comprehension. It states both advantages and disadvantages of each technique.

The practical part involves three chapters: Introduction, Stages of our test construction (including setting the purpose, writing specifications, writing the test, marking and pre-testing), Test analysis and finally the Conclusion. In the first chapter the focus of the research and the reason for choosing the topic for the thesis is explained. In the second chapter individual stages of the test construction are described. At this place testing techniques that were chosen for our research are presented in a very detailed way. The reasons for that choice are clearly stated.

In the third chapter the three tests are introduced and than their results analysed. We assessed the readers' success to deal with different test texts and test formats. Their successfulness is summed up in the last chapter of the practical part of the thesis.

At the very end of the work information gained from the theoretical part is then compared to the results of our research.

1. TESTING

1.1 Testing and teaching

As Madsen states at the very beginning of his book, "testing is an important part of every teaching and learning experience" (Madsen 1983:3). However, a large number of examinations in the past have encouraged a tendency to separate testing and teaching (Heaton 1988:5). Heaton argues that both testing and teaching are so closely interrelated that it is virtually impossible to work in either field without being constantly concerned with the other.

"The effect of testing on teaching and learning is called backwash." (Hughes 1989:1) It can be harmful or beneficial. He adds that we cannot expect testing only to follow teaching; what we should demand of it is that it should be supportive of good teaching and, where necessary, exert a corrective influence on bad teaching (Hughes 1989:2).

1.2 Reasons for testing

According to Madsen, testing helps not only the students but also to teachers. He mentions two ways how the well-made tests can help learners. First, he says, such tests can help create positive attitudes in terms of motivation and efficient instruction. This means that a sense of accomplishment should be taken into account. Madsen believes that tests of appropriate difficulty, announced well in advance and covering skills scheduled to be evaluated, can also contribute to a positive tone by demonstrating your spirit of fair play and consistency with course objectives (Madsen 1983:4). The second way that students can benefit from tests is by helping them to master the language. They can confirm what each person has mastered and they point up those language items that

need to be studied further. It can help learners to adjust their own personal goals (Madsen 1983:4).

As mentioned earlier in the text, testing also helps teachers. Due to testing they can be able to answer the important questions, such as:

- Have I been effective in my teaching?
- Are my lessons on the right level?
- Am I aiming my instruction too low or too high?
- Am I teaching some skills effectively but others less effectively,
- What areas need more work?
- Which points need reviewing?
- Should I spend more (or less) time on this material with next year's students?
- Were the test instructions clear?
- Was everyone able to finish in the allotted time?
- Did the test results reflect accurately how my students have been responding in class and in their assigned work?

(Madsen 1983:5)

In other words, testing can be used to diagnose both teachers' and students' effort. It can confirm progress that has been made and show how to redirect our future efforts. Madsen adds that good tests can sustain or enhance class morale and aid learning (Madsen 1983:5).

Heaton presents reasons for testing as follows:

- finding out about progress
- encouraging students
- finding out about learning difficulties
- finding out about achievement
- placing students
- selecting students
- finding out about proficiency

(Heaton 1990:9–18)

1.3 Test characteristics

1.3.1 Validity

As Heaton states, "the validity of a test is the extent to which it measures what it is supposed to measure and nothing else" (Heaton 1991:159). In other words, a test is said to be valid if it measures accurately what it is intended to measure. The concept of validity can be approached from a number of perspectives. The relationship between these is interpreted in a number of ways in literature (Weir 1990:22). Hughes and Weir

agree on content validity, criterion-related validity, construct validity and face validity; Weir adds washback validity. What he means by this term is the washback of the test on teaching and learning, which was already discussed earlier in this paper, and that is why we do not dwell on it again in this section. Hughes claims “a test is said to have content validity if its content constitutes a representative sample of the language skills, structures, etc. with which is meant to be concerned” (Hughes 2002:22). He also explains that another approach to test validity is to see how far results on test agree with those provided by some independent and highly dependable assessment of the candidate’s ability. He says that such independent assessment is then the criterion measure against which the test is validated (Hughes 2002:23). Hughes presents two kinds of criterion-related validity: concurrent and predictive. Concurrent validity is when the test scores are correlated with another measure of performance; usually an older established test, taken at the same time (Weir 1990:27). Predictive validity concerns the degree to which a test can predict a candidates’ future performance. “If a test has a construct validity, it is capable of measuring certain specific characteristics in accordance with a theory of language behaviour and learning“ (Heaton 1991:161). In Hughes’ words, the test has construct validity if it measures just the ability which it is supposed to measure (Hughes 2002:26).

Hughes claims “a test is said to have face validity if it *looks* as if it measures what it is supposed to measure (Hughes 2002:27). He gives an example of a test that pretends to measure pronunciation ability but does not require the testee to speak.

1.3.2 Reliability

As Heaton states, “reliability is a necessary characteristic of any good test: for it to be valid at all, a test must first be reliable as a measuring instrument” (Heaton 1991:162). Weir describes the concept of reliability as a fundamental criterion against which any language test has to be judged (Anastasi in Weir 1990: 31). He explains that the concern is “how far can we depend on the results that a test produces or, in other words, could the results be produced consistently” (Weir 1990:31).

Three aspects of reliability are usually taken into account. As Weir presents, the first aspect of reliability concerns the consistency of scoring among different markers. The second aspect refers to how to enhance the agreement between markers by establishing and maintaining adherence to, explicit guidelines for the conduct of

marking. The third aspect of reliability is “of parallel-forms reliability, the requirements of which have to be borne in mind when future alternative forms of a test have to be devised” (Weir 1990:32).

Hughes suggests ways how to make tests more reliable. They are:

- take enough samples of behavior
- do not allow candidates too much freedom
- write unambiguous items
- provide clear and explicit instructions
- ensure that tests are well laid out and perfectly legible
- candidates should be familiar with format and testing techniques
- provide uniform and non-distracting conditions of administration
- use items that permit scoring which is as objective as possible
- make comparisons between candidates as direct as possible
- provide a detailed scoring key
- train scorers
- agree acceptable responses and appropriate scores at outset of scoring
- identify candidates by number, not name
- employ multiple, independent scoring

(Hughes 2002:36-42)

1.3.3 Reliability versus validity

Hughes claims that there will always be some tension between reliability and validity (Hughes 2002:42). Valid test must provide consistently accurate measurements; it must therefore be reliable. A reliable test, however, may not be valid at all. Hughes shows that on an example:

> As writing test we might require candidates to write down the translation equivalents of 500 words in their own language. This could well be a reliable test; but it is unlikely to be a valid test of writing.
>
> (Hughes 2002:42)

Hughes concludes that we should be careful with reducing test validity in our efforts to make tests more reliable (Hughes 2002:42). Weir explains that this “inevitable tension exists in the sense that it is sometimes essential to sacrifice a degree of reliability in order to enhance validity” (Weir 1990:33). He claims “the two concepts

are, in certain circumstances, mutually exclusive, but if a choice has to be made, validity after all, is the more important“ (Weir 1990:33).

1.3.4 Test efficiency

Even valid and reliable test can be of little use when it is not a practical one. The term practicality here involves question of economy, ease of administration, scoring, and interpretation of results. Weir points out that “the longer it takes to construct, administer and score a test, and the more skilled personnel and equipment are involved, the higher the costs are likely to be” (Weir 1990:34). The duration of the test has to also be taken into consideration. Weir concludes, “there is clearly an imperative need to try and develop test formats and their evaluation criteria that provide the best overall balance among reliability, validity and efficiency in the assessment of communicative skills” (Weir 1990:34).

1.4 Types of tests

Madsen presents following test classification:

CONTRASTING CATEGORIES OF ESL TESTS

Knowledge tests……Performance (or Skills) tests

Subjective tests……. Objective tests

Productive tests…….Receptive tests

Language subskill tests…….Communication skills tests

Norm-referenced tests……Criterion-referenced tests

Discrete-point tests…….Integrative tests

Proficiency tests…….Achievement tests

(Madsen 1983:8)

Hughes, however, notes a difference between kinds of tests and kinds of testing. He distinguishes four types of tests (according to the use of the test results): proficiency tests, achievement tests, diagnostic tests and placement tests. To the contrary he describes distinctions between direct and indirect testing, between discrete point and integrative testing, between norm-referenced and criterion-referenced testing and finally between objective and subjective testing. We will follow his categorization in this paper.

Proficiency tests can according to Madsen measure overall mastery of language (Madsen 1983:9). In other words they show how the testee is prepared to use the

language. Hughes states, “proficiency tests are designed to measure people’s ability in a language regardless of any training they may have had in the language” (Hughes 2002:9). He further explains that the content of a proficiency test is not based on the content or objectives of language courses which people taking the test may have followed; rather it is based on a specification of what candidates must be able to do in the language to be considered proficient (Hughes 2002:9).

Achievement tests, on the other hand, measure progress or development in mastering particular skills. In contrast to proficiency tests, achievement tests are linked directly to particular courses and to the achievement of their objectives. There are two kinds of achievement tests: final and progress. Final achievement tests are administered at the end of a course of study by ministries of education, official examining boards, or by members of teaching institutions. Progress achievement tests are intended to measure the progress that students are making. (Hughes 2002:11-12)

Diagnostic tests are constructed to show students’ strengths and weaknesses. Hughes adds that diagnostic tests are intended primarily to ascertain what further teaching is necessary (Hughes 2002:12).

As the name suggests, placement tests provide information which will help to place students at the stage or in the part of the teaching programme most appropriate to their abilities (Hughes 2002: 14). Most often they are used to rank students to classes or courses at different levels. Hughes remarks that it is possible to buy placements tests, however he does not recommend it:

> The placement tests that are most successful are those constructed for particular situations. They depend on the identification of the key features at different levels of teaching in the institution. They are tailor-made rather than bought of the peg. This usually means that they have been produced “in house”. The work that goes into their construction is rewarded by the saving in time and effort through accurate placement.
>
> (Hughes 2002: 14)

1.5 Types of testing

Distinguishing types of testing means distinguishing between two approaches to test construction. They are: Direct versus indirect testing, Discrete point versus integrative testing, Norm-referenced versus criterion-referenced testing and Objective testing versus subjective testing.

"Testing is said to be direct when it requires the candidate to perform precisely the skill which we wish to measure" (Hughes 2002:15). Hughes gives following examples:

> If we want to know how well candidates can write compositions, we get them to write compositions. If we want to know how well they pronounce a language, we get them to speak.
>
> (Hughes 2002:15)

Even though the test situation does not allow the tasks to be really authentic, teachers should try to find and use tests that are as authentic as possible. Hughes points out the problem of direct testing of receptive skills such as reading and listening. He comments on it: "with listening and reading, it is necessary to get candidates not only to listen or read but also to demonstrate that they have done this successfully" (Hughes 2002:15). We will look closer at this problem in the part called Testing Reading.

While direct testing intends the candidate to perform precisely the skill, which we wish to measure, indirect testing attempts to measure the abilities that underlie the skills, in which we are interested (Hughes 2002:15).

As Hughes claims, discrete point testing refers to the testing of one element at a time, item by item (Hughes 1989:16). That could be a series of items each testing a particular grammatical structure. Integrative testing, by comparison, requires the candidate to combine many language elements. That could be used in writing a composition, making notes while listening to a lecture, taking a dictation, or completing a cloze passage. Hughes also points out the relation within direct and indirect testing; he says that discrete point tests will almost always be direct, while integrative testing methods, such as the cloze procedure, are indirect (Hughes 2002:17).

Concerning validity and reliability, Harris asserts that discrete item tests support high reliability (Harris et al. 1994:34). Such formats include short answers only, so there can be more of them and that is why reliability is increased. Nonetheless, there are certain disadvantages too, for instance multiple-choice tests cannot be considered as real communication tests, so the validity is quite low (Harris et al. 1994:34). Harris concludes that both discrete item and integrative test formats have their advantages and disadvantages and suggests to mix them both and use integrative tasks especially for testing productive skills and discrete item tasks for testing receptive skills (Harris et al. 1994:35).

Another set of contrasting tests is that of norm-referenced and criterion-referenced exams. Madsen explains these two types of testing as follows: “Norm-referenced tests compare each student with his classmates, but criterion-referenced exams rate students against certain standards, regardless of how other students do” (Madsen 1983:9).

Hughes adds that in the case of norm-referenced tests we cannot say directly what the student is able to do in the language. Criterion-referenced tests are designed to do so; they provide the information about what the student can actually do in the language (Hughes 2002: 18). Hughes sums up that criterion-referenced tests have two positive merits:

1) they set standards meaningful in terms of what people can do, which do not change with different groups of candidates
2) they motivate students to attain those standards

(Hughes 2002:18)

A final classification of types of testing is objective and subjective testing. According to Madsen’s opinion, subjective tests, like translation or essay, have the advantage of measuring language skill naturally, almost the way English is used in real life (Madsen 1983:8). Many English teachers, however, cannot score such tests quickly and consistently. On the other hand, objective tests can be scored very quickly and consistently. Hughes claims that the only distinction between these two methods is in scoring and nothing else (Hughes 2002:19). He explains that if no judgement is required on the part of scorer, the scoring is objective, while if judgement is called for, the scoring is said to be subjective. Hughes remarks that there are different degrees of subjectivity in testing: “The impressionistic scoring of a composition may be considered more subjective than the scoring of short answers in response to questions on a reading passage” (Hughes 2002:19). To conclude Hughes states that many testers seek after objectivity in scoring for it brings greater reliability.

Heaton looks at the problem from a different point of view; he points out that objective tests are often criticized because they are said to be simpler to answer than subjective tests. However, he claims, items in an objective test can be made just as easy or difficult as the test constructor wishes (Heaton 1991:26). Heaton disputes the fact that objective tests *are* easier only because they may generally *look* easier. Hughes

suggests a way in which the test constructor can calculate the approximate degree of difficulty of the test:

> Objective tests can be pre-tested before being administered on a wider basis … Standards may then be compared not only between students from different areas of schools but also between students taking the test in different years.
>
> (Heaton 1991:26)

Another criticism Heaton discusses is that objective tests of the multiple-choice type encourage guessing. However, he says, if four or five alternatives for each item are offered, it sufficiently reduces the possibility of guessing (Heaton 1991:26).

In despite of earlier mentioned disadvantages of objective tests, Heaton concludes that good objective tests may be useful - "provided that such tests are never regarded as measures of the students' ability to communicate in the language (Heaton 1991:27). Heaton describes a very poor objective test as a test where items are poorly written, where irrelevant areas and skills are emphasized in the test simply because they are "testable" and when it is confined to language-based usage and neglects the communicative skills involved (Heaton 1991:27).

To sum up, Heaton declares:

> It should never be claimed that objective tests can do those tasks which they are not intended to do… They can never test the ability to communicate in the target language, nor can they evaluate actual performance. A good classroom test will usually contain both subjective and objective test items.
>
> (Heaton 1991:27)

1.6 Stages of test construction

Weir presents four stages in the development of a test which are presently accepted as the "best practice". They are: test design, test development, operation and monitoring (Weir 1990:36-41). To the contrary, Hughes distinguishes the stages of test construction as follows:

1.6.1 Setting the purpose

Hughes believes that the essential first step in testing is to make perfectly clear what a teacher wants to find out and for what purpose. He points out that it is necessary to answer the following questions:

- What kind of test is it to be? Achievement (final or progress), proficiency, diagnostic, or placement?
- What is its precise purpose?
- What abilities are to be tested?
- How detailed must the results be?
- How accurate must the results be?

- How important is backwash?
- What constraints are set by unavailability of expertise, facilities, time (for construction, administration and scoring)?

(Hughes 2002:48)

1.6.2 Writing specifications

When the first step is done, then next steps can follow. Hughes recommends starting with writing a set of specifications for the test. That includes information on content, format and timing, criterial levels of performance, and scoring procedures (Hughes 2002:48).

1.6.2.1 Content

Firstly, as Hughes states, the content of a test should include samples for not a single version of a test but for more versions. Samples of such potential content will then appear in individual versions of the test. Hughes believes that: ”the fuller the information on content, the less arbitrary should be the subsequent decisions as to what to include in the writing of any version of the test” (Hughes 2002:49). He warns that in the desire to be highly specific, one may go beyond the current understanding of what the components of language ability are and what their relationship is to each other. He assumes that the best choice would be to include in the content specifications only those elements whose contribution is well established (Hughes 2002:49).

1.6.2.2 Format and timing

Secondly, format and timing should specify test structure and item types or elicitation procedures, with examples. It also should say how each component would be evaluated; in case of reading how many passages will be presented and how many items there will be in each component (Hughes 2002:50). Alderson states that some items of a test may be more important than others. Such items should, therefore, carry more weight. This process is called weighting. The easiest method of weighting is, however, to give the same “weigh” to each item (Alderson et. al. 1995:149).

1.6.2.3 Criterial levels of performance

Thirdly, the required level of performance for success should be defined. It is necessary to determine what performance is satisfactory and what is not. Alderson believes that it is appropriate to set a pass mark in fixed percentage (Alderson et

al.1995:155). Hughes states that this could involve a simple statement (for instance: to mastery, 80 per cent of the items must be responded to correctly) or it could be more complex. He gives an example of a test evaluating the students' performance from different points of view; the test assesses accuracy, appropriacy, range, flexibility, and size (Hughes 2002:50-51).

1.6.2.4 Scoring procedures

Finally, Hughes notes that the test constructors should know precisely how high scorer reliability could be achieved (Hughes 2002:51). The question of reliability was discussed earlier.

1.6.3 Writing the test

Writing the test includes three areas: sampling, item writing and moderation, and writing and moderation of scoring key.

First area is to sample. Making choices of samples is inevitable at this point. "For content validity and beneficial backwash, the important thing is to choose widely from the whole area of content" (Hughes 2002:51). We should avoid choosing those elements that are only easy to test. In other words, the test should sample widely and unpredictably.

As for the item writing and moderation, "the writing of successful items (in the broadest sense, including, for example, the setting of writing tasks) is extremely difficult" (Hughes 2002:51). Therefore, it is highly advisable to cooperate with other colleagues. Hughes claims that teamwork is essential at this stage.

> Colleagues must really try to find fault; and despite the seemingly inevitable emotional attachment that item writers develop to items that they have created, they must be open to, and ready to accept, the criticisms that are offered to them. Good personal relations are a desirable quality in any test writing team.
>
> (Hughes 2002:51)

Here are some critical questions Hughes believes should be asked:

- Is the task perfectly clear?
- Is there more than one possible correct response?
- Can candidates show the desired behavior (or arrive at the correct response) without having the skill supposedly being tested?
- Do candidates have enough time to perform the task(s)?

(Hughes 2002:51-52)

The described process is called moderation. In addition to that it is appropriate to try to administer the test to native speakers. They should score 100 percent, or close to it. Otherwise the items that proved to be too difficult should be revised or replaced.

The next step, after choosing the right items, is to write the scoring key. Sometimes only one correct response is possible; however in some cases there may appear some alternative acceptable responses. In such case it is necessary to decide how these would be awarded. Again, help of colleagues is welcomed here (Hughes 2002:52).

1.6.4 Pretesting

To complete test construction pretesting should be done. Even though it has proved to be very helpful, it may not be possible to practise it every time. The aim of pretesting is to identify possible problems of the test. (Some may occur even after careful moderation.) That is why the test should be first tested on a different but similar group of those testees for whom the test is intended. In spite of indisputable advantages of pretesting, it is not always feasible. For instance, a suitable group for pretesting may not be available, or the security of the test might be put at risk (Hughes 2002:52).

1.6.5 Marking

As Harris claims, it is essential to think of marking at the very beginning of our test construction. We should have already made a decision about marking when we are choosing the most suitable test technique. Otherwise, as Harris adds, it can be the most time-consuming stage of our test construction. Objective tests are usually marked very quickly and easily. Marking subjective tests, however, can be much more difficult and can take a longer time. That is why Harris suggests using rating scales where scoring is precisely described.

Alderson agrees with Harris's differentiation of two types of marking: objective and subjective marking. Objective marking is used when there is a clear difference between right or wrong response. As an example we can mention testing formats as multiple-choice items or true / false statements. In other words, an examiner compares the test-taker's answers to answers in a *key* or *mark scheme* As Alderson explains, the term *key* is used when there is only one correct answer for each item, while the term mark scheme is used when there is more than one possible response for an item (Alderson et al. in Potůčková 2003:41).

To mark tests of speaking and writing subjective marking is usually used. Examiners evaluate the student's performance according to a rating scale. As Alderson et al. states, there are two types of these rating scales: a holistic scale and analytic scale. The holistic scale, which is sometimes called an impression scale, is implied when we want to judge a student's performance as a whole. The analytic scale, on the contrary, is used when more components are taken into account and assessed separately (Alderson in Potůčková 2003:41).

A last note on scoring that is worth pointing out refers to scoring reading tests. Hughes believes that:

> … errors of grammar, spelling or punctuation should not be penalized, provided that it is clear that the candidate has successfully performed the reading task which the item set. The function of a reading test is to test reading ability. To test productive skills at the same time simply makes the measurement of reading ability less accurate.
>
> (Hughes 2002:131)

2. READING

> Reading is a constant process of guessing and what one brings to the text is often more important than what one finds in it. This is why, from the very beginning, the students should be taught to use what they know to understand unknown elements, whether these are ideas or simple words. This is best achieved through a global approach to the text.
>
> (Grellet 1991:7)

Grellet sums up this kind of approach in the following way:

Study of the layout:	→	Making hypotheses	+	anticipation of where
title, length, pictures,		about contents		to look for confirmation
typeface of the text		and function		of these hypotheses
				according to what one
				knows of such text types
				↓
Second reading	←	Further prediction	←Confirmation	←Skimming
for more detail			of revision of	through the
			one's guesses	passage

(Grellet 1991: 7)

Wallace explains what reading means as follows:

> The most important resource that any potential reader possesses, whether reading in a first or any other language, is an awareness of the way in which we use language. For reading is above all to do with language.
>
> (Wallace 1992:3)

Smith and Barrett discuss different definitions of reading and their implications for instruction as follows:

> First, differing definitions of reading have differing implications for instruction; second, the question provides an avenue for presenting a definition which recognizes the importance of affective behaviours in the reading act and finally, the discussion will hopefully cause the reader to take an accounting of his or her definition of reading and its implications for instruction.
>
> (Smith and Barrett 1976:96)

They assume that the latter reason is the most important one, for the definition of reading a teacher maintains determines the goals of the reading program he or she carries out. Smith and Barrett conclude that there are differing definitions of reading which have differing implications for instruction (Smith and Barrett 1976:96).

Here are some of the definitions of reading according to Smith and Barrett:

- Reading is decoding

- Reading involves word identification and comprehension
- Reading involves interactions between thought and language
- Reading involves perceptual, cognitive, and affective responses

Regarding the first definition, Smith and Barrett present decoding as the ability to produce the phonemes or sounds represented by graphemes or written letters in English (Smith and Barrett 1976:96). They quote Bloomfield who put it in the following manner: ”The letters in a piece of English writing do not represent things, or even words, but sounds. The task of the reader is to get the sounds from the written or printed page“ (Bloomfield in Smith and Barrett 1976:96). Bloomfield also clearly states the place of comprehension in his definition:

> A person who can read aloud in a text that is before his eyes, but cannot reproduce the content or otherwise show his grasp of it, lacks something other than reading power, and needs to be taught the proper response to language, be it presented in writing or in actual speech. The marks on the page offer only sounds of speech and words, not things or ideas.
>
> (Bloomfield in Smith and Barrett 1976: 32)

Obviously, comprehension is not involved in this definition. Bloomfield claims that comprehension is not uniquely inherent to reading, but diffuses all use of language. It is argued that the decoding definition is applicable only to the beginning stages of reading (Smith and Barrett 1976:97).

Another definition of reading has over the years emerged as the most widely accepted one. It says that reading involves word identification and comprehension. Smith and Barrett present a representative of this type of definition offered by DeBoer and Dallman:

> In reading we employ visual symbols to represent auditory symbols. The basic task in reading is therefore to establish in the mind of the reader automatic connections between specific sights and the sounds they represent. Since the sounds themselves are symbols of meanings, the process of reading involves a hierarchy of skills ranging from auditory and visual discrimination to such higher-order mental activities as organizing ideas, making generalizations, and drawing inferences.
>
> (DeBoer and Dallman in Smith and Barrett 1976:97-98)

This definition includes word identification and comprehension as integral parts of reading. DeBoer and Dallman suggest that the reader brings meaning to the printed page and that his or her intent and his or her background of information permit the reader to develop new understandings and modify old concepts as a result of what an

author writes (Smith and Barrett 1976:98). In this approach emphasis is put on such things as:

a) the use of context clues as an aid in the identification of words
b) silent reading for a purpose
c) efforts by the teacher to stimulate students to think about and react to what they read in a variety of ways
d) prepared oral reading by students to inform or entertain classmates in an audience setting

(Smith and Barrett 1976:98)

The third type of definition of reading claims that reading involves interaction between thought and language. Goodman believes that reading includes not only perception but also the use of syntactic and semantic information:

> Reading is a selective process. It involves partial use of available minimal language cues selected from perceptual input on the basis of the reader's expectation. As the partial information is processed, tentative decisions are made to be confirmed, rejected, or refined as reading processes… More simply stated, reading is a psycholinguistic guessing game. It involves an interaction between thought and language. Efficient reading does not result from precise perception and identification of all elements, but from skill in selecting the fewest, most productive cues necessary to produce guesses which are right the first time.
>
> (Goodman in Smith and Barrett 1976:98)

There would be emphasis on meaningful silent reading, with students being encouraged to predict outcomes ahead of reading. Smith and Barrett add that the use of context clues, both semantic and syntactic, would be emphasized as the basic approach to world identification (Smith and Barrett 1976:99). They created their own definition:

> Reading involves the visual perception of written symbols and the transformation of these symbols into their audible or inaudible oral counterparts. The audible or inaudible oral responses act as stimuli for thoughtful reactions on the part of the reader. The types or levels of thought induced by the stimuli are determined, in part, by the syntactic and semantic accuracy of the audible or inaudible oral responses; the general language sophistication, intent, and background of the reader; and the nature of materials. In addition, the effort expended in the perceptual and intellectual acts is partially controlled by the reader's interest in a specific selection and by his attitude toward reading in general.
>
> (Smith and Barrett 1976:99-100)

They explain that this definition is three-dimensional in nature. It recognizes word identification and comprehension as integral parts of the reading act and in this

respect, is similar in intent to the two previous definitions. What makes this definition different is the underlined portion which clearly recognizes that affective responses are involved in reading (Smith and Barrett 1976:100).

They give reasons for including affect in the definition. These reasons come from two sources. First one comes from the teachers who observed the influence of interests and attitudes on students' reading performances; second, from researchers. They studied the relationships between reading interest and reading comprehension in sixth graders and concluded that reading interest may enable most students to read beyond their measured reading ability.

Smith and Barrett sum up that a reading program designed to the specifications of the three-dimensional definition would resemble the programs based on the previously cited definitions. However, they add, the most striking feature one would see in such a program would be the attention given to the development of interests in, attitudes toward, and valuing of reading (Smith and Barrett 1976:100).

Reading is an active skill. As mentioned earlier, it constantly involves guessing, predicting, checking and asking oneself questions. Grellet supposes this should be considered when constructing reading exercises.

The second aspect of reading as an active skill, Grellet brings about, is its communicative function. Exercises should be meaningful and should correspond as often as possible to what one is expected to do with the text. "We rarely answer questions after reading a text, but we may have to:

- write an answer to a letter
- use the text to do something (e.g. follow directions, make a choice, solve a problem)
- compare the information given to some previous knowledge" (Grellet 1991:9)

> The students must be taught how to approach and consider the text in order to become independent and efficient readers. It is also important to remember that meaning is not inherent in the text, that each reader brings his own meaning to what he reads based on what he expects from the text and his previous knowledge. This shows how difficult it is to test competence in reading comprehension and how great the temptation is to impose one's own interpretation on the learners.
>
> (Grellet 1991: 9).

2.1. Reading and reading comprehension

Grellet points out that reading comprehension is a communication skill and should not be separated from the other skills. As he sees it, there are not any situations

in real life when we do not talk or write about what we have read or when we do not relate what we have read to something we might have heard. It is therefore important, he emphasizes, to link the different skills through reading activities chosen: reading and writing; reading and listening; reading and speaking (Grellet 1991:8).

"Understanding a written text means extracting the required information from it as efficiently as possible" (Grellet 1991: 3) In other words, according to the purpose of our reading we use different reading strategies. It is therefore very important to consider following elements: What, why and how do we read?

2.1.1. What do we read?

Here are the main text-types we usually come across (we will talk about the text types in more details later in the thesis):

- Novels, short stories, tales; other literary texts and passages (e.g. essays, diaries, anecdotes, biographies
- Plays
- Poems, limericks, nursery rhymes
- Letters, postcards, telegrams, notes
- Newspapers and magazines (headlines, articles, editorials, letters to the editor, stop press, classified ads, weather forecast, radio/TV/theatre programmes)
- Specialized articles, reports, reviews, essays, business letters, summaries, précis, accounts, pamphlets (political and other)
- Handbooks, textbooks, guidebooks
- Recipes
- Advertisements, travel brochures, catalogues
- Puzzles, problems, rules for games
- Instructions, directions, notices, rules and regulations, posters, signs, forms, graffiti, menus, price lists, tickets
- Comic strips, cartoons and caricatures, legends (of maps, pictures)
- Statistics, diagrams, flow / pie charts, time-tables, maps
- Telephone directories, dictionaries, phrasebooks

(Grellet 1991:3-4)

Grellet believes that it is very important to use authentic texts whenever possible. He gives three reasons for it:

- simplifying a text often results in increased difficulty
- the difficulty of a reading exercise depends on the activity which is required of the students rather than on the text itself

- authenticity means that nothing of the original text is changed and also that its presentation and lay out are retained (pictures, newspaper headlines, etc. - using these non-linguistic clues help the readers a lot)

(Grellet 1991:7)

2.1.2 Why do we read?

Grellet presents two main reasons for reading: reading for pleasure and reading for information. By reading for information it is meant to read in order to find out something or in order to do something with the information we get (Grellet 1991:4). Unlike Grellet, Wallace distinguishes three reasons for reading: reading for survival, reading for learning and reading for pleasure (Wallace 1992:6-7). She explains that: "some reading is almost literally a matter of life and death, for example a 'stop' sign for a motorist. Survival reading serves immediate needs or wishes, such as 'ladies', 'gentlemen' or "'exit'" (Wallace 1992:6). Moreover, she adds that it has been found that children from all social backgrounds willingly acquire an understanding of print, related to the ways they perceive their day-to-day needs and interests. They get these from sources like TV, advertising or street signs. This is sometimes called "environmental print" (Wallace 1992:6-7)

2.1.3 How do we read?

The main ways of reading according to Grellet are skimming, scanning, extensive and intensive reading. To briefly explain the terms (they will be explained in more details in the next chapter): skimming means running over a text to get the gist of it; scanning refers to going quickly through a text to find a particular information; by extensive reading we understand reading longer texts, usually for pleasure (it is a fluency activity involving mainly global understanding); finally, the term intensive reading is used for reading shorter texts to extract specific information (it is more an accuracy activity involving reading for detail) (Grellet 1991:4).

Grellet explains that these different ways of reading are not mutually exclusive. He uses the example of when we first want to skim the text to see whether it is worth scanning a particular part for the information we are looking for.

Our reading purposes change very often and that is what we should bare in mind when preparing reading tests. We should give our students the opportunity to answer

different questions, try different activities according to the type of text and the purpose in reading it.

2.2 Sensitizing

The way the text is perceived by readers is called sensitizing. It has the following phases: inference, understanding relations within the sentence and linking sentences and ideas.

"Inferring means making use of syntactic, logical and cultural clues to discover the meaning of unknown elements" (Grellet 1991:14). Grellet believes that it is better not to explain difficult words from a new text beforehand. They could be then used to such help and would not try to cope with a hard passage on their own. The students should be, instead of that, encouraged to first make a guess what the unknown word could mean. It is better than looking the word up in a dictionary immediately. It is therefore significant to develop the skill of inference from the very beginning (Grellet 1991:14)

As inability to infer the meaning of unknown elements makes readers discouraged, a similar problem arises when readers do not understand the sentence structures. It is, therefore, very important to give the students enough opportunity to practise looking for the "core" of the sentence (subject and verb).

Another area Grellet believes it is essential to prepare our students in is recognizing different ways that are used to form textual cohesion, particularly the use of reference and link-words (Grellet 1991:15). By reference they mean all the devices allowing lexical relationship within a text. (For instance, reference to an element previously mentioned – anaphora, reference to the element mentioned below – cataphora, use of synonymy, hyponymy, comparison, nominalization, etc.)

> It is important for the students to realize that a text is not made up of independent sentences or clauses, but that it is a web of related ideas that are announced, introduced and taken up again later throughout the passage with the help of references.
>
> (Grellet: 1991:15)

Grellet points out that if the reader does not understand some words of the passage, some of the facts and ideas will probably escape him. Though if the reader does not understand inter-or intra-sentential connectors, he or she may fail to recognize the communicative value of the passage since those words are like signals indicating the

function of what follows, for example announcing a conclusion, supposition, etc. (Grellet 1991:16). On that account students should be taught not only to know what these connectors mean but also to look for them in a text, which will they need and appreciate especially when skimming.

2.3 Reading techniques

According to Grellet one of the most important things to bear in mind when we are teaching reading comprehension is that there is not just one type of reading. There are several types of reading and they differ in reasons or purposes for which the texts are read. To read efficiently students must learn to adapt their reading speed and technique to their aim when reading. If they read all texts in the same way, they would waste their time and not remember what is important for them because they would have to absorb too much information that is irrelevant for them at the time (Grellet 1991:17). Grellet suggests a few techniques that could be useful for readers who want to become really effective in their reading: predicting, previewing, anticipation, skimming and scanning.

Predicting is rather a skill than a technique and it is basic to the process of reading generally. A reader can use grammatical, logical and cultural hints to predict or guess what will come next. We can train our students in it by, for example, giving them unfinished passages to complete or by going through a text little by little, stopping after each sentence in order to predict what is likely to come next (Grellet 1991:17).

Previewing is, unlike predicting, a very specific reading technique which involves using the table of contents, the appendix, the preface, the chapter and paragraph headings in order to find out where the required information is likely to be. Grellet adds that previewing is particularly useful when skimming and scanning and as a study skill (Grellet 1991:18).

When we are motivated to read a text it means that we are also expecting to find answers to many questions and specific information or ideas we are interested in. “This ‘expectation’ is inherent in the process of reading which is a permanent interrelationship between the reader and the text.” (Grellet 1991:18) Unfortunately, this cannot always be guaranteed in the classroom; students are often given a text with a topic they do not know much about, they cannot put the text situation in a more general cultural context and what is most important, they have no particular desire to read. In such situations it

is almost impossible to expect that the students will improve their reading. Grellet suggests letting students choose the topic they would be interested in as often as possible. However, this is not always possible. The teacher then can try to introduce a new topic and hope it will catch his or her students' attention (Grellet 1991:18).

As Grellet claims: "Both skimming and scanning are specific reading techniques necessary for quick and efficient reading" (Grellet 1991:19). By the term skimming we mean going through the reading material quickly in order to get the gist of it, to know how it is organized, or to get an idea of the tone or the intention of the writer. In scanning, on the other hand, we try to locate specific information and often we do not even follow the linearity of the passage to do so. Grellet describes the process as follows: "We simply let our eyes wander over the text until we find what we are looking for, whether it be a name, a date, or a less specific piece of information" (Grellet 1991:19). Skimming is therefore more thorough; it requires an overall view of the text and signifies definite reading. On the contrary, scanning only picks what information is relevant to our reading purpose.

Moreover, we can apply both techniques together. For instance, we can first only skim through the article to get to know whether it is worth reading or not. When we find the text interesting, then we can read it more carefully. It is also possible to scan the text after careful reading in order to note particular information we want to remember (Grellet 1991:20).

2.4 Skills involved in reading

As Grellet states, reading involves a variety of skills. He offers a list of the main ones (the list was taken from John Munby's *Communicative Syllabus Design*):

- Recognizing the script of language
- Deducing the meaning and use of unfamiliar lexical items
- Understanding explicitly stated information
- Understanding information when not explicitly stated
- Understanding conceptual meaning
- Understanding the communicative value (function) of sentences and utterances
- Understanding relations within the sentence
- Understanding relations between parts of a text through lexical cohesion devices
- Understanding cohesion between parts of a text through grammatical cohesion devices
- Interpreting text by going outside it
- Recognizing indicators in discourse

- Identifying the main point or important information in a piece of discourse
- Distinguishing the main idea from supporting details
- Extracting salient points to summarize (the text, an idea etc.)
- Selective extraction of relevant points from a text
- Basic reference skills
- Skimming
- Scanning to locate specifically required information
- Transcoding information to diagrammatic display

(Grellet 1991:4-5)

Heaton tried to identify these specific skills involved in reading as follows:

- recognize words and word groups, associating sounds with their corresponding graphic symbols
- deduce the meaning of words by
  a) understanding word formation
  b) contextual clues
    - understand explicitly stated information
    - understand relations within the sentence, especially

  a) elements of sentence structure
  b) negation
  c) fronting and theme
  d) complex embedding
    - understand relations between parts of a text through both lexical and grammatical cohesive devices, especially anaphoric and cataphoric reference and connectives
    - perceive temporal and spatial relationship, and also sequences of ideas
    - understand conceptual meaning, especially:

  a) quantity and amount
  b) definiteness and indefiniteness
  c) comparison and degree
  d) means and instrument
  e) cause, result, purpose, reason, condition, addition, contrast, concession
    - anticipate and predict what will come next in the text
    - identify the main idea and other salient features in a text

- generalize and draw conclusions
- understand information not explicitly stated by

a) making inferences (i.e. reading between lines)

b) understanding figurative language

- skim and scan
- read critically
- adopt a flexible approach and vary reading strategies according to the type of material being read and the purpose for which it is being read

(Heaton 1991:105-106)

Heaton emphasizes the fact that reading aloud is not included in these specific reading skills. He argues that it is because reading aloud is considered to be a unique skill since it involves different skills from silent reading (Heaton 1991:106).

According to Heaton, there are two different kinds of complementary reading activities to which students are usually asked to deal with; that are intensive and extensive reading. Short extracts which contain features requiring detailed study form basis for intensive reading practice. On the other hand, whole articles, chapters and books are used for extensive reading practice. As Heaton claims, most reading tests, unfortunately, concentrate on intensive reading. He presumes that the reason for that is probably: "because it is more economical to have a large number of items based on a short reading extract than a few items based on a much longer one" (Heaton 1991:106).

As regards reading tests, Heaton suggests similarly to Grellet to include a variety of text types for reading comprehension; for instance newspaper articles, instructions for using appliances and machinery, directory extracts, public notices, timetables and maps, advertisements, etc. He states that:

> The inclusion of such text types will not only provide a more realistic and reliable means of assessment but will also help to motivate students by demonstrating how the target language is used in real-life situations.
>
> (Heaton 1991:107)

3. TESTING READING

Heaton introduces testing reading comprehension as follows:

> Until recently the many and diverse reading skills and strategies for use in everyday situations have been largely subordinate to a narrower range of skills required for dealing with simplified readers, especially at the elementary levels.
>
> (Heaton 1991:105)

Moreover, first efforts to deal with complex reading skills often come too late, not earlier than at the tertiary level. This is when the students have to cope with professional and technical literature in the foreign language.

Next, Heaton points out that before constructing reading tests in the second or foreign language, we must be aware of the students' first language reading skills. "Clearly, there is often little purpose in testing in the second language those basic reading skills which the students have not yet developed in their own language" (Heaton 1991:105). He adds, that despite this fact, mastering some reading skill in the first language does not necessarily mean that the student is able to use those skills in reading another language.

Alderson claims that a reading score may be high or low because of item difficulty rather than text difficulty and vice versa. He distinguishes between item effects and passage effects (Alderson 2001:86).

3.1. Factors affecting the difficulty of reading test items

3.1.1 Language and types of questions

First of all, it is a language of questions. If the language of questions is harder than the language used in a text, we cannot tell whether the reader's low score is due to his not understanding the text or the questions. It is advisable to use simple language for the questions, easier than it is in the text. Alderson poses a question concerning second-language reading: "Should the question be in the target language, the language of the passage, or in the first language of the reader?" (Alderson 2001:86).

The next factor affecting the difficulty of reading test items Alderson brings up is the issue of types of questions. Pearson and Johnson identify three different types of questions and point out that they may vary in their difficulty (Pearson and Johnson in Alderson 2001:87). They are: textually explicit questions, textually implicit questions and script-based (scriptally implicit) questions.

Textually explicit questions are those where the question information and the correct answer can be found in the same sentence. Textually implicit questions, on the contrary, demand to combine information between sentences. Script-based questions require readers "to integrate text information with their background knowledge since correct responses to the questions cannot be found in the text itself" (Alderson 2001:87). This is not, however, the only categorizing of types of questions; Bensoussan et al distinguish local and global questions (for more information see Alderson 2001:88).

3.1.2 Role of grammar and vocabulary in reading tests

As for the role of grammar in reading tests, Alderson states:

> An issue for many developers of second-language reading tests is whether their test linguistic competence, and particularly grammatical competence as well as, or indeed more than, reading comprehension.
>
> (Alderson 2001:98)

Alderson believes that: "Tests of vocabulary are highly predictive of performance on tests of reading comprehension" (Alderson 2001:99). In other words, he assumes that vocabulary plays a very important role in reading tests; "Clearly vocabulary is important to text comprehension, and thus to test performance" (Alderson 2001:99).

3.1.3 Use of dictionaries in reading tests

To reduce the effect of vocabulary knowledge on measures of reading comprehension, Alderson suggests allowing students to compensate for lack of vocabulary by consulting dictionaries. Some test constructors, however, do not agree with this idea for, as they claim, it "invalidates the test since the dictionaries provide some of what is being tested" (Alderson 2001:99-100). What is more, students would waste time looking up words that would be better spent on reading the text. Bensoussan et al. investigated the effect of dictionary usage on EFL test performance and concluded, "the use of dictionaries had no effect on students' test scores, regardless of whether the dictionary was bilingual or monolingual" (Bensoussan in Alderson 2001:100). Nesi and Maera confirmed the fact that using dictionaries does not have a significant effect on test scores, but they found that students using dictionaries spent much more time to complete the reading tests (Nesi and Maera in Alderson 2001:100).

So, to conclude, using dictionaries during reading tests is still an open question. (For more information see Alderson 100-101)

3.2 Factors affecting the difficulty of reading test texts

Many aspects of text that could influence the reading process have been studied. Although linguistic is the major source of understanding the language of the text the concern for the reader should be considered as well. Texts analysts from different backgrounds have contributed to better insight into the factors that influence the reading process. These variables concern text content, text types or genres, text organisation, sentence structure, lexis, text typography, layout, the relationship between verbal and non-verbal text, and the medium in which the text is presented (Alderson 2001:61).

3.2.1 Text and discourse

Wallace points out two facts about language: first, she states, we use the language for a purpose; second, the language only makes sense in context.

Wallace distinguishes two ways how we can look at written language. She uses two different terms - *text* and *discourse*. According to her, text is an output of a writer which can be recorded and studied, while a discourse approach to reading focuses less on the text as product and more on the reader's process of constructing meaning from it (Wallace 1992:8). At this point Wallace mentions some difficulties of the English writing system. She states that correspondences between sound and written symbol are less consistent than in languages such as Spanish, Urdu, Hindi or Arabic (Wallace 1992:9). Another facet she mentions in connection to reading and text are features of connected text. She looks at texts in three different ways:

1) in terms of formal features (at ways features of grammatical system are used to link sentences or paragraphs)
2) in terms of their prepositional meaning (how ideas or concepts are expressed and related to each other)
3) in terms of their communicative function (the ways in which sections of a text can be interpreted in relation to other sections and of the function of any text as a whole)

(Wallace 1992: 11)

Wallace concludes that readers are helped in their interpretation of texts both by their knowledge of the principles of word formation and cohesion, and by their ability to attribute an appropriate communicative function to texts and parts of texts (Wallace 1992: 14).

Concerning the term *discourse*; Wallace uses this term to describe the meaning which the reader constructs from the text during the reading process. She adds that it has been argued that there is not just a single discourse but a number of them. These discourses are rather social than personal for they relate to social practices and beliefs. However, they are not just socially determined; they also are culture specific (Wallace 1992: 14-17).

3.2.2 Function of the text

As regards function of the text; understanding the function of a passage is crucial to its comprehension. That is why we should train our pupils to find out the aim of the text. As Grellet explains, the very form of the passage, the way it is printed, laid out, or the place where it was found, helps us to recognize the function. (Grellet 1991:20). For that reason teachers should lead their students to get used to looking at the text from this point of view at the very first moment they see it.

3.2.3 Organization of the text

The way the text is organized is one of the things that distinguish various text types or genres. Text organization, the way paragraphs are related to each other and the relationships between ideas are signaled or not, has been studied for a long time. Alderson adds that even within one genre different organizations might lead to different outcomes or processes (Alderson 2001: 67).

As Alderson states, Urquhart proved the effects of chronological and spatial ordering in text. He showed that for both native and non-native readers of English, “texts organized according to the sequence of events could be read faster and were easier to understand than texts whose temporal sequencing was disturbed” (Alderson 2001:67). Urquahart also pointed out that texts with “consistent spatial organization”, such as following a clear logical sequence, from outside in, or left to right, were easier to understand and more memorable.

According to Meyer we can distinguish five different types of expository text; in other words five different ways in which topics can be organized: “collection (lists, causation (cause and effect), response (problem – solution), comparison (compare and contrast) and description (attribution).” (Meyer in Alderson 2001:67)

Concerning the effects of cohesion on understanding and recalling, Alderson states that cohesion is not a key variable in readability; however, “conjunctions do

facilitate discourse processing for average-ability readers when the topic is less familiar" (Alderson 2001:68). He adds that the reason for the weak effects of cohesion can be caused by the readers' ability to make bridging inferences.

As Grellet describes, "the organization of a passage is not always determined by its contents and by the nature of the information to be conveyed" (Grellet 1991:20). Very often the writer chooses the thematic pattern. And when the students recognize the pattern it helps them to guess what will probably follow. They use their reading strategies to predict what can come next. For instance, if the text looks as an argumentative one, the reader expects to find (further in the text) arguments, counter-arguments and finally some conclusion drawn from these arguments (Grellet 1991:20).

One of the kinds of organization of the text is building around a main idea, which is further developed throughout the text. We can find it in newspaper articles, where the first paragraph very often sums up the main idea and then the rest of the text is analyzed.

3.2.4 Understanding the meaning of the text

Though understanding the function and organization of the text is very important, understanding the content is vital. According to Grellet, the comprehension is usually checked through different types of questions; for example open questions, right or wrong, multiple-choice questions. They can involve the students actively into reading. Other activities to help or check readers' comprehension suggested by Grellet are divided into two categories:

- To make the students active in the reading process by presenting them with decision-making activities (e.g. drawing a diagram with the information given in the text, solving the problem, completing a table which reorganizes the information).
- To devise activities which are as natural as possible, i.e. as close as possible to what one would naturally do with the text (e.g. completing a document, comparing several texts, etc.)

(Grellet 1991:22)

The above suggested activities Grellet divides into two categories: non-linguistic response to the text and linguistic response to the text.

Regarding non-linguistic response to the text, Grellet states several comprehension activities that do not require any complex verbal response on the part of

the learners (Grellet 1991:22). In such exercises pupils are asked to relate the text to what is added (a document, a diagram, a picture). We can divide these tasks as follows:

- a comparison (e.g. comparing texts and pictures, matching passages of the text and diagrams)
- a transposition of the information (transcoding the information into the form of diagram, completing or labelling a document)
- using the information in the passage to find solution, make a decision or solve a problem

(Grellet 1991: 22)

The last type of exercise proved to be one of the most useful ones because it reflects situations, the purpose of reading from real life. The fact that the student is able to make a decision is a proof that he or she thought about the text and understood it (Grellet 1991:22).

To practise linguistic response to the text we can, as Grellet presents, use exercises suggested in the following categories: reorganizing the information, comparing several texts, completing a document and study skills.

Reorganizing the information is in other words presenting the information in a different way or according to a different pattern. The students can be asked to, for instance, complete a table or draw a chronological list of the events mentioned in the passage. The good point of this kind of exercises is that it underlines the fact that one piece of information can be presented in more than just one way.

Comparing several texts is very natural activity. As Grellet claims, we are used to compare different versions of the same event or incident in every day situations (Grellet 1991:23). We compare, for example the text from a guidebook to what our friends tell us about a country they visited.

As for completing a document, students are required to, for instance, answer a letter, fill in the evaluation card, an application form or leave a note. We can use this category for simulations or role-play. A student can, for example, identify with one of the characters from the text and then react in different new situations.

Study skills include tasks like use of a dictionary, note-taking, summarizing or others. Grellet presents note-taking as a basic skill to remember what one reads or listens to; moreover, he points out that: “when taking notes, it is necessary to establish the structure of the text and its key ideas and to learn to leave out unessential information” (Grellet 1991:23). In addition, it is a difficult activity which sums up most of the reading strategies mentioned earlier in this work.

To compare it to writing a summary, it is necessary to see both the same as well as different characteristics. What is in common in note-taking and summarizing is the need of refusing secondary details. However, there is some dissimilarity, as Grellet states:

- a summary is usually written in one's own words
- it does not necessarily imply outlining the structure of the passage, as note-taking usually does
- it should be an accurate and objective account of the text, leaving out our reactions to t (whereas note-taking can be supplemented by note-taking, i.e. briefly jotting down one's reactions and ideas about the passage)

(Grellet 1991:24)

3.2.5 Text topic and content

It is commonly assumed that text content will affect how readers process text. Abstract texts will be harder to understand than describing real objects, events or activities. The more concrete, imaginable and interesting, the more readable the text will be. Alderson adds that texts located in familiar settings, on everyday topics, are likely to be easier to process than those that are not (Alderson 2001:62). The quantity of information and the density of propositions in a text affect understanding and recall. Non-specialist texts in the arts and humanities will be easier to process for more people of equivalent educational background than scientific texts. As Alderson presumes this is because more people will have read fiction, popular journalism, advertisements and simple expository texts, than will have read technical or scientific texts. It is part of most people's education to read the literature and contemporary journalism of their tongue. Alderson speculates that future generations, due to growing emphasis on science education and the increasing role of technology in society, however, may be more familiar with broadly scientific texts.

Test designers have drawn to a conclusion that it is more appropriate to take texts from popular fiction and non-fiction on the grounds that they are likely to be less biased in terms of difficulty, and therefore more suitable for tests of reading.

To conclude, it is advisable to be aware that variation in the text content may lead to different test results. Therefore, the readers should be assessed for their ability to understand texts in a range of topics. Bachman and Palmer argue that the approach one takes must take into consideration the presupposed background knowledge (topical knowledge) of the test-takers (Bachman and Palmer in Alderson 2001:63). They suggest

three ways how to decide on the approach that is the most appropriate to a particular testing situation: excluding background knowledge from the construct; including both background knowledge and language ability in the construct; and defining background knowledge and language ability as separate constructs.

3.2.6 Text type and genre

Certain topics are associated with certain types of text. What causes difficulty in texts, according to Alderson, is less the actual content than the way the text is written: "Its style of the text or the features make one text different from another …" (Alderson 2001:64). That leads to different classifications of text type; Alderson explains the differences between expository and narrative texts, literary and non-literary texts.

Expository texts are generally assumed to be more difficult to process. Alderson sees the reason for that in "the greater variety of relationships among text units, possibly due to greater variety of content" (Alderson 2001:64). To the narrative texts he points out that they appear to induce visualisation; readers report seeing scenes in their head when reading such texts. The interesting fact is that different people visualise different scenes (this probably depends on their prior experience as well as on their expectations. The visualisation, however, becomes an important part of readers´ understanding. Alderson concludes that text variables only have a crucial role when materials are conceptually more difficult or unfamiliar and when readers are relatively less able (Alderson 2001:65).

Concerning literary and non-literary texts; literary texts are assumed to be more difficult to process. Alderson gives two reasons for that: first, there are multiple layers of meaning; second, the range of language which is wider and more complex. However, significant differences between literary and non-literary texts are disputable. We cannot think of literary texts as of a homogeneous whole since there is a number of genres (including, for example, fiction as well as non-fiction). "Rather there might be a cline of ´literariness´ on which texts might be placed, and whose features might be identifiable empirically." (Alderson 2001:66)

3.2.7 Text readability and text simplification

Researches have been concerned with features that make text readable for a long time. It has been important especially in educational links. Many attempts have been made to come up with pattern (based on empirical research into difficulty) that could be

used to estimate text readability. The lexical load can be judged by checking how many words of the text appear in a word frequency list. In English word frequency is very roughly related to word length – more frequent words tend to be shorter. One way reflects the number of syllables, another way is called Flesch and this formula provides a reading-ease score:

$$RE = 206.835 – (0.846 \times NSYLL) – (1.015 \times W/S)$$

where NSYLL is the average number of syllables per 100 words and W/S is the average number of words per sentence (Davies in Alderson 2001:71). The number of words on average per sentence can also be counted as a measure of readability. As Alderson states short sentences are syntactically simpler than long sentences, although some researches points out the fact that adding (not deletion) of words to sentences make them easier to comprehend (Alderson 2001:72).

To measure text readability cloze techniques were developed. Many studies have shown high correlation between readability measured by formulae and cloze. Although Taylor argues that cloze could provide a more accurate estimate of readability since it involves real readers processing texts, Alderson and Harrison warn of uncritical acceptance of cloze test results. They suggest the combination of expert judgement and readability formula (Alderson 2001:72).

Research into text readability has been accompanied by research into text simplification. When the text had been considered too difficult for the intended readers the question how to simplify it arose. Different methods have been studied. Davies and Widdowson, for example, distinguish between ‘simplification’ and ‘simple’: “a simple account is an authentic piece of discourse, a simplified account may or may not be authentic, and is usually pedagogic in intent. It may, however, not be simple.” (Davies and Widdowson in Alderson 2001:72) Strother and Ulijn discovered that simplifying texts syntactically but not lexically does not necessarily make the texts more readable. They therefore suggest using a conceptual rather than a syntactic strategy since it involves processing content words and hence requires lexical and content knowledge. Alderson concludes that it has long been known that vocabulary load is the most significant predictor of text difficulty and adds quotation from Chall: “Once a vocabulary measure is included in a prediction formula, sentence structure does not add very much to the prediction.” (Chall in Alderson 2001:73)

To conclude, the testers should be aware of different components that affect text difficulty, such as topic, syntactic complexity, cohesion, coherence, vocabulary and readability. All these should be taken into account when selecting texts. We can simplify the texts, however, we should be careful not to actually make them harder (Alderson 2001:74).

3.2.8 Typographical features

Researches are still interested in what features of print, fonts and layout might be important in causing reading ease or difficulty. Alderson mentions her for example the fact that the top half of normally mixed-case print is more informative than the bottom half or that the first half of English words is more informative than the second part. Although the effect of such variables in reading is disputable, testers are advised to make sure the texts are suitably presented and legible. It is not desirable to penalize the reader due to bad layout or copy (Alderson 2001:76)

3.2.9 Verbal and non-verbal information

Concerning the use of non-verbal or graphic information in text, Alderson indicates that text that contains only verbal information will be “not only intimidating but also more dense and therefore much more difficult to process” (Alderson 2001:77). Research into the relationship between verbal and non-verbal information has specifically concentrated on advertisements. There is a disjunction between text and illustration in many advertisements, “such that one or the other appears surprising, contradictory or humorous, thereby attracting the readers’ attention and becoming more memorable” (Alderson 2001:77). Tables, diagrams or other forms of presentation of data are used in many genres in order to offer an alternative and complementary way of processing information. The information presented in tables, diagrams or other forms, however, very often provides support of the verbal information. In other words, to understand the text completely, readers need to read both. Alderson therefore recommends maintaining the relationship between the verbal and non-verbal in test texts. He also sums up that “testers should consider assessing a reader’s ability to understand that relationship, as well as their ability to use the graphic information to understand the verbal, and vice versa.” (Alderson 2001:77)

3.2.10 The medium of text presentation

Discussing the effect of text variables on reading the medium by which the text is presented should also not be forgotten. Information can be presented on overhead slides or on TV or computer screens. Even though with the development of the Internet and the World Wide Web more and more information is now processed on screen, many readers prefer to print out texts. They can process them at leisure. Alderson mentions one significant limitation of this medium and that is "readers can only process one screen at a time and scrolling forward and backwards is more time-consuming and less efficient than turning pages" (Alderson 2001:78).

3.2.11 Presence of text while answering questions

Similarly to the question of using dictionaries the question of presence of text while answering questions has been discussed widely. Should we allow students to look back at the passage when answering questions or should we remove the text before allowing the students to respond? Alderson presumes that removing the text increases the role of memory in the responding, although not in the comprehending process (Alderson 2001:106). Davey and Lasasso found an interaction between question type and the removal of text. Alderson describes what they found:

> When subjects were allowed to look back at the text, they performed better than when not allowed, but also there was an interaction with item type. When subjects were allowed to look back at the text there were no significant differences between selected response and constructed response items. However, when subjects were not allowed to look back at the text, selected response items were easier than constructed response items.
>
> (Alderson 2001:106-107)

Alderson further delineates the Johnston study devoted to this issue. He sums up that we can find advantages and disadvantages in both variants; central questions are, according to him, easier to answer without the presence of a text (probably because main ideas are incorporated into schemata and reinterpreted on retrieval), whereas peripheral questions are easier to answer in the presence of the text since readers can use matching or search-and-retrieval strategies (Alderson 2001:109).

3.2.12 Text length

A question all reading-test constructors think of is how long the text should be. Alderson points out that: "Text length is a surprisingly underresearched area" (Alderson 2001:108). After discussing Engineer's research into this area and some others, he concludes:

> This points up the sort of compromise one is often presented with in testing, in this case between maximizing authenticity by using the sort of long texts that students might have to read in their studies, on the one hand, and minimizing content bias by using several shorter passages, on the other hand.
>
> *(Alderson 2001:109)*

3.3 Test techniques

Alderson gives reasons why he uses the terms “test method”, “test technique” and “test format” more or less synonymously: “The testing literature in general is unclear as to any possible difference between them” (Alderson 2001:202).

Weir explains that test methods are used to construct tests but are not tests in themselves. Unlike tests we cannot evaluate test methods as good or bad, valid or invalid. A multiple-choice procedure, for instance, might produce a valid test in one case but not in another. And this is the same for all methods (Weir: 1990:42).

Alderson claims that different testing techniques allow the measurement of different aspects of the construct to be assessed. “Therefore”, Alderson stresses, “it is important to consider what techniques are capable of assessing, as well as what they might typically assess” (Alderson 2001:202).

Weir remarks that there is some evidence that test format might affect student performance. We would like to introduce main kinds of test formats and stress their potential advantages as well as disadvantages. Weir further mentions the main condition for testing within a communicative framework: “the test task should as far as possible reflect realistic discourse processing and cover the range of contributory enabling skills” (Weir 1990:42). What is more, it is also very important that tests developed according to this model should have a strong washback effect on practice in the language classroom.

Alderson discusses what many books on language teaching assert; that is that there is a significant difference between teaching techniques and testing techniques. However, he believes that this distinction is overstated, and that the design of a teaching exercise is in principle similar to the design of test item. Moreover, he adds:

> The primary purpose of a teaching / learning task is to promote learning, while the primary purpose of an assessment task is to collect relevant information for purposes of making inferences or decisions about individuals – which is not to say that assessment tasks have no potential for promoting learning, but simply that this is not their primary purpose.

(Alderson 2001:203)

Before we start describing individual test techniques it is important to say that there is no one "best method" for testing reading. "No single test method can fulfill all the varied purposes for which we might test" (Alderson 2001:203). However, as Alderson wants to emphasize, certain methods are common solely for reasons of convenience and efficiency, often at the expense of validity. "And it would be naive to assume that because a method is widely used it is therefore 'valid'" (Alderson 2001:204).

To conclude this passage we can use Alderson' words:

> It is now generally accepted that it is inadequate to measure the understanding of text by only one method, and that objective methods can usefully be supplemented by more subjectively evaluated techniques. Good reading tests are likely to employ a number of different techniques, possibly even on the same text, but certainly across the range of texts tested. This makes good sense, since in real-life reading, readers typically respond to texts in a variety of different ways.
>
> (Alderson 2001:206)

After a thorough consideration of all related aspects the following test techniques were chosen for our research: selective deletion gap filling, short answer questions, dichotomous items and information transfer. Therefore they will be described in more details in the practical part. At this point, however, we will introduce them – but only very briefly.

We talk about *selective deletion gap filling* when the test constructor chooses items for deletion. Alderson describes this testing method as an alternative technique to cloze (Alderson 2001:209). In the cloze exercises some words are deleted from a text. Candidates are to fill in the missing words. The words are deleted regularly; the deletion rate, as Weir refers to it, is every fifth to eleventh word. (Weir 1990:46).

Short answer questions require candidates to write down specific answers in spaces provided on the question paper.

In regards to *dichotomous items*; this testing technique is very popular, mainly because the construction is not very difficult. Students are to decide whether the statement related to the text is "true" or "false". An obvious disadvantage of this test format is the 50 per cent chance of getting the answer right by guessing alone.

In an attempt to avoid the problem of involving writing in the testing of reading comprehension, several Examination Boards in Britain came up with tasks where the

information transmitted verbally is transferred to a non-verbal form, for instance by labeling a diagram, completing a chart or numbering a sequence of events (Weir 1990:50). This technique is called *information transfer*.

3.3.1 Multiple-choice questions

Multiple-choice test items are usually set up in a way that requires a candidate to choose a correct answer from a number of options.

This testing technique or method has many advantages. First, marking cannot be affected by personal judgement. It means that the marking is reliable. It is also very simple and therefore quick and effective. It is also quite easy to pre-test multiple-choice tests. Potential ambiguities or mistakes can be revealed and then corrected. Another benefit is that the format of the multiple-choice test item and its intentions are very clear and obvious. Therefore, candidates know precisely what they are asked to do. The final positive point of multiple-choice questions, according to Weir, is that unlike in more open-ended formats, such as short answer questions, where the testee has to use the skill of writing as well. Alderson adds that this testing technique allows testers to control the range of possible answers to comprehension question and, to some extent, to control the students' thinking when responding (Alderson 2001:211).

On the other hand, we should also mention some of the disadvantages of this test method. One of them is that we do not know whether a candidate's failure is due to lack of comprehension of the text or of the question. He or she can also identify the right answer by eliminating wrong answers, which, as Weir states, is a different skill from being able to find the right answer in the first place (Weir 1990:44). Alderson comments on this as follows:

> Thus it is possible to get an item correct for the "wrong" reason – i.e. without displaying the ability being tested – or to get the item wrong (choosing a distractor) for the "right" reason – i.e. despite having the ability being tested.
>
> (Alderson 2001:212)

We also have to admit that in both multiple-choice tests and in true-false statements tests there is quite a high chance of guessing the right answer (instead of finding it). Another fact we have to bare in mind before choosing a suitable test format is that preparing multiple-choice tests takes a much longer time, is much more

expensive and difficult than more open-ended exams, such as compositions. Alderson assumes that: “By virtue of the distractors, they may present students with possibilities they may not otherwise thought of” (Alderson 2001:211). This can lead to tricking the students and consequently to false measure of their understanding. Some researches argue that the ability to answer multiple-choice questions is a separate ability:

> Students can learn how to answer multiple-choice questions, by eliminating improbable distractors, or by various forms of logical analysis of the structure of the question.
>
> (Alderson 2001:211)

Multiple-choice tests are written by specially trained item writers and pre-tested before use in a formal examination. Each item is thoroughly edited to ensure that:

- There is no superfluous information in the stem.
- The spelling, grammar and punctuation are correct.
- The language is concise and at an appropriate level for candidates.
- Enough information has been given to answer the question.
- There is only one unequivocally correct answer.
- The distractors are wrong but plausible and discriminate at the right level.
- The responses are homogenous, of equal length and mutually exclusive and the item is appropriate for the test.

(Weir 1991:44)

Weir notes that it is extremely time-consuming and demanding to get the requisite number of satisfactory items for a passage, especially for testing skills such as skimming (Weir 1990:44). He also points out the problem of coming up with suitable distractors for items testing the more extensive receptive skills. Therefore Heaton suggests setting simple open-ended questions rather than multiple-choice items for these activities. “Multiple-choice items are for students much easier for they do not have to keep in mind four or five options while going through the text” (Heaton in Weir 1990:44).

Alderson agrees with Weir’s opinion that construction of multiple-choice questions is a very skilled and time-consuming work. He adds: “To write plausible but incorrect options that will attract the weaker reader but not the better reader is far from easy” (Alderson 2001:212).

Another problem with multiple-choice items Weir mentions: ”A further objection to the use of multiple-choice format is the danger of the format having an undue effect on measurement of the trait” (Weir 1990:44). The last mentioned weak point of the discussed test format is, according to Weir, the question of validity. He supposes that:

> There is considerable doubt about their validity as measures of language ability. Answering multiple-choice items is an unreal task, as in real life one is rarely presented with four alternatives from which to make a choice to signal understanding. Normally, when required, an understanding of what has been read or heard can be communicated through speech or writing. In a multiple-choice test distractors present choices that otherwise might not have been thought of.
>
> (Weir 1990:44)

Alderson presents an interesting alternative on multiple-choice. He gives an example where the testee has to read the information about Lancaster University (ten numbered paragraphs) and then find the paragraphs where the answers to ten given questions can be found (for the test, see Alderson 2001:213-214). What Alderson points out is the fact that the test-taker has the same set of options to choose from for each item. (There is a note in the assignment that some paragraphs contain the answer to more than one question.) What is more, Alderson claims:

> Since the response is not a short-answer question, the reader has to read and understand the relevant paragraphs and cannot get the item correct from background knowledge alone. In addition, the questions that are asked are of the sort that a reader reading a text like this might plausibly ask himself about such a text, thereby enhancing at least face validity of the test.
>
> (Alderson 2001:212).

3.3.2 C-tests

C-test is an alternative to cloze as well as to selective deletion gap filling. It intends to test comprehension of the more specifically linguistic elements in a text. This adaptation of the cloze has been developed in Germany by Klein-Braley in 1981. In the C-test every second word in a text is partially deleted. To make it easier for the students the first half of the deleted word can be given.

A variety of texts are recommended for this technique. “Large number of items that can be generated on small texts further enhances the representative nature of the language being sampled” (Weir 1990:49).

Another favorable thing about C-tests is objective scoring. It is not very probable that there could be more than one correct answer for any of the gaps. Beside that, the C-test technique is economical and the results are reliable and valid. As Weir presumes, it could represent a viable alternative to cloze procedure and selective deletion gap filling (Weir 1990:49).

Unfortunately, there is only a little empirical evidence of the value of this technique since it appeared relatively recently. We do not know yet if the public will accept it as a measure of language proficiency. Moreover, students find this technique irritating for they have to process heavily mutilated texts and therefore and the face validity is not very high then (Weir 1990:49).

3.3.3 Cloze elide

On the contrary to the three preceding testing techniques, where candidates were to fill in deleted words, in cloze elide candidates are required to find the words which do not belong to the text. In other words, words that do not belong to the text are inserted into it and the testees have to indicate them. It must be said, that this is not a new technique; Davies was using it much earlier only it was known as the intrusive word technique (Weir 1990:50).

When we compare it to multiple-choice or short answer testing technique, we realize that the reader does not have the problem with understanding the question here. It certainly is an advantage. On the other hand, there can be a problem in scoring for the reader may delete items that are correct but redundant (Weir 1990:50).

; they helped each other to be understood. Now, nevertheless, in a testing situation, the original text can become much harder if not impossible to understand for the reader since the information from the graphic test is missing. In such a case the test constructor should consider adding some information to the original text to make sure the information necessary for completion are clearly stated therein (Alderson 2001:248).

In conclusion Weir recommends using short answer questions together with selective deletion gap filling for testing reading comprehension. He explains:

> The C-test is an interesting alternative to the latter and its acceptability to students and validity are worthy of further investigation. If we are to develop the

communicative nature of our tests it is perhaps important to focus on performance tasks in reading tests, and the use of information transfer techniques and other restricted response formats is advocated.

3.2.8 Multiple matching

One of the objective techniques testing reading is, according to Alderson, multiple matching. Two sets of stimulation are being matched against each other, for instance headings for paragraphs to their corresponding paragraphs, titles of books, extracts from each book or others. However, it can be argued that in matching as well as in multiple-choice exercises, candidates may be distracted by choices they would not otherwise have considered (Alderson 2001:219).

3.3.4 Ordering tasks

As for the ordering tasks, testees are given a scrambled set of word, sentences, paragraphs or texts and have to put them in the right order (Alderson 2001:219). Even though such tasks look temptingly and seem to offer the possibility to test the ability to detect cohesion, overall text organization or complex grammar, they are eminently difficult to prepare. Alderson shows it on an example where ordering though different from the original text can be acceptable (Alderson 2001:221). He also brings up a question of partially correct answers and difficulty of evaluating them and concludes that the effort made in constructing as well as in answering the item may not be worth it, especially if only one mark is given for the correct version (Alderson 2001:221).

3.3.5 Free-recall tests

In free-recall tests candidates are asked to read a text, then put it aside and then write down everything they remember from the text. These tests are sometimes called immediate-recall tests and are examples of what Bachman and Palmer call an extended production response type (Alderson 2001:230). The advantage of this technique, as Alderson explains, is primarily that it provides a pure measurement of comprehension. No questions intervene between the reader and the text. Alderson presents Bernhardt's opinion that this technique provides a picture of learner processes: "Recalls reveal information about how information is stored and organized, about retrieval strategies and about how readers reconstruct the text" (Bernhardt in Alderson 2001:230). At this

point, however, it is essential to say that the recall needs to be in the first language; otherwise it becomes a test of writing as well as reading.

A more familiar alternative of the free-recall test is the summary. Students are asked to read the text first and then to sum up the main ideas. As Alderson states, it is believed that students need to understand the main ideas of the text and to distinguish between important and less important ideas to be able to summarize it. As Alderson states the problem with summary tests is that students may understand the text, but may not be unable to express their ideas in writing adequately, especially within the time available for the task (Alderson 2001:236). In other words, "summary writing risks testing writing skills as well as reading skills" (Alderson 2001:236). One solution, Alderson suggests, could be allowing test-takers to write the summary in their first language rather than the target language or to offer multiple-choice summaries, where the reader chooses the best summary of the given ones.

Clearly, scoring the summary test is problematic. In some cases marking includes a scheme where main ideas get two points and subordinary ideas one point. Another way of making the scoring more objective is to let the test constructors write their own summaries and then accept as the main ideas only those written by an agreed proportion of respondents. "Experience suggests, however, that this often results in a lowest common denominator summary which may be perceived by some to be less than adequate" (Alderson 2001:233).

Alderson claims that: "One way of overcoming both these objections to summary writing is the gapped summary" (Alderson 2001:240). Students read the text, then a summary of the same text but with deleted key words. They are to fill in the missing words. The condition is that the missing words cannot be restored without reading the text and without understanding the main ideas of the original text. As for the scoring, Alderson remarks that it is relatively straightforward and the risk of testing readers' writing as well is unlike with short-answer questions out of question here.

Alderson concludes that such tests are difficult to write and need much pretesting, but can eventually work well and are easier to mark (Alderson 2001:242).

3.3.6 "Real-life" methods (the relationship between text types and test tasks)

The disadvantage of all the methods discussed so far is that they bear little or no relation to the text whose comprehension is being tested nor to the ways in which

people read texts in normal life. Indeed, the purpose for which a student is reading the test text is simply to respond to the test question. Since most of these test methods are unusual in “real-life reading”, the purpose for which readers on tests are reading, and possibly the manner in which they are reading, may not correspond to the way they normally read such texts. The danger is that the test may not reflect how students would understand the texts in the real world (Alderson 2001:248-249).

We have already discussed the importance of purpose in determining the outcome of reading. Yet but still very often the only purpose our students see in reading, is to answer our questions to show they either understood the text or not. Therefore, the test constructors should be challenged to differ reader’s purposes by creating test methods that would be more realistic than cloze tests and multiple-choice formats (Alderson 2001:249). Short answer question are closer to our lives, although we usually do not answer someone else’s question about our reading, we do make and answer our own questions. That is why Alderson suggests (to test constructors before deciding what method to use) asking: “What might a normal reader do with a text like this? What sort of self-generated questions might the reader try to answer?” (Alderson 2001:249) He gives an example when the reader is given a copy of a television guide and is asked to answer “real-life” short answer questions, for instance: You like folk songs. Which program will you probably watch? or give the name of one program which will be televised as it happens and not recorded beforehand (se Alderson 2001:250). Alderson claims that we should try to match test tasks to text type in an attempt to measure “normal” comprehension; in other words to devise tasks which more closely mirror “real-life” uses of texts (Alderson 2001:250).

Alderson believes that thinking about the relationship between texts and potential tasks is useful discipline for test constructors. It also presents possibilities for innovation in test design and measurement of reading. Consequently Alderson suggests that: “giving thought to the relationship between text and task is one way of arriving at a decision as to whether a reader has read adequately or not” (Alderson 2001:255). Earlier approaches to the assessment of reading did not pay much attention to the relationship between text and test question. “Most test developers probably examined a text for the ‘ideas’ it contained … and then used text content as the focus for test questions”

(Alderson 2001:255). A more recent alternative approach is, as Alderson presents, to decide what skills one wishes to test, select a relevant text, and then intuit which bits of the text require use of the target skills to be read.

> I suggest that a 'communicative' alternative is, first, to select texts that target readers would be plausibly read, and then to consider such texts and ask oneself: what would a normal reader of a text like this do with it? Why would they be reading it, in what circumstances might they be reading the text, how would they approach such a text, and what might they be expected to get out of the text, or to be able to do after having read it?

(Alderson 2001:256)

Such an approach has become increasingly common. Tests also include graphic texts – tables, graphs, photographs or drawings. The texts are taken from authentic, non-literary sources and are presented in their original length and format. As Alderson notes, "they often include texts of a social survival nature: newspapers, advertisements, shopping lists, timetables, public notices, legal texts, letters and so on" (Alderson 2001:256).

3.3.7 Informal methods of assessment

Until now we have discussed techniques that can be used in the formal assessment of reading. There are, however, other techniques that are frequently used in more informal assessment of a reader. These are suitable in the first place for those who are learning to read, those with particular reading disabilities or for students in adult literacy programmes. An extensive discussion of these informal methods of assessment is unfortunately beyond the scope of this work. (For more information see Alderson 2001:257-270)

All three completed out chapters: testing, reading and testing reading provide a theoretical base for the research which is described in detail in the following part.

4. RESEARCH

4.1 INTRODUCTION TO THE RESEARCH

I have been interested in the topic of reading since I have experienced very little of both teaching and testing reading comprehension during my own learning as well as during my whole clinical year experience. If a teacher wanted to give a mark from reading pupils were asked to read text (that they could prepare at home) aloud. Therefore, however, their ability to pronounce certain words was assessed. As Grellet states, he too, unfortunately came across teachers who thought that the way to test reading ability of their pupils is to let the pupil read aloud. He assumes that reading aloud is an extremely difficult exercise, highly specialized and it tends to give the impression that all texts are to be read at the same speed (Grellet 1991:10). He tries to explain that when we read, our eyes do not follow each word of the text one after the other – at least in the case of efficient reading. On the contrary, we skip a lot of words or expressions; we go back to check something, or forward to confirm some of our hypotheses. Such tactics become impossible when reading aloud, and this reading activity therefore tends to prevent the students from developing efficient reading strategies (Grellet 1991:10).

As it was stated in the theoretical part, first efforts to deal with complex reading skills often come not earlier than at the tertiary level (Heaton 1991:105) That is when the students have to cope with professional and technical literature in the foreign language. As I was given the opportunity to teach in sixth, seventh, eighth and ninth grades, I tried to evaluate my pupils' reading abilities. For this purpose we often used the magazine Rainbow that children like a lot.

The research focused on the pupils at ninth grade of elementary school. The group was chosen since it was the most suitable sample for our research. There was a variety of learners' levels of English with different reading experience and also a good atmosphere within the class. These children are from fourteen to fifteen years old, they attend a class specialized in sport, they have been learning English for almost six years (three 45 minute-lessons per week) and I have been teaching them since the last year. (They had three different English teachers before me.) There are thirteen of them, five girls and eight boys, and we meet on Wednesdays, Thursdays and Fridays the second or third lesson in the morning in their classroom, which gives my students the opportunity

to sit at one desk each (in test situations). The classroom is quite nice, roomy and light, with flowers and pleasing decoration; some of our English project posters are displayed there too.

The aim of the research was to find out and then evaluate how the pupils perceive some of the techniques testing their reading skills. (The criteria for choosing the particular testing techniques will also be described later in this part.) The pupils were asked to fill in three tests; each test contained of five testing techniques. The tests were based on three different texts; the first was an interview with a famous Czech actor Jiří Macháček, the second was a story from war called "How we met" and the last one was an article about koalas. When choosing the texts all the factors stated in the theoretical part were thoroughly considered as it is further explained.

As it was said many times before in the theoretical part, reading is a complex skill and reveals general knowledge of language. The test was announced to the students but in order to prepare themselves for the testing situation not to learn or revise specific part of grammar, for example. I believe that it is less stressful for the students when they expect the test, when it is not new, very often shocking, unpleasant news for them. They can then do much better in the test.

The tests were administered within three weeks in May. I assume it is important to add that at that time the pupils had already passed their entrance examinations at secondary schools. This fact could and very probably did influence their motivation. We wrote the tests on Wednesdays, discussed them on Thursdays and the Friday after the last test we had a discussion over all the tests but this time focusing on the reading techniques. The students had the opportunity to recollect their tests for ten minutes and then we started the discussion.

Before distributing the test I tried to motivate them, announced the time, summed up what they were going to do in the tests and suggested some ways to proceed. Maximum afford to create a positive atmosphere was made; we cleaned and aired the room, the pupils were asked about listening to music during the test. After distributing the tests to all students we went through the test together. I made sure they understood what they were supposed to do, checked all of them separately and helped with understanding the instructions to weaker students. The time was announced as

usually; that means when they were in half and then two minutes before the time was up.

After finishing each test pupils were to fill in a questionnaire, after finishing all three tests we had a short discussion about the tests, questionnaires and other aspects. With respect to the pupils' level of English the questionnaire was written and also the discussion lead in Czech language.

Although many interesting conclusions could be drawn out of our research, we will focus on the testing techniques – specifically, how the pupils perceive them. We will also compare the pupils' opinions to their test results.

The findings of the research will be used in my and my colleagues' further teaching.

4.2 Stages of our test construction

4.2.1 Setting the purpose

According to Hughes the essential first step in testing is to make perfectly clear what a teacher wants to find out and for what purpose. The questions he points out are stated earlier at this work, the chapter *Stages of test construction).* They were taken as a hint before our tests were being constructed. All our tests were final achievement tests. As it is stated earlier, achievement tests measure development in mastering particular skills. They are administered at the end of a course of study. In our case they revealed how students master reading at the end of the elementary school. Concerning types of testing; As Hughes states, while direct testing intends the candidate to perform precisely the skill we wish to measure, indirect testing attempts to measure abilities that underlie the skills (Hughes 2002:15). Integrative testing requires the testee to combine many language elements. While norm-referenced testing compare each student with his classmates, criterion-referenced testing rate students against certain standards (Hughes 2002:18). The testing is objective when no scorer's judgment is required. Our testing was indirect, integrative and criterion-referenced. Concerning subjectivity and objectivity, there were parts in the tests that could be assessed objectively (T/F statements, for example) as well parts that needed a subjective judgment (short answer questions, for example).

4.2.2 WRITING SPECIFICATIONS

After the first step is done (setting the purpose), the next step should follow. As Hughes recommends, at that moment testers should write a set of specifications for the test; that means information relevant to content, format and timing, criterial levels of performance and scoring procedures (Hughes 2002:48). As the objective of our research is to assess the difficulty of particular testing techniques, we will start this chapter by a closer look at those.

4.2.2.1 Testing techniques

Before we start describing individual test techniques it is important to say that there is no one "best method" for testing reading. "No single test method can fulfill all the varied purposes for which we might test" (Alderson 2001:203). However, as Alderson wants to emphasize, certain methods are common solely for reasons of convenience and efficiency, often at the expense of validity. "And it would be naive to assume that because a method is widely used it is therefore 'valid'" (Alderson 2001:204).

> It is now generally accepted that it is inadequate to measure the understanding of text by only one method, and that objective methods can usefully be supplemented by more subjectively evaluated techniques. Good reading tests are likely to employ a number of different techniques, possibly even on the same text, but certainly across the range of texts tested. This makes good sense, since in real-life reading, readers typically respond to texts in a variety of different ways.
>
> (Alderson 2001:206)

When considering what techniques should be involved into our research, the survey of test formats displayed in the *Assessment* by Harris and McCann, 1994, was very helpful to me. Pros and cons of different test formats are compared there (p.36). It is a very well arranged table. However, more information was needed to decide what the most suitable test formats for our purpose would be. The most consulted sources were Alderson, 2001 and Weir, 1990.

4.2.2.1.1 Cloze

To understand the technique called *selective deletion gap*, it is necessary to first learn about technique called *cloze*.

In the cloze exercises some words are deleted from a text. Usually a few sentences in the beginning of the text are without missing words to enable the reader to get involved into the text. Candidates are to fill in the missing words. The words are

deleted regularly; the deletion rate, as Weir refers to it, is every fifth to eleventh word. According to Alderson it is every n-th word, when *n* is usually a number between 5 and 12 (Alderson 2001:207). In other words, the deletion rate is mechanically set (Weir 1990:46).

Engineer compares cloze and multiple-choice testing technique and concludes that they are measuring different aspects of the reading activity (Engineer in Weir 1990:46). While a timed cloze measures the process of reading (the ability to understand the text while reading it), a multiple-choice measures the product, "namely the reader's ability to interpret the abstracted information for its meaning value" (Engineer in Weir 1990:46).

Weir uses a few authors and their quotations to support the idea of using cloze test procedures:

> Up to now, in the main, the results of research with cloze tests have been extremely encouraging. They have shown high validity, high reliability, objectivity, discrimination and so on.
>
> (Klein-Braley in Weir 1990:46)

> As demonstrated in this and other studies, it can be a valid and reliable test of overall second language proficiency.
>
> (J.D.Brown in Weir 1990:46)

> The last decade, in particular, has seen a growing use of the cloze procedure with non-native speakers of English to measure not only their reading comprehension abilities but also their general linguistic proficiency in English as Foreign Language. … The general consensus of studies into and with cloze procedure for the last twenty years has been that it is a reliable and valid measure of readability and reading comprehension, for native speakers of English. … As a measure of the comprehension of text, cloze has been shown to correlate well with other types of test on the same text and also with standardized testing of reading comprehension.
>
> (Alderson in Weir 1990:46)

W.L.Taylor first introduced the term *cloze* in 1953. He took it from the gestalt concept of *closure* which refers to the tendency of individuals to complete a pattern once they have grasped its overall significance (Weir 1990:46). Taylor describes it as follows:

> A cloze unit may be defined as: any single occurrence of a successful attempt to reproduce accurately a part deleted from a "message" (any language product), by deciding from the context that remains, what the missing part should be.
>
> ***(Taylor in Weir 1990: 46)***

According to Alderson the reader comprehends the mutilated sentence as a whole and completes the pattern. He adds: "the cloze procedure becomes a measure of the similarity between the patterns that the decoder is anticipating and those that the encoder had used" (Alderson in Weir 1990:46).

Taylor first applied the cloze procedure to assess the readability of a text. Later, however, it became a measure of testing reading comprehension and even a measure of overall language proficiency. Heaton thought that cloze tests measure the reader's ability to "decode interrupted or mutilated messages by making the most acceptable substitution from all context clues available" (Heaton in Weir 1990:47).

Weir points out following advantages of cloze testing technique:

- Cloze tests are easy to construct and score.
- They are claimed to be valid indicators of overall language proficiency.
- With a fifth word deletion a large number of items can be set on a relatively short text.
- Cloze tests are often considered to be valid and uniform measures of reading comprehension.

(Weir 1990:47)

Despite the arguments adduced in favour of cloze testing technique, a number of doubts have been expressed. Weir mentions the following ones:

- The students find cloze tests irritating and unacceptable.
- The specialists doubt the underlying assumption that it randomly samples the elements in a text.
- Concerning construct validity, it was found that cloze tests fail to ensure random deletion of elements in a text.

Cloze procedure, according to Alderson, is not a unitary procedure:

> since there is a marked lack of comparability among the tests it may be used to produce. The fact emerges clearly that different cloze tests, produced by variations in certain of the variables, give unpredictably different measures, particularly of proficiency in English as a foreign language.

Weir adds that if one changes the text, the deletion rate, begins at a different place or modifies the scoring, then one gets a different test in terms of reliability, validity and overall test difficulty.

- The evidence about the differing scoring methods is contradictory.
- Cloze procedure may seem to produce more successful tests of syntax and lexis at sentence level than of reading comprehension in general or of inferential or deductive abilities, what Darnell calls higher order abilities. Alderson describes his findings concerning that as follows:

> cloze is essentially sentence bound. … Clearly the fact that cloze procedure deletes words rather than phrases or clauses must limit its ability to test comprehension of more than the immediate environment, since individual words do not usually carry textual cohesion and discourse coherence (with the obvious exception of cohesive devices like anaphora, lexical repetition and logical connectors).

- According to Weir, the most crucial qualification against cloze tests is the question of what performance really tells us about a candidate's language ability. "It is difficult to translate scores on a cloze test to a description of what a candidate can or cannot do in real life."

(Weir 1990:47-48)

4.2.2.1.2 Selective deletion gap filling

When the test constructor chooses items for deletion, we then talk about selective deletion gap filling. Alderson describes this testing method as an alternative technique (to cloze) for those who wish to know what they are testing (Alderson 2001:209). Linguistic reasoning is used to decide which items should be deleted. Therefore, it is easier to state what is the aim of each test; in other words, what the test is intended to measure. Support for this technique is increasing especially after recent negative findings on mechanical deletion cloze (Weir 1990: 48).

As it was previously mentioned, one of the advantages of selective deletion gap filling is that it enables the test constructor to determine where deletions are to be made and to focus on those items which have been selected a priori as being important to a

particular target audience (Weir 1990:48). Another benefit of this testing technique is relative simplicity for the writer to alter the test items after analyzing them.

Weir stresses that this technique restricts one to sampling a much more limited range of enabling skills than do the short answer and multiple-choice formats. “If the purpose of a test is to sample the range of enabling skills including the more extensive skills such as skimming, then an additional format is essential” (Weir 1990:48).

A problem with this test format Alderson points out is that the test constructor knows which words have been deleted and so may tend to assume that those words are essential to meaning. Therefore, as he believes, pre-testing followed by careful analysis of responses is necessary (Alderson 2001:210).

Concerning scoring which was said to be very clear and easy to prepare, Alderson adds that in some scoring procedures, credit may also be given for providing a word that makes sense in the gap, even if it is not the word which was originally deleted (Alderson 2001:207).

Alderson offers a variant on both cloze and gap-filling procedures by supplying multiple choices for the students to select from. Two versions, as Alderson states, are common: one, when the options (usually three or four) for each blank are inserted in the gap and the students are to choose from these; and the other version, when the choices are placed after the text (either all together in one blank or separately grouped into fours and identified against each numbered blank by the same number). This cloze procedure is called “banked cloze” or “matching cloze” and is quite difficult to construct, since one has to make sure that a word which is intended to be a distractor for one blank is not a possible correct word for another blank. That is probably, as Alderson claims, one of the reasons why many test designers prefer the variant mentioned earlier in this paragraph where the options are presented with each blank. (Alderson 2001:210).

For our purpose we used gapped summary to avoid the possibility that pupils would remember the certain words after reading the completed text first. (They had to have the completed text for they needed it for other testing techniques.)

4.2.2.1.3 Short answer questions

Short answer questions require candidates to write down specific answers in spaces provided on the question paper. The technique is very useful especially for testing reading and listening comprehension.

Here are some advantages of this testing method:

- in comparison with multiple-choice questions, answering correctly shows that the reader really understands the text; he or she cannot use their guessing skills
- if the question is formulated carefully, candidate's response can be brief and so more questions can be asked and broader view covered
- if the number of acceptable answers is explicitly stated, it is then possible to give the examiners quite precise instructions how to mark the test
- the right answer must be sought in the text; not just being one of those provided like in activities such as inference, recognition of a sequence or comparison

However, there also are certain disadvantages of short answer questions format. Probably the main one is that it requires the candidate not only to read but also to write. So it can happen that the student understands the text but is not able to answer the question, to construct a sentence. Weir discusses the importance of limiting the possible acceptable responses together with the extent of writing required. At that point he also remarks that the number of correct answers can lead to unreliability of those who score the tests.

4.2.2.1.4 Dichotomous items

In regards to dichotomous items, this testing technique is very popular, mainly because the construction is not very difficult. (In our case, however, only pre-testing revealed that one of the T/F statements was not proper since the wanted information was not explicitly stated in the text.) Students are to decide whether the statement related to the text is "true" or "false". An obvious disadvantage of this test format is the 50 per cent chance of getting the answer right by guessing alone. To counterbalance this, it is necessary to have a large number of such items. Sometimes, to reduce the possibility of guessing, an option "not stated", "not given" or "the text does not say" is

added. Though these extra options, especially when used with items intended to test the ability to deduce meaning, may lead to significant confusion (Alderson 2001:222).

4.2.2.1.5 Information transfer

The problem of involving writing in the testing of reading comprehension has been discussed earlier. Weir describes that in an attempt to avoid this, several Examination Boards in Britain came up with tasks where the information transmitted verbally is transferred to a non-verbal form, for instance by labeling a diagram, completing a chart or numbering a sequence of events (Weir 1990:50). Alderson expounds this testing procedure as follows:

> The student's task is to identify in the target text the required information and then to transfer it, often in some transposed form, on to a table, map or whatever. Sometimes the answers consist of names and numbers and can be marked objectively; other times they require phrases or short sentences and need to be marked subjectively.
>
> (Alderson 2001:242)

The information transfer technique is particularly suitable for testing an understanding of process, classification or narrative sequence. (That also was the reason why this technique was chosen; it suited to our texts.) It is a realistic task for different situations so its interest and authenticity gives it high face validity. For some students, however, non-verbal tasks can be difficult and even though they understand the text they do not have to be able to understand what is expected from them in the transfer stage. Also, there is a danger of cultural and educational prejudices, so some students may have a disadvantage (Weir 1990: 50). Alderson explains this problem by giving an example, where a candidate may be asked to read a factual text and then to identify relevant statistics missing from a table and to add them to that table. However, when the student is not familiar with this kind of presentation of statistical data, the task is extremely difficult for him or her to do. "This may be more an affective response than a reflection of the true cognitive difficulty of the task" (Alderson 2001:248). On the other hand, it can be claimed, that since we all have to deal with such tasks in real life they should not be excluded from testing experience. Furthermore, such tasks indicate the validity of the test.

A possibly related problem, Alderson gives, is that tasks can be too complicated. The reader then has to spend a lot of time and effort on comprehension of what is required from him or what should go where in the table. In other words, Alderson

remarks, “the information transfer technique adds an element of difficulty that is not in the text” (Alderson 2001:248).

> Test constructors very often take graphic texts already associated with a text (for instance: a table of data, a chart or illustration) and then delete information from the graphic text. These two texts (the original and the graphic) were complementary at the beginning
>
> (Weir 1990:51)

At this phase of the test construction the answer key was prepared. We did our best to state as many acceptable responses as we could think of.

4.2.3 WRITING THE TEST AND MARKING

Writing the test includes three areas: 1) sampling, 2) item writing and moderation and 3) writing and moderation of scoring key. All these aspects were born in mind when preparing the test. Also some of the critical questions Hughes (2002) suggests were considered (see our chapter 1.6.3).

Even though marking has been introduced as the last step of the test construction in the theoretical part of this work, we will insert information concerning marking of our tests at this place. The reason is that the marking was considered in such order in our test construction; it was due to Harris’ advice to think of it already at the initial phases for otherwise, it can be the most time-consuming part of the construction.

Last year during my clinical experience I did a little research into marking system of my colleagues. As I had no experience I asked them how they evaluate their pupils performance. Obviously each test requires specific marking scheme, however, for some language tests assessment expressed in percentage can be applied. Unfortunately, I did not find their marking scheme suitable for me – it is too strict and it can discourage the students, in my opinion. That is why I adjusted the scheme according to my experience and needs and used it during the second year of my teaching. The reading tests presented in this paper were marked according to this scheme as well.

My colleagues´ scoring		**My scoring**	
100% - 91%	1	100% - 90%	1
90% - 78%	2	89% - 75%	2
77% - 65%	3	74% - 55%	3
64% - 50%	4	54% - 35%	4
49% - 0%	5	34% - 0%	5

4.2.4 Pre-testing

As it was stated in the theoretical part pretesting is a very important and useful stage of test construction. The aim of pre-testing is to identify possible problems of the test. As Hughes remarks, even after careful moderation some problems might appear. Therefore, if it is possible, the test should be tested first on a different but similar group. That is why our test was pre-tested on one of my colleague's group before using them with my target group. They were pupils from ninth grade as well, they were eleven of them, they wrote the tests on Tuesday, the third lesson in the morning in an English classroom where they spend most of their English lessons. Each student chose one of the three tests and had 45 minutes to work on it. After finishing the test they were answering the questions from the questionnaire. The questionnaire was constructed in Czech with respect to their knowledge of language. During their work I checked their understanding of instructions and helped to weaker students. They also chose to work while listening to music they agreed on. I believe maximum effort has been invested to motivate the students to do their best and to provide both pleasant and working atmosphere in the class.

The pre-testing revealed several facts:

- questionnaire was too long; two questions were misunderstood; a sign that it is double-sided needed to be added; there was not enough space for response
- completion of the information in the table was not explicitly instructed
- one T/F statement was tricky (the information was not stated explicitly in the text)
- drawing in the test about koalas was too time-consuming and for some students too difficult
- two items that were to fill in the gapped summary could be done without understanding the text

As it is seen pre-testing had revealed some very important facts. Therefore necessary changes were made and the tests were then used for the target group of our research (see appendices).

4.3 Test analysis

Concerning the texts, they were chosen after thorough consideration of all related aspects that were mentioned in the theoretical part, in the chapter 3. Testing Reading. There we mentioned factors that affect the difficulty of reading tests items and reading test texts. At the first place we tried to look for the tests whose theme would interest my students. Then the role of vocabulary and grammar, the genre together with the text length was considered. In case of the first test, however, the text had to be shortened. As we learned in the chapter 3.2.7 (Text readability and text simplification), it could be taken as a kind of simplification of the text since the more difficult parts were skipped. Even though we were aware of the risk of reverse effect, the risk that the test will be made even less readable, yet we decided to shorten the text. High attention was paid not to disrupt important relationships in the text; moreover a native speaker agreed on the shortened version also in terms of cohesion and coherence. All the three texts that were chosen in the end were already accompanied by one or two pictures, which appeared to be very useful in terms of saving time for other stages of the tests construction. The need for both verbal and non-verbal or graphic information in text was discussed as one of the factors affecting reading difficulties (for more detailed information see 3.2.9 Verbal and non-verbal information).

Recalling Alderson's warning concerning typographical features we tried to prepare the layouts nice and legible (see 3.2.8 Typographical features).

To the question of presence of the text during the test situation, we decided to allow the presence of the text while working the test out. As Alderson states removing the text increases the role of memory in the readers' responses although not in the comprehending process (Alderson 2001:106). He further admits that both variants have their advantages and disadvantages and gives advise to let answer the central questions without the presence of a text while peripheral questions should be answered in the presence of the text (Alderson 2001:109).

Aspects of text type and genre, organization of the text as well as the issue of weighing difficulty of the text and test items will be discussed in the specific analysis of each test.

The tests were administered in the following order:

1. Interview with Jiří Macháček
2. How we met
3. Koalas

The reason for that order was the assumption that the students will enjoy the text, an interview with their favourite actor Jiří Macháček, and therefore they will be motivated to read it in order to get new or funny information. Also a pleasant experience with the first test would be important for another testing situations.

INTERVIEW WITH JIŘÍ MACHÁČEK

The text was taken from an English magazine Rainbow (R&R), number 2 from October 2003. Originally, the interview was longer; for our purpose, however, it was shortened as it was explained earlier in the research part of the thesis. I chose it for I thought the pupils would like to get to know more information about their favourite actor; also the text type, interview, should be motivating for the students since it is close to real life. Other aspects (like role of vocabulary, grammar, etc.) were considered as well, of course, as it is mentioned earlier in this part.

To reduce the effect of vocabulary knowledge on measures of reading comprehension, Alderson suggests allowing students use dictionaries. Other constructors of the test, however, claim that it invalidates the test since the dictionaries provide some of what is being tested (Alderson 2001:99-100). Moreover, as I can also add from my own teaching experience, most of the students would waste time looking up words that should be rather spent on reading the text. Therefore my students were not allowed to use dictionaries. They could, however, use a box with some words they might not know and would need them for better understanding. This, in my opinion, could solve the problem of using dictionaries in some measure at least.

Although the results of the pre-tested group were very good, my students found this test very difficult. They stated it in the questionnaire, confirmed it in the discussion and the results showed it too (see p.93 and 94). The students claimed that even though the text was interesting for them they lost their motivation since there were too many unknown words for them and they got easily lost. Another thing some of them agreed on was not a very clear distinction between the question of the interviewer and Jirka's answer. I should have added, for example: I. for the interviewer's part and JM for

Jirka's part. The worse results from the three tests could also be caused by not knowing such a testing situation – having one text and five testing techniques to do. It took a long time to get to understand what they were supposed to do. However, the main reason for my students´ failure, in my opinion, was that as they had two substituted lessons before our lesson (they watched video with the first teacher, went out with the second teacher) they were not in the mood to write a test with me at all. The average successfulness was only 40%. Therefore the results from this test were not inserted into our little research of testing techniques since I believe the outcome would not be objective.

HOW WE MET

This text was taken from an English textbook (Headway, elementary, workbook). The story is touching and the pupils like to read people's life stories. As questionnaires, discussion and the results showed the story about two young people who fell in love during the war took my pupils' interest. However, I did not expect them to find this text the most interesting of all the three. They stated that it was quite simple, not too long and the words they did not understand were either easy to guess or deduce from the context or not essential for general understanding of the text.

All this is reflected in the results; they were much better than the results from the first test. While the average successfulness in the first test was only 40%, now it was 83%. I suppose that another component affected such good results; the pupils knew already what the test would look like, what they were going to do; they had experience from the previous test and could become more familiar with the test format. As they stated in the discussion, the second test was easier for them also for they knew what was expected from them, they could better lay out their time and had no trouble with the instructions. What is important to emphasize is that it proved the necessity of not only interest and motivation to obtain good results from a test but also knowing the test format. It can significantly impact the pupils´ results.

CONCERNING INDIVIDUAL TESTING TECHNIQUES, THE RESULTS SHOWED THAT THE PUPILS WERE MOST SUCCESSFUL IN TABLE COMPLETION AND IN GRAPHIC INFORMATION TRANSFER. GAPPED SUMMARY AND QUESTIONS, ON THE OTHER HAND, WERE THE MOST DIFFICULT FOR THEM. NOT ONLY THE RESULTS SHOW IT BUT

ALSO THE PUPILS CONFIRMED THIS FACT BOTH IN THEIR QUESTIONNAIRES AND DURING THE DISCUSSION. THEY WERE ALSO VERY GOOD AT SELF EVALUATION; THERE WAS A QUESTION IN THE QUESTIONNAIRE ASKING FOR THEIR GUESS ABOUT THEIR SUCCESS IN PARTICULAR TEST FORMATS. THEY WERE TO ASSESS WHICH TEST FORMAT SUITED THEM BEST AND IN WHICH THEY EXPECT THE BEST RESULTS. THEY WERE ALSO ASKED TO EVALUATE WHAT TEST FORMAT WAS THE MOST DIFFICULT FOR THEM. AS IT WAS SUPPOSED THE TEST FORMAT THAT THEY WERE GOOD AT WAS ALSO THE ONE THEY MOST ENJOYED DOING.

KOALAS

This text was also taken from the magazine R&R (March 2004, number 7, section *Unique animals of Australia*). We have been working with this magazine for the second year and the students like the animal section very much. To compare it with the two previous tests; the pupils enjoyed reading about koalas, the text was longer than the war love story but shorter than the interview with Jiří Macháček. There were some unknown difficult words but not as many as in the interview. Since the students knew the lay out of the test perfectly they went through it quite easily and quickly. The average successfulness of this test was 78%.

This time there was a slight difference in the results. Graphic information transfer, table completion and questions were the most successful test formats. The pupils scored 90%. Gapped summary, however, was perceived as the most difficult again.

4.4 Conclusion of the research

The pupils´ individual results from the three tests are compiled in the tables attached in the appendices (see p.93). Even though there are thirteen pupils in my class one was absent for two lessons so I did not include him into this short research.

The purpose of each test was to find out about reading comprehension of pupils leaving elementary school after five years of English lessons three times a week.

The research aimed at comparing successfulness, difficulty (viewed by the pupils) and popularity of the tested techniques. The pupils viewed the gapped summary

as the most difficult technique and they also had the lowest score in it. No one stated this technique as the one he or she enjoyed doing. On the other hand the information transfer in form of graphic representation was the most successful and popular testing technique (for both reasons – the pupils enjoyed doing it and they found it quite easy). The table *Testing techniques – pupils' results* on p.95 reveals not only the most and least successful testing technique but also the results of other testing formats. We can see that completing a table was the second most successful technique and based on pupils' opinions it was also the second easiest and the second most popular. A relatively high number of students see T/F statements as rather easy to answer since, as they expressed in the discussion, they do not have to come up with words or formulate phrases; they just need to recognize the right answer and then circle it. As mentioned in the theoretical part, there is also a high chance of guessing instead of finding the right item (and my pupils used this strategy quite a lot as they confirmed during the discussion).

The table *Testing techniques – students successfulness* on p.95 also reveals that the individual student' results from a particular test format do not differ significantly; in other words, the results of certain testing techniques from both tests are very similar (60% in gapped summary and over 90% in information transfer, for example).

I believe that the results would be very similar in testing reading comprehension in students' mother tongue. In the cloze or deletion exercises students have to more than understand the text, they also have to come up with new words. Concerning questions they sometimes can be not very clear to the students; they might not understand what they are asked about. T/F statements can be very tricky and than it is hard to decide if we have only too options. Completing a table or drawing a picture after reading the text can also remind the testees more of a play than of a testing situation. They might feel less stressed and that can have a positive impact on their results.

To conclude, although the perception of different testing techniques is very individual, some of them can be considered to be more while others less difficult. In our research most of the students found the gapped summary the most difficult testing technique to answer and on the other hand information transfer in form of graphic representation was the least difficult for them.

CONCLUSION

The thesis focused on testing reading abilities. It was divided into two main parts: the theoretical and practical. In the theoretical part three central chapters were discussed: testing, reading and testing reading. In the practical part the research aimed at an assessment of chosen testing techniques from the pupils´ point of view was described. The objective of this work was to assess difficulties in testing techniques (as the pupils see it) and compare it with the pupils´ test results.

In the first chapter of the theoretical part testing in English as a second language is emphasized. The test characteristics such as validity, reliability and efficiency are described here. Types of tests, types of testing and individual stages of test construction are introduced in this chapter as well.

In the following part we are primarily concerned with reading skills and with the subskills involved in reading. Reading is introduced by means of definitions and answers to questions such as: What do we read?, Why do we read? And how do we read? This part is also devoted to reading techniques.

The last chapter of the theoretical part is devoted to different factors affecting reading tests. Significant space is given to aspects of text as such. Issues like function, organization, type, genre or typographical features of the text are discussed at this point. As the objective of the paper is to evaluate them, the testing techniques fill considerable space in our work too.

The practical part introduces the conditions under which the research was done. It closer explains the techniques that were chosen for our research. It further attempts to analyse the output of individual tests, relates them to the theoretical base and makes conclusions. The research was aimed at comparing successfulness, difficulty (viewed by the pupils) and popularity of the tested techniques. The pupils viewed the gapped summary as the most difficult technique and they also had the lowest scores in it. On the

other hand the information transfer in form of graphic representation was the technique the pupils were most successful at.

As it was stated at the very beginning of this paper testing is an important part of teaching and learning process; it cannot be separated. It helps both students and teachers. The washback effect should also not be ignored.

Moreover, we should consider many aspects before preparing a reading test; they were discussed earlier in the paper. We should also try to make our tests as reliable, valid and efficient as possible. If feasible we should pretest our tests since it has proved to be very helpful. We should also be aware of the undesirable psychological states brought about by stress and seek out ways to reduce it.

Concerning reading in particular the fact that it is an active communication skill which involves many specific subskills should be considered as well – both in teaching and testing. This paper offers many techniques that can be used to test reading skills. However, no best method can be recommended. Each teacher should choose the technique that best suits his or her students´ needs and requirements.

RESUMÉ

Tato diplomová práce se zabývá problematikou jazykových testů zaměřených na řečovou dovednost čtení. Je rozdělena na dvě hlavní části: část teoretickou a praktickou. Teoretická část obsahuje tři klíčové kapitoly: „Testování“, „Řečová dovednost čtení“ a „Testování řečové dovednosti čtení“. V praktické části je analyzován výzkum týkající se žáků deváté třídy základní školy a jejich chápání obtížnosti testovacích technik představených v části teoretické. Cílem předkládané práce je pokusit se zhodnotit obtížnost jednotlivých testovacích metod a porovnat ji s úspěšností žáků v jejich řešení.

V první kapitole teoretické části práce je kladen důraz především na vymezení významu testování čtení ve výuce anglického jazyka. Jsou zde představena hodnotící kritéria testů, typy testování, typy testů a fáze procesu jejich tvorby.

V následující kapitole teoretické části práce se věnuji řečové dovednosti čtení jako takové. Představuji tuto dovednost pomocí definic a odpovědí na otázky typu: „Co čteme?“, „Proč čteme?“ a „Jak čteme?“. Pozornost je zde věnována i technikám čtení a řečovým dovednostem čtení.

V závěrečné kapitole teoretické části jsou nejprve uvedeny jednotlivé faktory ovlivňující obtížnost testu jako takového, textu i jednotlivých položek testu. Důraz je v této kapitole kladen i na přiblížení jednotlivých testovacích technik čtenáři.

Část praktická v úvodu seznamuje čtenáře s prostředím a podmínkami, za kterých byl výzkum proveden. Dále pak analyzuje výsledky jednotlivých testů, vztahuje je k poznatkům z teorie a vyvozuje závěrečná zjištění.

Úvodní kapitola teoretické části, „Testování“ (Testing), je dále rozčleněna do šesti podkapitol. První z nich poukazuje na úzký vztah procesu učení a testování. J.B. Heaton jej hodnotí takto: „testování a učení (vyučování) je vzájemně tak blízce propojeno, že je prakticky nemožné pracovat v jedné oblasti aniž bychom se neustále zajímali i o tu druhou“ (vlastní překlad, Heaton, 1988:5). Jev nazývaný Backwash (do

češtiny se nepřekládá) s výše zmíněným vztahem velmi úzce souvisí. Jde totiž o vliv, který má testování na učení a vyučování. Tento vliv může být jak pozitivní tak i negativní. Dále je pozornost věnována důvodům, proč vůbec testovat. Je zde zdůrazněno, že testování pomáhá nejen testovaným, ale i testujícím. Testování kupříkladu pomáhá studentům orientovat se v jejich silných i slabších stránkách a dále se tak v jazyce zdokonalovat. Učitelé pak mohou z výsledků vhodně vytvořených testů čerpat informace, které následně využijí ve své práci.

Třetí část poukazuje na základní vlastnosti dobrého didaktického testu. Jsou zde vysvětleny pojmy jako validita, reliabilita a praktičnost. Ve stručnosti, je-li test validní, pak testuje výhradně a pouze to, co dle původního záměru měl. Odborníci uvádějí několik druhů validity. Tvrzení, že daný test je reliabilní, znamená, že je nejen spolehlivý (a to v tom smyslu, že by měl za stejných podmínek poskytovat stejné nebo velmi podobné výsledky), ale také přesný (při měření výsledků by nemělo docházet k velkým chybám). Učitel při vytváření testu zvažuje i jeho praktické výhody, jako jsou např. snadné použití či jednoduchá a rychlá oprava.

Ve své další fázi se tato diplomová práce zabývá rozdělením testů a typů testování. Arthur Hughes rozlišuje čtyři typy testů, a to podle využití jejich výsledků: testy úrovně (proficiency tests), testy výkonové (achievement tests), testy diagnostické (diagnostic) a testy zařazovací (placement tests). K čemu dané testy slouží, je zřejmé: testy úrovně měří jazykové dovednosti nezávislé na předchozí výuce, zatímco testy výkonové odhalují, jak žák zvládl konkrétní společně probrané učivo. Dělíme je na průběžné (progress achievement test) a výstupní (final). Chce-li učitel zjistit stav jazykových schopností žáků, použije test diagnostický, podle nějž pak upravuje další výuku. Zařazovacích testů pak využíváme při umístění studentů do nejvhodnější úrovně, např. do kurzů pro začátečníky, pro mírně pokročilé atd.

Z hlediska postupu při tvorbě testu rozlišujeme testování přímé a nepřímé, jednotlivé a integrující, z hlediska interpretace výsledků rozlišující a ověřující a vzhledem k míře objektivnosti hodnocení testy subjektivně a objektivně skórovatelné (Potůčková 2003: 67).

Poslední část kapitoly Testing je zaměřena na popis jednotlivých fází tvorby didaktického testu. Patří mezi ně stanovení účelu (cíle, záměru), upřesnění týkající se

obsahu, formátu, časového rozvržení, požadavků a způsobu hodnocení, dále pak sestavení testu, předtestování a známkování.

Druhá kapitola teoretické části nazvaná „Čtení“ nejprve vysvětluje, jak je tento pojem chápán ve vztahu k cizímu (druhému) jazyku. Porozumět psanému textu znamená, jak uvádí ve své definici Francoise Grellet, být schopen vybrat z něj (získat) požadované (potřebné) informace co nejefektivněji. Dodává, že při čtení je velmi důležité i odvozování neznámého, a také to, co čtenář sám do textu přináší. Proto, tvrdí, by studenti měli být od samého začátku vedeni k využívání toho, co již znají, aby pochopili neznámé, nové. V práci je zdůrazněno, že řečová dovednost čtení je komunikativní dovedností a neměla by být oddělována od jiných. Stejně jako jsou tyto dovednosti propojené v běžném životě, měly by být spojené jak při jejich učení tak při jejich testování. Grellet uvádí příklady, kdy na základě přečteného textu např. odepíšeme na dopis, rozhodneme se, vyřešíme problém, postupujeme podle návodu, řekneme o tom někomu dalšímu… Neodpovídáme na otázky, nevolíme mezi danými možnostmi jako při některých typech testových metod. V této části jsou řešeny i jiné otázky: Co čteme? Proč čteme? Jak čteme?

Následující kapitola popisuje techniky, které při čtení používáme a navrhuje strategie, jež by mohly být užitečné pro ty, kteří chtějí prohloubit svou schopnost četby s plným porozuměním. Záleží i na tom, za jakým účelem text čteme; zda se snažíme vyhledat jen určitou informaci, nebo jde-li nám o celkové pochopení textu. Grellet zde zdůrazňuje, že pojem čtení v sobě zahrnuje mnoho dovedností, např. odvozování neznámých slov, chápání větných vztahů, rozlišování podstatných informací od méně důležitých, vnímání signálů koheze, chápání komunikativní funkce textu atd. Značný prostor je zde věnován i kritice přístupu některých pedagogů k testování čtení; Grellet je přesvědčen, že hlasitým čtením nezjišťujeme, zda byl text pochopen, či ne. Hlasité čtení je velmi úzce specifikovaná dovednost, nepodává však informaci míře porozumění textu. Dále pak uvádíme dovednosti využívané při čtení.

Ve třetí, závěrečné kapitole teoretické části práce je pozornost věnována především problematice obtížnosti čtení a faktorům, které tento jev ovlivňují. Jednotlivé testovací techniky, kterých využíváme při zjišťování porozumění čtenému textu, jsou zde následně detailně popsány.

V této kapitoly teoretické části je pozornost věnována textu. Nejprve je zde vymezen rozdíl mezi pojmy *text* a *promluva* (discourse). Catherine Walace uvádí, že zatímco text je výstupem, výsledkem autorova snažení, promluva se týká více čtenáře (ne textu jako produktu) a soustředí se na proces rozkrývání významu ze strany čtenáře (Wallace 1992:8). Pochopení funkce a organizace textu považuje Grellet za velmi důležité a domnívá se, že by studenti měli být vedeni k jejich rozpoznávání od samého začátku. Může jim to velice pomoci při porozumění textu jako celku. Zatímco pochopení funkce a organizace textu je důležité, pochopení obsahu je zcela zásadní. Grellet navrhuje rozličné aktivity, jak žákům pomoci naučit se číst s porozuměním, a dělí je do dvou kategorií: 1) aktivity, které čtenáře „vtáhnou“ do procesu čtení, učiní ho aktivním (kupříkladu čtenář bude muset řešit nějaký problém, rozhodnout se atd.), 2) aktivity, které se co nejvíce podobají běžným přirozeným situacím.V samém závěru této kapitoly se pak zmiňujeme o možnostech přístupu k textu. Grellet rozlišuje přístup lingvistický a „nelingvistický“ (nelingvistické aktivity nevyžadují ucelenou slovní odpověď, doplňování diagramů, tabulek, porovnávání a přiřazování obrázků atd., zatímco lingvistické ano, pracuje se zde s částmi textu, se slovy, frázemi apod., které se kupříkladu doplňují do textu).

Na obtížnost testu má vliv nejen text, ale i jednotlivé položky testu, jako např. jejich jazyková obtížnost, typ otázek, svou roli zde hraje i mluvnice a slovní zásoba. Diskutujeme zde i o možnosti používat slovník, o spojitosti čtení a inteligence. J. Charles Alderson uvádí, že volba textu má zcela zásadní vliv na výsledky žáků. Domnívá se, že výběr testovací techniky tak důležitou roli nehraje. Na místě je jistě i zamyšlení nad otázkou, zda by žák měl či neměl mít text k dispozici po jeho přečtení (při odpovídání, vyplňování testu) a jak rozsáhlý by vlastně text měl být.

V kapitole „Testovací techniky“ se čtenář podrobně seznamuje s výhodami i nevýhodami jednotlivých technik[1] Jsou zde popsány následující techniky: testové úlohy s výběrem odpovědí (multiple-choice questions), otevřené otázky (short answer questions), doplňovací testy (cloze, selective deletion gap filling, C-test and cloze elide), přiřazovací úlohy (multiple matching), informační transfer (information transfer), uspořádací testové úlohy (ordering tasks), dichotomické testové úlohy (dichotomous

[1] Termíny *testovací technika*, *testovací metoda* a *testovací formát* jsou v práci používány synonymně.

items, T/ F statements) a testy požadující shrnutí, stručnou formulaci obsahu, výtah z textu (free-recall tests, summary, gapped summary).

Na základě prostudovaných informací týkajících se teoretické části problematiky řečové dovednosti čtení byl uskutečněn výzkum zaměřený na vnímání obtížnosti jednotlivých testovacích technik čtecích testů z hlediska žáků deváté třídy základní školy. Praktická část je rozdělena do čtyř kapitol.

V první je představen cíl výzkumu, shrnutí důvodů k tomuto výběru a stručný popis jeho průběhu. V další je pak čtenář podrobně seznámen s testovou situací; dozvídá se zde potřebné informace o testovaném vzorku respondentů a o jednotlivých fázích tvorby testu. Důležitou součástí této kapitoly je poměrně rozsáhlé seznámení čtenáře s těmi testovacími technikami, které byly pro výzkum vybrány. Je zde zmíněna i tvorba a úprava bodování a známkování testu. Dále je pak představen a zhodnocen jeden z kroků předcházejících samotnému psaní testu – předtestování. Ve třetí kapitole jsou vyhodnoceny jednotlivé testy.

Výsledky potvrdily hypotézu, že to, jak žáci vnímají obtížnost jednotlivých testovacích technik, se odráží i v úspěšnosti jejich řešení. Jinými slovy, žáci byli nejméně úspěšní při testovací technice, kterou v dotazníku a v následné diskusi označili za nejobtížnější. Naopak, nejvíce správných odpovědí dosáhli žáci v testovací technice, která pro ně byla nejjednodušší a zároveň je i nejvíce bavila.

V celkovém závěru se práce pokouší komplexně zhodnotit získané informace z obou částí. Z nich pak vyplývá, že není možné určit ideální metodu, kterou lze hodnotit jakýkoliv test. Navrhujeme zde, aby tvůrci testů zvažovali všechny potřebné informace na samém začátku vytváření testů, aby poskytovali testovaným různé možnosti reakcí, odpovědí a řídili se individuálním přístupem k žákům. Testovací úloha, která vyhovuje jednomu z nás (třeba právě a jenom učiteli) nemusí vyhovovat jinému. Proto bychom se měli pokusit, tak jako v celkovém přístupu k učení, nabídnout svým žákům co možná nejširší možnost volby; ta má na žáka ohromný vliv, umocňuje jeho motivaci a významně tak napomáhá procesu učení.