universität wien

DIPLOMARBEIT

Titel der Diplomarbeit

A comparative study on differences in language output between mainstream and CLIL students at two Austrian colleges of engineering, crafts and arts.

Verfasserin

Silvia Jexenflicker

Angestrebter akademischer Grad

Magistra der Philosophie (Mag. phil)

Wien, am 22.12.2009

Studienkennzahl laut Studienblatt:	A 343
Matrikelnummer laut Studienblatt:	8551209
Studienrichtung laut Studienblatt:	Diplomstudium Anglistik und Amerikanistik
Betreuerin:	Ao. Univ. Prof. Mag. Dr. Christiane Dalton-Puffer

Acknowledgements

The preparation of this study would not have been possible without the cooperation and generous support of a number of people to whom I would like to express my gratitude. I am particularly obliged to all those who agreed to take part in the tests, at both the trialling stage and the operational stage, as well as the staff at the colleges involved who made the tests possible. I am particularly indebted to Ms Veronika Schindelegger and Ms Nadia Wilhelmer, who administered the tests at the two colleges participating in the main study.

Moreover, I would like to express my gratitude to all the lecturers at the English Department at Vienna University whose teaching has provided me with the insights, knowledge and skills required to carry out this study. My special thanks go to my tutor, Prof. Christiane Dalton-Puffer, for her continued support and encouragement as well as her patience and understanding when progress was slow.

Finally, I would like to thank my mother for supporting me at times that were particularly stressful and my son Robert, who is my greatest source of inspiration.

Table of contents

Appendices:

List of Figures

List of Tables

1. Introduction: The purpose of this paper

Globalisation and internationalisation are making increasing demands on the foreign language skills of European citizens. In reaction to this, a trend has emerged in schools throughout Europe to use English (and other foreign languages) as a medium of instruction, not as an elitist project but also in mainstream education. In these so-called CLIL (Content and Language Integrated Learning) classes a language other than the L1 of the students is used in teaching a non-language subject matter, the aim being to increase the students' exposure to the language and to create a motivating, low-anxiety environment in which attention is paid to the message conveyed rather than the accuracy of the linguistic forms used. In this way the language competence of the students is to be enhanced and they are to be better prepared for life and work in a globalised society and economy, where English, in particular, dominates as the Lingua Franca of today's business world.

While the basic idea underlying CLIL, i.e. to provide students with more language input and thus to further their language proficiency, seems compelling, the question arises to what extent increased exposure translates into tangible improvements in the quality of language output and what aspects of language proficiency are most likely to be affected. The main aim of this paper is to investigate the impact of CLIL provision on the language output produced by students at two Austrian upper-secondary engineering colleges. For this purpose a test was designed and administered to two sets of students, i.e. those who have undergone CLIL instruction and those who have not. The aim was to see whether any differences could be observed in the quality of the language output of the two groups, the main focus being put on the students' general language ability as well as their written output.

The first part of the paper deals with basic issues in language testing. It first gives an overview of different purposes of language tests and key terminology used to describe different types of tests. Moreover, the key qualities tests should have, in particular validity and reliability, are discussed briefly before focusing on recent developments in language testing. This part charts key developments in the field of language testing from the 1960s until today, focusing on three main types of tests, i.e. discrete-point tests, integrative tests and communicative tests. With regard to the communicative approach to

language testing, a short overview is also given of the underlying concept of 'communicative competence' and its key components, referring briefly to the models designed by Canale and Swain as well as Bachman. Finally, the first part of the paper is completed with a discussion of the various steps to be taken in developing a language test, involving test design and construction, test trialling and finally, test administration. Together these constitute the 'testing cycle', a term used to emphasise the fact that information obtained at each stage feeds back into the test development process.

The second part of this thesis deals with Content and Language Integrated Learning (CLIL). First the term is defined and the relationship between the two main elements, i.e. content and language, is discussed. Following this, the rationale for CLIL is presented, focusing in particular on the claim that CLIL creates more natural conditions for the acquisition of communicative competence in the language used as medium of instruction. Involving students in discourse about subject-matter concepts is said to be more 'real' and meaningful than the activities students typically engage in in foreign language classes. Krashen's Monitor model, which is often referred to as the theoretical underpinning of CLIL, is sketched briefly in this context. Some qualifications to the claims made about CLIL are also introduced, however, based on research done on naturalistic classroom discourse within the context of CLIL lessons as well as the fact that the classroom setting itself puts certain limits on the development of communicative competence. Finally, reference is made to selected studies investigating the outcome of CLIL instruction, focusing on language ability.

While the first two parts are based on a review of the relevant literature, the third part presents a case study designed to show if those students who have undergone CLIL instruction at two Austrian upper-secondary engineering colleges outperform their non-CLIL peers in terms of general language ability and writing skills. First, the situation of language instruction at Austrian engineering colleges ('HTLs') is discussed briefly and the two participating schools are introduced. Following this, the paper focuses on the test administered, which consists of two parts, i.e. a C-test, which is an integrative test based on the cloze test, and a writing task in the form of a letter or e-mail sent to a prospective host family. The purpose of the C-test is to provide a single score which serves as a global measure of a test taker's general language proficiency. The design principle is explained and the various steps involved in test development are described including the selection of texts, the trialling stage consisting of two pilot tests and the final administration of the

test. Following this, the results are presented and the scores obtained by the CLIL and non-CLIL groups are compared and analysed with regard to the statistical significance of any differences observed.

While the C-test offers the advantages of objective assessment, it does not test communicative language ability as such. Therefore the second part of the test used in this case study is designed to meet the requirements of communicative language testing in so far as it puts test takers into an authentic communicative situation, asking them to write an e-mail or letter to a prospective host family in New York. The texts produced are assessed on the basis of an analytic rating scale consisting of the four sub-categories of task fulfilment, organisation, grammar and vocabulary. These categories are introduced and the results obtained analysed as to their statistical significance. In addition to this qualitative assessment in the fields of grammar and vocabulary a few quantitative ratios are also calculated to measure aspects of accuracy (e.g. error frequency) and complexity in the grammatical and lexical field. In addition to presenting the overall results, examples are provided from the texts produced by test takers in order to illustrate the main points.

Finally, based on a summary of the main findings, conclusions are drawn regarding the study itself and what it suggests about the strengths and weaknesses of current CLIL programmes. Naturally, the limited scope of this study makes any general conclusions difficult. Therefore reference is also made to the findings and suggestions put forward in a recent report to the Ministry of Education concerning the further development of CLIL in Austria.

2. Language testing: an overview of key issues

2.1. The purposes of language tests

Due to the fact that testing is "a universal feature of social life" (McNamara 2000: 3) with an important gatekeeping function, we all have been subject to it and have developed some intuitive notion of what a test, or, more specifically, what a 'language test' is. Generally speaking, the latter can be defined as

> a procedure for gathering evidence of general or specific language abilities from performance on tasks designed to provide a basis for predictions about an individual's use of those abilities in real world contexts (McNamara 2000: 12).

This basic definition already draws our attention to some key features of language, or indeed any kind of, tests. Note that McNamara clearly distinguishes between the test itself, which is an instrument that allows us to 'gather evidence', on the one hand, and the underlying abilities of the test taker, on the other. Clearly, it is the latter that we are interested in when administering a test. However, as these underlying skills do not lend themselves to direct observation, we need to use test performance in order to make inferences about how the test taker would perform in what is called the criterion situation, i.e. "tasks in the real-world setting of interest" (McNamara 2000: 8) which require the use of language. The main challenge in language testing can thus be seen in devising measurement instruments that help us form a clear picture of the 'true' abilities of the examinee, and to predict his or her performance 'in the real world', on the basis of language samples specifically gathered for this purpose.[1] This directly leads on to key issues in language testing such as the validity and reliability of inferences made on the basis of language tests. A brief account of these issues will be provided in Chapter 2.3. below.

Apart from this general purpose of language testing, different language tests may serve different purposes. At one end of the spectrum there are very specific, typically occupationally oriented, 'special purpose' tests administered to a very specific target

[1] What distinguishes tests from other forms of evaluation (e.g. continuous assessment in a language class on the basis of observation and verbal description, portfolio assessment etc.) is that they serve to elicit a specific sample of (language) behaviour and involve a process of quantifying the evidence obtained from it (cf. Bachman 1990: 20-23).

group and at the other end there are large-scale tests which try to establish the 'general language proficiency' of a large number of test persons. The purposes of language tests are thus manifold. One typical, and rather general, classification distinguishes between four types of tests, i.e. proficiency, achievement, diagnostic and placement tests (e.g. Hughes 2003: 8-17).

Proficiency tests are those that are not based on any specific syllabus or language programme but aim to ascertain the level of language ability attained by the test taker independent of any specific language training. This also means, of course, that test designers need to define what 'proficiency' means in any given context and to lay down a "specification of what candidates have to be able to do [...] to be considered proficient" (Hughes 2003: 11). In contrast to this, achievement tests relate to a specific programme or language course and intend to measure the progress made by a test taker who has attended the course, thus trying to measure "how much has been learned of what has been taught" (Davies 1977: 45). They can be administered at the end or during a language course. In this context, McNamara (2000: 7) points out that while achievement tests are backward-looking, proficiency tests are forward-looking as they aim to predict a test taker's use of language in future 'real life' situations.

As the terms suggest, diagnostic tests serve to provide information on the strengths and weaknesses of individual test takers in order to "ascertain what learning still needs to take place" (Hughes 2003: 15) while the purpose of placement tests is to place students into appropriate courses. In general, tests can be used in the context of educational programs as well as for research purposes (cf. Bachman 1990: 54), with the latter having an indirect influence on educational settings, for example, by testing theoretical models underlying language tests or by shedding light on the relative effectiveness of different teaching methods (as is the case with the empirical study described in Chapter 4 of this thesis).

2.2. Characteristics of tests: some relevant terms

Before discussing different language tests and the qualities they need to fulfil it might be useful to clarify a number of terms. One classification referring to the primary purpose of the test has already been introduced. However, there are a number of further aspects according to which different types of tests can be distinguished:

2.2.1. Direct vs. Indirect tests

A direct test can be defined as one "which claims to measure ability directly by eliciting a performance approximating authentic language behaviour" (Davies et al 1999: 47). Such tests are typically associated with tests of the productive skills, i.e. speaking and writing. The basic, and rather straightforward, idea is to test the speaking or writing skills of examinees by getting them to speak or write so that an extended sample of spoken or written language can be elicited. Typical examples are oral interviews or written composition tasks. As for the receptive skills, it is debatable whether or not these skills lend themselves specifically to direct testing in view of the fact that they "are essentially unobservable" (Davies et al 1999: 47). On the other hand, direct tests typically require the integration of various skills so that Weir, for example, includes receptive skills when describing the features of direct tests as follows:

> Direct testing requires an integrated performance from the candidate involving communication under realistic linguistic, situational, cultural and affective constraints. Candidates have to perform both receptively and productively in relevant contexts (Weir 1990: 12).

Indirect tests, on the other hand, target "the abilities that underlie the skills in which we are interested" (Hughes 2003: 18). Thus, instead of eliciting a direct sample of speaking or writing, these tests try to assess different components of what is seen as 'writing/ speaking ability' separately. The main criticism of indirect tests is that the relationship between mastering test items and performing in the criterion situation may be weak (Hughes 2003: 18).

While the classification into direct and indirect tests is often presented as a dichotomy, it should be remembered that there is, in fact, a continuum and that tests can thus be described as more or less direct. It will not always be possible, for example, to meet the requirements listed by Weir in the quote above and compromises may be necessary. It may also be pointed out that while there is a close relationship between this classification and the distinction between discrete-point and integrative tests (see Chapter 2.4. below) they are not synonymous. The cloze procedure, for example, is an integrative method of testing, but, at the same time, indirect (e.g. Hughes 2003: 19).

2.2.2. Pencil-and-paper tests vs performance tests

This distinction refers to the circumstances under which a test is carried out. As the name suggests, pencil-and-paper tests are those where students work on test papers specifically set for the purpose of the test. They typically involve standardised tests employing traditional methods of testing different skills or components separately. A performance test, on the other hand, aims to assess the skills of a candidate "in an act of communication" (McNamara 2000: 6) and is therefore always a direct test. Typically, test designers aim to simulate 'real-world' tasks and to put examinees into realistic contexts, specifying, for example, the concrete situation, purpose and audience of the communicative act. The aim is thus to "replicate their language performance in non-test situations" (Bachman 1990: 77). Generally speaking, with the ascent of the communicative approach to both language teaching and testing, it has become more common for test designers to include "performance features" (McNamara 2000:7) in their tests.

2.2.3. Subjective vs objective tests

This classification refers to the procedures involved in scoring tests. A test is defined as 'objectively scored' if no judgement is required of scorers as to whether or not a response is correct. Instead this is laid down in predetermined criteria and marking keys. Subjectively scored tests, on the other hand, require scorers to make a personal, and thus subjective, judgement. While 'subjective' is thus equated with 'requiring judgement', 'objective' means 'automatic', involving a scorer who "is reduced, for the purpose of marking, to the status of a machine" (Pilliner 1970: 20f). [1] In the case of subjective tests it is, of course, essential to have suitable procedures in place to control scorer variability. (see Chapter 2.5.3.) Generally speaking, it can be said that traditional 'pencil-and-paper tests' and indirect tests typically rely on tasks that can be objectively scored, an obvious example being the multiple choice test (see Chapter 2.4.1.). However, also other, more integrative, tests such as cloze tests or C-tests can be scored objectively if scoring keys are provided which lay down precisely which responses are to be judged correct (see Chapter 2.4.2. on cloze tests). The assessment of direct and performance tests, on the other hand, typically involves the use of rating scales and thus requires subjective

[1] In this context, Pilliner emphasizes that even in 'objective' tests there is plenty of room for subjectivity in other parts of the test cycle i.e. the test compilation and the test answering phases and that it is only at the assessment stage that a distinction can be made between subjective and objective tests.

judgements "since there is no feasible way to objectify the scoring procedure" (Bachman 1990: 76) in this case.

2.2.4. Norm-referenced vs criterion-referenced tests

Here we look at the frame of reference which is used in interpreting test scores. In the case of norm-referenced tests it is other test takers and their scores which provide this frame of reference. In other words, these tests "are designed and developed to maximize distinctions" (Bachman 1990: 75), which means that care is taken to include items which discriminate between those achieving high overall results and those whose results are poor.[1] Criterion-referenced tests, on the other hand, define certain levels of ability and assess the performance of candidates with reference to these levels. Thus "we learn something about what he or she can actually do in the language" (Hughes 2003: 20) rather than how his or her performance compares with that of other test takers.[2] In this case, items are selected that are seen as representative of specific content domains or specific abilities deemed relevant (Bachman 1990: 75f). While criterion-referenced tests are often seen as preferable, one advantage of norm-referenced ones is that statistical methods for analysing the test scores have been available for some time and are well-established, while more work still needs to be done in establishing such procedures for criterion-referenced ones (Hughes 2003: 21f).

2.3. Key qualities of tests

In order to be useful to test takers and users, and to ensure fairness, tests need to meet certain criteria. Most importantly, test designers must prove that the tests they have developed are valid and reliable, with reliability a necessary precondition for validity. Moreover, there are practical considerations such as the resources available which must not be underestimated. Therefore, in addition to being valid and reliable, tests also need to be efficient in terms of time and effort required.

[1] See Chapter 4.2.3. on the trialling phase in the development of the C-test used in the empirical study described in this paper.

[2] Of course, the scores can afterwards also be compared with those of other test takers, introducing a normative element. In the empirical study presented in Chapter 4, for example, test takers' performance on the writing task is assessed on the basis of rating scales with descriptors which allow us to describe the candidates' performance verbally. The scores obtained by different groups are then compared to establish the effect of CLIL provision on language output.

Before dealing with these requirements in a little more detail, one should perhaps add that the amount of time and effort spent on quality control will, of course, depend on the type of test one is dealing with. Clearly, it would be unrealistic to expect teachers to produce evidence of a comprehensive validation process when they set in-class tests. Nevertheless also at this level the key ideas about validity and reliability apply as obviously teachers are interested in setting tests that are both meaningful and fair. Developers of large-scale standardised tests, of course, will be expected to submit suitable evidence if they wish to convince potential test users of the qualities of their tests.

2.3.1. Validity

There seems to be general agreement that validity is "the central concept in testing and assessment" (Fulcher and Davidson 2007: 3). Put simply, deciding on a test's validity involves asking "whether a test measures what it is intended to measure" (Weir 1990: 1). In other words, we need to consider the underlying abilities and skills a test is supposed to measure and then ascertain whether or not inferences can be made about these on the basis of the scores obtained in a test. Validity can thus be defined as "the extent to which the inferences or decisions we make on the basis of test scores are *meaningful*, *appropriate* and *useful*" (Bachman 1990: 25). It follows that validity is not a feature of a test or test scores, but refers to how these scores are interpreted and used. Test validation thus aims to investigate "the defensibility of the inferences" (McNamara 2000: 10) and involves two elements, which McNamara (2000: 54) refers to as "speculation and empiricism". On the one hand, test developers need to use reasoning in considering what inferences can be made from the test right from the beginning, and throughout the different stages, of test development. Such 'a priori' considerations are then to be complemented with empirical investigations involving a number of statistical procedures using data from test performances. In this context, Weir (1990: 23) warns against overemphasising statistical aspects, regarding the analysis of empirical data as "a necessary but not a sufficient condition for establishing the adequacy of a test for the purpose for which it was intended." Moreover, it should be pointed out that a priori validation is essential in any test situation, even where comprehensive statistical evaluation is not feasible due to time constraints (cf. Weir 1990 :24).

Traditionally, different types of validity were distinguished, i.e. content validity, construct validity and criterion-related validity with its sub-forms of concurrent validity and

predictive validity (e.g. Fulcher and Davidson 2007: 4f; Bachman 1990: 236f). As the name suggests, content validity is concerned with test content, considering both content relevance and content coverage in order to demonstrate "that a test is relevant to and covers a given area of content or ability" (Bachman 1990: 244). As the content to be included is laid down in the test specifications a comparison between these specifications and the actual content of the test will form the basis for an investigation into content validity (Hughes 2003: 26f).

Construct validity concerns itself with theoretical 'constructs'[1] and the relationships between them that form part of the underlying theory of language ability. Such a construct can be described as "the specific definition of an ability that provides the basis of a given test or test task and for interpreting scores derived from this task" (Bachman and Palmer 1996: 21). The first step in the validation process thus requires a clear definition of this construct or ability. Test developers then need to prove "whether or not such a distinct ability exists, can be measured, and is indeed measured in that test" (Hughes 2003: 31). In demonstrating the construct validity of a test, it is thus the underlying theoretical model which is put to the test. As a consequence, the theory might be considered confirmed or may have to be modified or even abandoned (Hughes 2003: 32).

Establishing criterion-related validity involves determining "the extent to which test scores correlate with a suitable external criterion of performance" (Weir 1990: 27). In the case of concurrent validity, this could be another test administered at the same time. Typically, a newly developed test is validated against an older, well-established test assumed to provide "independent and highly dependable assessment of the candidate's ability" (Hughes 2003: 27) by investigating how well they correlate. In the case of predictive validity, on the other hand, researchers investigate the relationships between the scores obtained on a test and some future criterion, such as, for example, later academic success (Fulcher and Davidson 2007: 5).

More recently, the idea that validity splits into various sub-types has been abandoned in favour of what is called a "unitary concept of validity". This goes back to Messick, who defines validity as

[1] Fulcher and Davison (2007: 7) define constructs as concepts defined operationally so that "we can measure them in a test [...] by linking the term to something observable".

> an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the *adequacy* and *appropriateness* of *inferences* and *actions* based on test scores (Messick 1989: 13, quoted in Bachman 1990:236).

Under this integrated approach construct validity is seen as "the superordinate concept embracing all other forms of validity" (Weir 1990: 22). Evidence for the validity of inferences may be gathered in different ways, including analysis of test content or correlations with external criteria, but, as Bachman (1990: 237) points out, "none of these by itself is sufficient to demonstrate the validity of a particular interpretation or use of test scores". In addition, Messick emphasised the importance of consequential validity, which considers the values inherent in tests and test uses and the consequences of using test results for particular purposes (cf. Bachman 1990: 281-4).

Finally, another, fairly controversial, concept that is often discussed under the heading of 'validity', i.e. face validity, should be introduced. There seems to be general agreement these days that this is not an aspect of validity in the sense described above as it does not have any theoretical basis. Rather, the face validity of a test can be described as

> the degree to which a test appears to measure the knowledge or abilities it claims to measure, as judged by an untrained observer (such as the candidate taking the test or the institution which plans to administer it) (Davies et al 1999: 221).

In other words, a test is said to have face validity if it 'looks good' to its stakeholders. Bachman (1990: 287), for example, criticises the continued reference to this concept by practitioners, arguing that direct tests, in particular, may be given preference on the basis of their surface appeal only. On the other hand, while it should not replace true validation of test inferences, test appearance has an important role to play in ensuring that the test is accepted by test takers and users. As Bachman (1990: 288f) points out, it determines "whether test takers will take the test seriously enough to try their best and whether test users will accept the test and find it useful".

2.3.2. Reliability

Another key concern in language testing is the reliability, or consistency, of the scores obtained from a test. Just like validity, it is not a characteristic of the test itself, but refers to test scores. Moreover, it is a necessary precondition for making valid inferences on the

basis of these scores. In other words, the inferences made on the basis of test scores cannot be valid if the scores themselves are not reliable. Put simply, reliability refers to the "extent to which we can depend on the test results" (Weir 1990: 1). In other words, a test produces reliable results if we can be sure that these results are largely independent of factors other than the ability we wish to make inferences about. On a more technical level, reliability can be defined as "the proportion of the observed score variance that is true score variance" (Bachman 1990: 170). In other words, we need to estimate how much of the variance in the test scores can be put down to actual differences in ability and how much is due to measurement error. This will allow us to estimate how close actual scores are to the 'true scores' of test takers, which, of course, cannot be observed directly (Hughes 2003: 40).

Examples of factors that may influence actual scores are test takers' state of health, fatigue, interest and motivation, etc. (cf. Bachman 1990: 160). While these are largely unsystematic and thus impossible to predict, there are also other factors which may affect test scores in a more systematic way. Of particular relevance in this respect are so-called test method facets, or characteristics of the test method employed, as well as attributes of test takers that are independent of the ability to be tested. Test method facets refer to aspects of the test method such as the testing environment, the test rubric, the format of and language used in the input provided or the expected response and the relationship between input and response.[1] Care should be taken to ensure that such influences are reduced to a minimum as they distort test scores in a systematic manner. In fact, they erroneously increase reliability measures.

As has been pointed out above, reliability is a necessary (albeit not a sufficient) precondition for validity. In fact, the two concepts can be seen as "two complementary objectives in designing and developing tests: (1) to minimize the effects of measurement error, and (2) to maximize the effects of the language abilities we want to measure" (Bachman 1990: 161). Again, as with validity, investigating the reliability of test scores involves a two-step process including logical reasoning and statistical analysis of empirical data. While logical reasoning should help to identify potential sources of measurement error, statistical analysis will help to estimate the extent to which they influence scores (cf. Bachman 1990: 161f).

[1] For a more comprehensive overview of test method facets see Bachman (1990: 119).

A detailed overview of statistical methods used to calculate reliability measures would go beyond the scope of this thesis. Traditional measures are based on correlation coefficients and help to estimate internal consistency between test items, stability (or consistency over time) or equivalence estimates based on alternate forms of a test (Bachman 1990: 172). Another aspect of reliability particularly relevant in subjectively scored tests is rater consistency. The main concern in this field is, on the one hand, to establish how consistently individual raters allocate scores (intra-marker reliability) and, on the other, how much variability there is between raters (inter-marker reliability) (cf. Weir 1990: 32, see also Chapter 2.5.3.).

2.3.3. Economy and efficiency

While validity and reliability are undoubtedly the most essential aspects of language tests from both a theoretical and practical point of view, other, more pragmatic, considerations need to be taken into account as well. As Weir (1990: 34) points out, a "valid and reliable test is of little use if it does not prove to be a practical one". The resources available need to be considered in terms of financial viability, the need for skilled personnel or time constraints. In practice, therefore, compromises may have to be found, taking aspects of validity, reliability as well as efficiency and economy into account.

2.3.4. Test usefulness as a comprehensive concept

As an integrated approach to the required qualities a test should exhibit, Bachman and Palmer (1996: 18) suggest considering the overall concept of "test usefulness", which according to them includes the elements of reliability, construct validity, authenticity, interactiveness and practicality. The idea is to try and maximize the overall 'usefulness' of the test, to which all these factors contribute. The relative importance of these contributing factors, however, is said to change with the context of testing and thus "must be determined for each specific testing situation" (Bachman and Palmer 1996: 18). However, as Fulcher and Davidson (2007: 15) point out, this concept has not been taken up extensively by other authors.

2.4. Recent developments in language testing: from discrete-point testing towards communicative competence

Having clarified a few relevant classifications and the main properties tests should exhibit, we may now turn to the tests themselves. As with any other field of human activity, language testing has been subject to certain trends and fashions, which means that what is considered desirable in language testing has changed over time as different schools of language testing have come to dominate the field. This chapter is designed to briefly chart key developments in language testing from the 1960s up to the 21st century. It will focus on three main categories of tests, i.e. discrete-point tests, integrative tests and communicative tests. To illustrate the differences typical examples of such tests will be discussed briefly.

2.4.1. Discrete-point testing

Following what has been called the "traditional or pre-scientific phase" (Spolsky 1995: 353), in which little attention was paid to statistical analysis, the 1960s saw the rise of 'discrete-point testing', which is typically associated with the work of Lado. The underlying approach was structuralist and language tests were based on a view of language being made up "of small elements which can be learnt and tested in isolation" (Davies et al 1999: 48). Discrete-point tests therefore aimed to assess one element of knowledge at a time and typically consisted of a number of separate and decontextualised test items, each of which served to test one particular aspect of language knowledge. The main advantage of this approach was seen in the fact that the results could easily be quantified and subjected to statistical analysis, allowing test developers to achieve high reliability of test scores, a major concern at the time. It is due to this preoccupation with psychometric characteristics of tests that this period has been referred to as the "psychometric-structuralist period" of language testing (McNamara 2000: 14). In the meantime, however, the validity of such tests has been questioned. In 1979 Oller, for example, pointed out that testing language by means of atomistic, isolated test items failed to capture the very essence of language:

> Discrete point analysis necessarily breaks the elements of language apart and tries to teach them (or test them) separately with little or no attention to the way those elements interact in a larger context of communication. What makes it ineffective [...] is that crucial properties of language are lost when its elements are separated.

> The fact is that in any system where the parts interact to produce properties and qualities that do not exist in the part separately, the whole is greater than the sum of its parts (Oller 1979: 212).

Similarly, Morrow, who refers to the psychometric-structuralist period as "The Vale of Tears", argues that

> knowledge of the elements of a language in fact counts for nothing unless the user is able to combine them in new and appropriate ways to meet the linguistic demands of the situation in which he wishes to use the language (Morrow 1979: 145).

While Oller's doubts led him to suggest integrative tests as an alternative, Morrow is associated with the development of a communicative approach to language testing (see Chapters 2.4.2. and 2.4.3. below).

Before we turn to these alternative approaches let us consider specific test methods advocated in the 'psychometric-structuralist' era. Generally speaking, as objectivity and reliability were seen as the most essential qualities of language tests, preferred test methods were objectively scored indirect tests. Probably the most typical representative of this type of test is the multiple-choice test, which was the dominant method in the 1960s, being "so highly regarded [...] that it almost seemed that it was the only way to test" (Hughes 2003: 76).

Multiple-choice items consist in a stem and a number of options, only one of which is correct, while the others serve as distractors (Hughes 2003: 75). They employ a fixed-response format as candidates only need to select a given option and need not produce any language themselves. Moreover, they can be characterised as indirect pencil-and paper tests and, as the correct answers are predetermined, they lend themselves to objective scoring.[1] In terms of marking, such tests offer great advantages, not only in terms of "almost complete marker reliability" (Weir 1996: 43), but also as regards effort and time expended on marking, which leads to higher efficiency. Moreover, the format tends to be clear to test takers and the task can be completed without writing, thus avoiding possible contamination of the scores.

[1] As pointed out before, this does not mean that setting multiple-choice tasks does not involve any subjective judgement on the part of the test developer. Selecting items to be included in the test, for example, is clearly a subjective process (cf. Weir 1990: 43).

On the other hand, multiple-choice questions are not without their problems (cf Weir 1996: 43f). While scoring is efficient, the same cannot be said of test development as the preparation of multiple-choice items requires specially trained item writers who need to take great care in designing suitable items. Other disadvantages are the fact that this test format may encourage test takers to take a chance and simply guess the right answer, or even to try to cheat, and that the test method may introduce a systematic measurement error, thus affecting reliability. Last, but not least, there are severe doubts about the validity of inferences made on the basis of multiple choice tests as we do not know whether a test taker who has successfully completed a multiple-choice item would also be able to produce the correct form in a real-life situation involving speaking or writing (Hughes 2003: 76). The tasks themselves seem highly 'unreal' in the sense that "in real life one is rarely presented with four alternatives from which to make a choice to signal understanding" (Weir 1996: 44).

While the multiple choice format has come under considerable criticism in view of these disadvantages, and has lost its status as the method of choice, it is still used, but may be complemented with other elements of a more integrative or communicative nature.

2.4.2. Integrative testing: the psycholinguistic-sociolinguistic era

Criticisms of discrete-point testing led to a new 'era' in language testing, referred to as "psycholinguistic-sociolinguistic" (Weir: 1990: 3). As Oller points out, it was John Carroll who first made the distinction between discrete-point testing and integrative testing, emphasizing that while the former focuses on testing individual points of the language system one at a time, integrative tests aim to "assess a learner's capacity to use many bits all at the same time" (Oller 1979: 37). In other words, integrative tests are those that integrate "knowledge of systematic features of language [...] with an understanding of context" (McNamara 2000: 15).

Another term closely associated with Oller is the "pragmatic test",[1] which needs to meet two main "pragmatic naturalness criteria" and is thus defined as

> any procedure or task that causes the learner to process sequences of elements in a language that conform to the normal contextual constraints of that language, and

[1] Oller (1979: 19) defines pragmatics as the field of linguistics which "is concerned with the relationships between linguistic contexts and extralinguistic contexts".

> which requires the learner to relate sequences of linguistic elements via pragmatic mappings to extralinguistic context (Oller 1979: 38).

Therefore, to be classified as 'pragmatic' a test must, firstly, involve on-line processing of language and, secondly, require the test taker to relate language to context and to activate his or her linguistic knowledge in order to produce and understand language in this context (McNamara 2000: 15). Oller (1979: 38) further points out that integrative tests are typically, but not necessarily pragmatic, while pragmatic tests always require the integration of various skills. The basic idea is that "language tests should place the same requirements on test takers as does language performance in non-test situations" (Bachman 1990: 300f). This may recall the distinction made between direct and indirect tests, and in fact the two classifications overlap to a large extent. However, it should be pointed out that integrative/ pragmatic tests, as advocated by Oller, are not necessarily direct ones (Hughes 2003: 19). For example, the cloze test, which is closely associated with Oller and integrative/ pragmatic testing, is an indirect test as it does not "test communication itself" (Weir 1990: 4).

An important claim made by Oller in this context, referred to as the Unitary Trait Hypothesis, was that different types of pragmatic tests all involved the candidate drawing on "the same underlying capacity [...] – the ability to integrate grammatical, lexical, contextual, and pragmatic knowledge" (McNamara 2000: 15). According to this line of reasoning, it does not matter whether one administers a number of productive and receptive tests or just one integrative/ pragmatic test such as a cloze test since the same underlying capacity is activated. The idea was that in this way more expensive, direct tests could be avoided and replaced by indirect ones such as cloze tests or dictation.

The cloze test can be seen as the most typical test format used in this era of language testing. The term 'cloze' was coined by Wilson Taylor in 1953 and is said to refer to the concept of 'closure' in Gestalt psychology as completing a cloze test was likened to this process of completing a pattern (e.g. Oller 1979: 42; Weir 1990: 46). Originally conceived as a test to measure the readability of different texts, the cloze test soon developed into a widely used test of 'overall language capacity' and was, for a time, seen "as almost a language testing panacea" (Hughes 2003: 189). The 'classical' form of the cloze test is essentially a gap-filling test, in which every nth word is deleted and replaced by a gap. The task of the test-taker consists in filling the gaps and thus reconstructing the

original text (e.g. Oller 1979: 42). Scoring can be exact, requiring the candidate to replace the original word, or acceptable, marking semantically and syntactically suitable solutions as correct (e.g. Oller 1979: 367-75, Klein-Braley 1985b: 8). A later variant of the cloze test, the C-test, was chosen for the case study that forms the empirical part of this thesis and will be described in Chapter 4. While it is held to be an improvement on the cloze test, it is based on the same principles.

The underlying principle of both the cloze test and the C-test is that of reduced redundancy, which is based on the fact that language is naturally redundant, containing more information than is, strictly speaking, required to encode or decode a message.[1] This is important to allow the listener to decode the message correctly even though conditions may be less than ideal (due to 'noise in the channel'[2] such as other people talking, typing mistakes, incomplete transmission of the signal, illegible parts of texts etc.). It is assumed that native speakers of a language are generally able to cope with reduced redundancy on the basis of an anticipatory "global language processing competence" (Grotjahn et al 2002: 93) referred to as "pragmatic expectancy grammar" by Oller and "general language proficiency" by Spolsky. On the basis of "their knowledge of the rules, patterns and idiom of their own language and culture" (Raatz and Klein-Braley 2002: 76) native speakers are able to form hypotheses or expectations about what information is to follow and, on this basis, are able to restore the damaged text. Non-native learners of the language, on the other hand, will find this more difficult as they "cannot supply from [their] knowledge of the language the experience on which to base [their] guesses as to what is missing" (Spolsky 1973: 170). Naturally it depends on the level of language ability attained how well they can cope with reduced redundancy. Therefore the underlying idea of reduced redundancy tests is to assess an individual's language competence by deliberately introducing noise into the channel to see how well the candidate copes with the reduction in the natural redundancy of the language. The cloze test, which is "the most important and well known operationalization of reduced redundancy" (Klein-Braley 1985b: 8), does this by deleting words in the way described above.

[1] It has been suggested, for example, that with regard to letters printed English is about 75% redundant, meaning that 25% of the letters used would suffice to transmit the same amount of information if language use were completely efficient. (Klein-Braley 1985b: 2).

[2] In this context, noise refers to "any disturbance in communication" (Klein-Braley 1985b: 1). It is held to be unpredictable and random.

The main advantages of the cloze test are that it is easy (and therefore comparatively inexpensive) to construct, administer and score and that it has shown impressive performance in terms of reliability and (concurrent) validity (e.g. McNamara 2000: 16; Weir 1990: 4). In the 1970s and early 1980s it "came to be generally accepted that cloze tests were reliable and valid tests of general language proficiency" (Klein-Braley 1985b: 9) and the cloze test enjoyed great popularity. However, since then the test has come under criticism. On the one hand, doubt was cast on Oller's Unitary Trait Hypothesis (e.g. Weir 1990: 5, Bachman 1990: 6). Other points of criticism include the length of the tests, possible bias due to the fact that only one text is used, evidence that test results vary with test design (i.e. deletion rate and starting point), and frequent failure of native speakers to restore the exact words (e.g. Raatz and Klein-Braley 2002: 77f). Moreover, cloze tests typically have low face validity as filling the gaps tends to be seen as a "highly artificial and 'untestlike' task" (Bachman 1990: 49) by candidates. Finally, Oller's claim that cloze tests often require test takers to consider the wider context of a gap has also been disputed by authors providing contradictory evidence suggesting that cloze tests are "essentially sentence-bound" (Alderson 1978: 99, quoted in Weir 1990: 48) and that one rarely has to move beyond the immediate context to restore a word (Hughes 2003: 189). Finally, with the onset of the 'communicative movement' in language teaching (and testing), the cloze test was increasingly seen as unsatisfactory due to the fact that it was not seen as a "credible communicative task" (Carroll 1980: 12).

2.4.3. Communicative language tests

2.4.3.1. Models of communicative competence

In the late 1970s and 1980s new ideas emerged based on a broader view of the use of language as "the creation of discourse, or the situated negotiation of meaning and of language ability as multicomponential and dynamic" (Bachman 2000: 3) rather than a unitary trait. An important distinction (similar to the difference between *langue* and *parole* or competence and performance) in this context is that between *usage* and *use* (Widdowson 1978: 3), with the former focusing on using language structures correctly and the latter on using language appropriately, given the communicative purpose one has in mind. Basically, as Carroll (1980: 7) points out, appropriate language use should be the goal, while usage contributes to this goal and should thus be seen as a means to an end. From this point of view, the cloze test, which was discussed in the previous chapter, "is

still essentially usage-based [...] [as it] does not represent genuine interactive communication" (Carroll 1980: 9f) and fails to provide evidence on the communicative effectiveness of a language user.

New models emerged reflecting the broadened view of language ability which can only be sketched briefly. A major cornerstone was Hymes' theory of communicative competence, which includes aspects of grammaticality as well as acceptability. In fact, Hymes argues that a competent speaker must take four types of judgement into account, i.e. whether something is formally possible, feasible, appropriate and actually performed. (Hymes 1972: 281). This theory clearly marked a shift to a sociocultural perspective, as it included both a linguistic ('language competence') and a sociolinguistic element ('ability for use'), setting a pattern for later models (Weir 1990: 9). Generally speaking, it was based on the belief that "knowing a language was more than knowing its rules of grammar" (McNamara 2000: 16) and stressed the importance of communicative context.[1]

From this a number of models of communicative competence or communicative language ability developed. While the terminology is far from uniform, these models typically involve three basic dimensions (McNamara 1996: 48):

a) knowledge about the formal aspects of a language,
b) knowledge of how to use language appropriately (i.e. what Hymes called 'ability for use'), and
c) actual language use

While a) and b) together are typically seen as constituting 'communicative competence' or 'communicative language ability', the third point, which can be called 'communicative performance' focuses on what a learner "can do with the language" (Weir 1990: 9). With regard to language testing, it follows that communicative tests should assess learners on all three dimensions and should thus "require actual performance as well as tasks or item types that measure knowledge" (Fulcher and Davidson 2007: 39).

One of the most influential models of communicative ability is that developed by Canale and Swain. It features three components of communicative competence, i.e. grammatical

[1] Language competence can thus be seen as a necessary, but not a sufficient condition for successful and effective communication. While other important aspects are added by models of communicative language ability it should, on the other hand, not be forgotten that "linguistic competence must be an essential part of communicative competence" (Weir 1990: 13).

or formal competence, sociolinguistic competence (including rules of use and of discourse) and strategic competence, which refers to strategies used to compensate for shortcomings in other competences such as paraphrasing (Canale and Swain 1980: 29-31). It was later adapted by Canale, who turned discourse competence (which in the previous model had been included in the sociolinguistic component together with sociocultural rules determining appropriacy) into a separate element and redefined strategic competence in more positive terms to include strategies that make communication more effective (Fulcher and Davidson 2007: 40f). The Canale and Swain model, as adapted by Canale, can be seen as "the basis for all further work in this field" (Fulcher and Davidson 2007: 41), and is illustrated in Figure 1.

<table>
<tr><td colspan="4">Communicative competence</td><td>Actual communication</td></tr>
<tr><td colspan="4">Knowledge and skill</td><td></td></tr>
<tr><td>Grammatical competence</td><td>Sociolinguistic competence</td><td>Strategic competence</td><td>Discourse competence</td><td>Instances of language use</td></tr>
</table>

Figure 1 The Canale and Swain model (as adapted by Canale)
(Source: Fulcher and Davidson 2007: 41)

Bachman (1990: 81), for example, further extends this model by trying to show how the various components "interact with each other and with the context in which language use occurs". According to this model, communicative language ability includes three main elements, i.e. language competence, strategic competence and psychophysiological mechanisms, as illustrated in Figure 2 below.

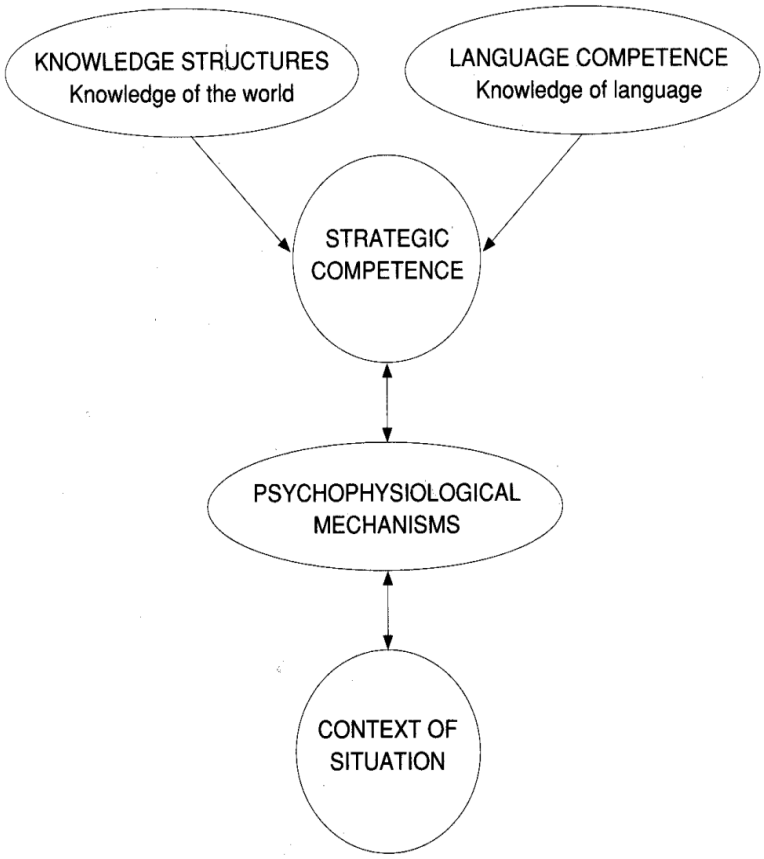

Figure 2: Components of communicative language ability in communicative language use
(Source: Bachman 1990: 85)

Language competence in turn consists of organizational (i.e. grammatical and textual) competence on the one hand and pragmatic (i.e. illocutionary and sociolinguistic) competence on the other (Bachman 1990: 86-90). Basically, organizational competence deals with formal language structures and allows the language user to produce and understand grammatically correct sentences and their propositional content and to order them in such a way as to produce (and understand) texts. Pragmatic competence, on the other hand, relates to the language functions intended by the speaker (illocutionary competence) and rules of appropriate use (sociolinguistic competence). Figure 3 illustrates the various components of language competence according to this model:

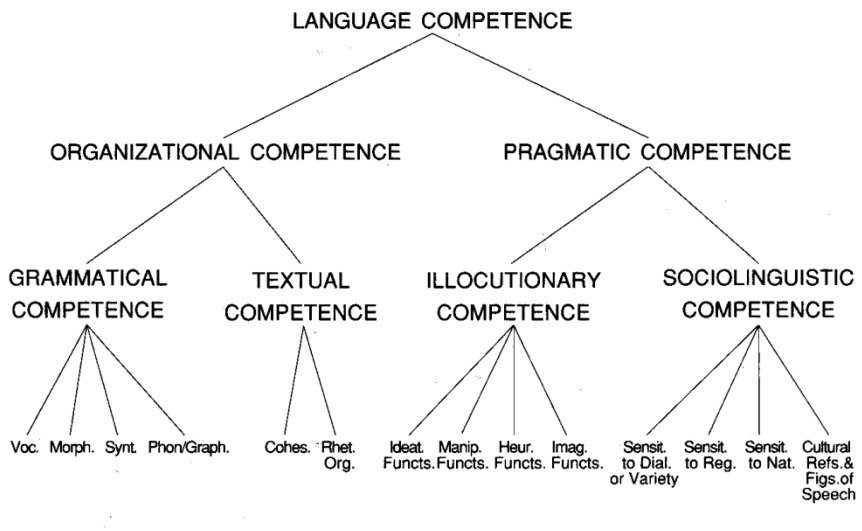

Figure 3 Components of language competence
(Source: Bachman 1990: 87)

Strategic competence is broadened again to include three components (ie. assessment, planning and execution) which are important in any kind of communicative language use as they serve to determine how a particular communicative goal can best be attained by drawing on and implementing the various elements of language competence (Bachman 1990: 84; 107f). In this sense, it can be seen as "a general reasoning ability" (McNamara 2000: 19) which allows the language user to use language to engage in the negotiation of meaning in a given context. Finally, psychophysiological mechanisms are required to actually produce and receive language signals.

What does all this mean for language testing? As Bachman (1990: 3f) points out, a "clear and explicit definition of language ability is essential to all language test development and use". After all, test developers should have a clear idea what exactly they are trying to test candidates on. Models of communicative language ability can form the basis on which test constructs are defined if the test in question is to be a 'communicative' one. Typically, of course, such general models will be too broad-based for most language tests and will have to be narrowed down and made more specific, leading to frameworks including the relevant parts of the model, and, in the next step, to test specifications which then provide the basis for constructing specific tests (cf. Fulcher and Davison 2007:36).

2.4.3.2. Characteristics of communicative language tests

Having laid down some theoretical concepts concerning communicative language ability we might now turn to the question what features typically characterise 'communicative language tests'. In spite of the popularity of what is termed the 'communicative approach', it has been suggested that "the nature of what makes a test 'communicative' is still not well understood" (Bachman 2000: 12). Nevertheless certain features are typically associated with communicative tests.

It has already been pointed out that communicative tests should assess candidates not only on the basis of their knowledge or competence, but also their ability to translate this competence into actual performance "in ordinary situations" (Morrow 1979: 149). As a consequence, communicative tests are characterised by two typical features. On the one hand, they must be performance tests (see Chapter 2.2.2.), involving candidates in "an extended act of communication" (McNamara 2000: 16) and on the other, they must consider the social roles assumed by test takers in these acts of communication As McNamara (2000: 17) points out, it is the latter point in particular which distinguishes communicative tests from the integrative/ pragmatic tests advocated by Oller by emphasizing the "external, social functions of language".

Weir (1990: 30f) lists a number of features that communicative tests typically exhibit and concludes that

> a test within the communicative paradigm [...] should be interactive; direct in nature with tasks reflecting realistic discourse processing activities; texts and tasks should be relevant to the intended situation of the target population; ability should be sampled within meaningful and developing contexts and the test should be based on an explicit a priori specification (Weir 1990: 36).

Similarly, Morrow (1979: 149f) stipulates a number of features that characterise communicative tests. First of all, communicative tasks need to be interaction-based. While this is most obvious in the case of oral interaction, where "what is said by a speaker depends crucially on what is said to him" (Morrow 1979: 149), it also applies to writing to a specified addressee, as the writer has to make certain assumptions concerning the expectations and needs of the recipient. Moreover, the outcome has to be, to some extent, unpredictable, depending on how the interaction evolves. Another requirement is that the task must be set in a particular, realistic context and must serve a particular

purpose so that the appropriacy of the candidate's language use can be assessed. Any language input should reflect normal performance conditions and be authentic. In other words, it should neither be idealised (e.g. by avoiding false starts) nor simplified. Finally, the assessment should be based on the communicative result, judging what the candidate can do with language. Concerning the last point, Morrow (1979: 150) even suggests that "strictly speaking no other criteria are relevant", but concedes that this extreme position is probably untenable as it would hardly be accepted by language teachers.

One particularly salient feature typically associated with communicative tests is that of authenticity. Basically, there are two approaches to the question of what constitutes an authentic task. On the one hand there is what Bachman (1990: 301) calls the "real-life" approach, the aim of which is to "mirror the 'reality' of non-test language use" as closely as possible so that inferences can be made about the candidate's future performance in non-test situations. Bachman (1990: 302) is highly critical of this approach as no distinction is made between the performance observed in the test and the underlying ability. Moreover, the question arises whether it is possible at all to replicate 'real-life' situations in a language test as all parties involved are aware of the main purpose of the test, i.e. assessing the candidate's language capacity, which is quite different from the communicative purpose of the task itself. In this sense, therefore, any language test is "by definition inauthentic" (Bachman 1990: 313). Moreover, there are practical constraints, such as the need to ensure fairness and to allow for easy administration and scoring or economic constraints, which may require "a sort of principled compromise" (McNamara 2000: 27) and put a limit on the test developer's quest to make the test as authentic as possible.

The second approach advocated by Bachman is referred to as the "interactional/ability" (IA) approach, which "focuses on what is sees as the *distinguishing characteristic* of communicative language use – the interaction between the language user, the context, and the discourse (Bachman 1990: 302). In this respect, Bachman builds on the ideas of Widdowson (1978: 80), who holds that authenticity is "a characteristic of the relationship between the passage and the reader and [...] has to do with appropriate response". In other words, whether a test task is considered authentic or not depends on how the test taker interacts with it. This also means that, under this approach, it is permissible, for example, to simplify or otherwise edit the input language in order to make it 'comprehensible' to the test taker as clearly the candidate will not be able to interact with the task in an

authentic way if he or she does not comprehend the input (Widdowson 1978: 82; Bachman 1990: 319).[1] To design an authentic language task test developers thus need to study and identify the key characteristics of the target language use (TLU) in terms of the abilities needed for effective communication and the characteristics of the situation. On this basis a framework can be developed in which the relevant language abilities and test method facets are laid down as a basis for test development (Bachman 1990: 356f) Authenticity is then seen as the "degree of correspondence" achieved between the TLU situation and the test situation (Bachman and Palmer 1996: 23).[2]

While multiple choice tests are typically associated with the psychometric-structuralist period and cloze tests with integrative/ pragmatic tests, communicative tests are probably most readily associated with tests of the productive skills, in particular speaking tasks, which "offer plenty of scope for meeting the criteria for communicative testing" (Weir, 1990: 73). Whether this scope can be realised depends, of course, on practical constraints such as the resources available. Examples would be oral interviews (free or controlled), interaction tasks based on information-gap activities or role plays. In the field of writing, free or controlled writing tasks are often used to assess written communicative ability. An example is the writing task in the case study described in Chapter 4, which is designed to meet the requirements of communicative language testing by putting test takers into an authentic communicative situation.

2.5. The testing cycle: Designing, trialling and administering tests

This chapter is designed to provide a brief overview of the basic stages of test development. Of course, the steps to be taken largely depend on the type of test, the test purpose and the conditions under which it will be administered (cf. Bachman 1990: 1). At one end of the spectrum, there are so-called high-stakes tests, i.e. those that "provide information on the basis of which significant decisions are made about candidates" (McNamara 2000: 133). To make sure such decisions are well-founded, the development process needs to be highly complex, involving a large team of test developers as well as extensive trialling and revision efforts. At the other end, the development of low-stakes

[1] Advocates of the 'real-life' approach, on the other hand, would insist on the use of "unsimplified language" to ensure authenticity (Weir 1990: 12)

[2] Bachman and Palmer (1996: 23-25) split the notion of authenticity into two qualities, i.e. authenticity and interactiveness, the latter focusing on the characteristics of the test taker and to what extent these are engaged by a test task (cf. Bachman 2000: 12f).

tests, such as regular quizzes administered in class, is typically a rather informal process (Bachman and Palmer 1996: 85).

Basically, we can distinguish between a number of different stages such as test design, test construction, trialling and operation (cf. McNamara 2000: 23). While on the one hand, test development can be seen as a linear process, with one stage followed by the next, it is, at the same time, an iterative one, with information from each stage feeding back into the process (cf. Bachman and Palmer 1996: 86; Fulcher and Davidson 2007: 89), so that it can be referred to as a "testing cycle" (McNamara 2000: 23).

2.5.1. Test design: specifying test content and method

2.5.1.1. Test content

Designing a test mainly involves deciding on the test content and the test method, with the latter including issues such as response formats and scoring. The test developer needs to decide on the language domain and the constructs to be focused on, ensuring that the demands of content relevance and content coverage are met. Obviously, this has important implications for the validity of the test (Bachman 1990: 244; see also Chapter 2.3.1.). Therefore it is important to keep a record of this process as this "constitutes validity evidence" (Fulcher and Davidson 2007: 89).

As McNamara (2000: 25) points out, there are two approaches to defining the test domain. On the one hand, this can be done "operationally", which means that the test designer tries to identify the most characteristic tasks from a particular real-world domain in order to include them in the test. This is particularly frequent in the case of tests with a clearly defined specific purpose, targeted at a specific group of test takers such as medical doctors or nurses who wish to practice their profession abroad. In this case a job analysis is typically carried out to help specify what tasks are of particular relevance. The other approach is to define the test domain in terms of a "more abstract construct" (McNamara 2000: 25), i.e. on the basis of a theory or model of language ability. In this context, as mentioned before, Fulcher and Davidson (2007: 36) suggest a three-step process. Having chosen an appropriate model, a framework document is drawn up which consists of the components of the model which are relevant for the domain in question. This then serves as a basis for developing test specifications, according to which a test can be constructed.

Tests designed to measure general language ability and targeting a wide range of test takers are more likely to be designed in this way.

2.5.1.2. Test method: response format and scoring

Once the test content has been laid down, a suitable test method has to be chosen. The two aspects are closely related and different methods are often associated with particular types of content (cf. McNamara 2000: 30). Test methods can be described on the basis of so-called test method facets (see Figure 4).

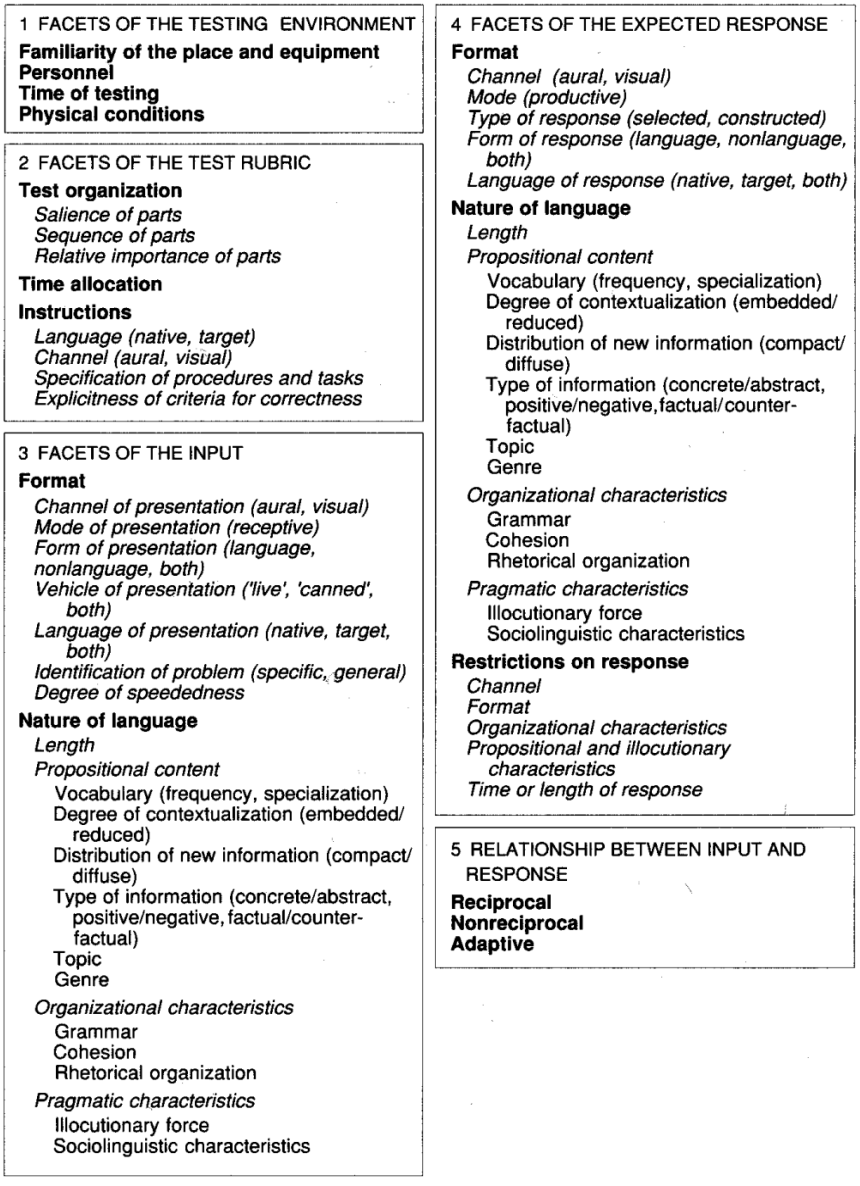

Figure 4: Categories of test method facet
(Source: Bachman 1990: 119)

Figure 4 shows the framework of test method facets suggested by Bachman (1990: 119), including facets of the testing environment (e.g. whether the test taker is familiar with the environment and the people involved in the test), the test rubric (e.g. the wording of instructions), the input (e.g. how the input is presented), the expected response (e.g. fixed vs. constructed response formats, see below) and the relationship between input and response (e.g. whether it is reciprocal or not). These facets need to be chosen carefully as they may potentially have an impact on the performance of candidates and may influence different candidates in different ways, an effect that needs to be taken into account in validity considerations.

One category of test method facet that deserves particular attention is the response elicited from the test taker. Here a distinction is commonly made between fixed and constructed response formats. While in a test using fixed response format (such as, for example, a multiple-choice test) the task of the examinee is to select the most appropriate answer from a range of options given, tests calling for a constructed response require test takers to formulate their own responses so that their ability to produce language can be assessed. A wide range of tests use this format, offering the test taker different degrees of freedom of choice (cf Davies et al 1999: 32). In a cloze test, for example, this choice is more limited while in a free-writing task or an extended speaking task the candidate has much more control over the language choices made. The test developed for the case study described in Chapter 4 consists of two tasks, i.e. a C-test and a writing task, which both require a constructed response from test takers, but provide them with different degrees of freedom to formulate their own responses.

The main advantages of the constructed response format thus are that it imposes fewer constraints on the test person, that the answer cannot as easily be guessed and, in general, that "the candidate assumes greater responsibility for the response" (McNamara 2000: 30). Typically, therefore, tests demanding a constructed response are seen as more challenging. On the other hand, they are more difficult to score, as they often involve subjective judgements and thus require the design of a more comprehensive marking scheme (e.g. including acceptable alternatives in a cloze or C-test) or rating scales (in the case of extended writing or speaking) (Davies et al 1999: 32). This leads us on to the next point to be considered when deciding on a test method, i.e. how performances are to be scored.

While in the case of objective tests (see Chapter 2.2.3.), scoring is rather straightforward as the acceptable responses are predetermined and laid down in scoring keys, designing an appropriate scoring system is more challenging when tests involve subjective scoring by raters. Nevertheless, rater-mediated assessment, which was frowned upon in the psychometric-structuralist period, has become more and more important due to the growing use of direct tests under the paradigm of communicative language teaching and testing (McNamara 2000: 36). In this case, rating scales are typically developed to provide guidance to raters.[1] Such scales are based on a number of criteria judged relevant (according to the underlying model of language ability) and descriptors, which describe "in words performances that illustrate each level of competence defined on the scale" (McNamara 2000: 40) and are designed to help raters distinguish between different levels of performance.

In developing rating scales, decisions need to be made about the rating method (holistic or analytic, see below), the number of levels on the scale and the wording of the descriptors. As this is a rather challenging task, Hughes (2003: 106) points out that for low-stakes tests it might be most useful to consider existing scales, which can then be adapted to suit the requirements of a given test. In any case, it is essential to consider scoring procedures at the stage of test design, together with test methods, and not "at some later stage" (Fulcher and Davidson 2007: 257).

Basically, there are two types of rating procedures, depending on whether test performances are assessed as a whole or according to a set of separate criteria. Although the terminology is not uniform,[2] these two approaches are commonly referred to as holistic and analytic rating. In the case of holistic rating, which is also referred to as "impressionistic scoring" (Hughes 2003: 94)), raters are asked to "record a single impression of the impact of the performance as a whole" (McNamara 2000: 43). Typically, a rating scale is provided as guidance, in which case the approach may also be called "focused holistic scoring" (Hamp-Lyons 1991: 244). As trained raters can score

[1] In fact, this is not the only function. Alderson (1991: 72-74) distinguishes between user-oriented, assessor-oriented and constructor-oriented functions of rating scales.

[2] For example, 'analytic' scoring is sometimes still taken to mean an assessment based on frequency counts, which goes back to the 1960s and 1970s. In this case 'multiple trait scoring' may be used to describe an approach that assesses test performances according to different categories (Fulcher and Davidson 2007: 254). On the other hand, 'holistic' is sometimes used to refer to any assessment based on verbal descriptions rather than frequency counts. Weir (1990: 67), for example, refers to an "analytic, holistic marking scheme" as one assessing different categories separately, based on verbal descriptors.

tests holistically at rather high speed, this scoring procedure is less time-consuming and thus also more economical even if each test is assessed by more than one rater.[1] On the other hand, there are a number of disadvantages associated with an assessment based on overall impression. One problem arises in case candidates show an "uneven development of subskills" (Hughes 2003: 102), which means that they exhibit different levels of progress in different aspects of competence such as fluency and accuracy. Moreover, holistic scoring procedures are not "designed to offer correction, feedback, or diagnosis" (Fulcher and Davidson 2007: 251) so that the scores cannot be explained to test takers or test users. Finally, raters may rate the overall impression of a test performance on the basis of only a few salient aspects so that "the prejudices and biases of the marker" (Weir 1990: 64) may have a considerable effect on scores. This could lead to low inter-rater consistency.

For these reasons, analytic rating may be preferred, under which raters are asked to "provide separate assessments for each of a number of aspects of performance" (McNamara 2000: 43). Therefore separate scales have to be developed for each aspect to be considered and, in addition, an overall score is often given when reporting results. In establishing such an overall score, the individual categories may be of equal importance or may be weighted (Hughes 2003: 103). Analytic scoring has a number of advantages over holistic scoring. As categories are assessed separately, no problems arise if candidates show different levels of proficiency in different fields. Moreover, diagnostic information can be provided to test takers and users so that "full profile reporting" (Weir 1990: 67) is possible. It also forces raters to consider a number of categories in assessing test performances, even though it is debatable whether the ratings on different categories are truly independent or subject to what is referred to as a "halo effect"(Hughes 2003: 102f), meaning that a score awarded on one dimension might influence the assessment of another category. Finally, analytic scoring is usually taken to lead to more reliable results (Hughes 2003: 102). The most important disadvantage of analytic scoring is that it is considerably more time-consuming and thus more expensive than a holistic scoring system. Another potential problem is that by focusing on various categories raters may lose sight of the overall effectiveness of a test performance. Therefore, they may also be asked to provide an overall score based on their general impression (Hughes 2003: 103).

[1] Research suggests that rater reliability is acceptable when each test is scored four times (e.g. Hughes 2003: 95; Weir 1990: 65).

2.5.2. Test specifications and test construction

The process of test design, which has set the general parameters, is followed by an operationalisation stage, in which detailed test specifications are drawn up.[1] These provide instructions for the actual construction of individual tests, "written as if they are to be followed by someone other than the test developer" (McNamara 2000: 31). On the basis of these specifications test items are written including a key or expected response. As this is a challenging task, a moderation process should be undertaken involving at least two people other than the author of the item, whose task it is "to try to find weaknesses in the items" (Hughes 2003: 63). In this way items that need to be revised or even abandoned can be identified. It should also be stressed that more items need to be prepared than are actually needed so that in the next phase, the test trialling phase, the less suitable ones can be weeded out and the most suitable ones selected.

2.5.3. Test trialling

Before a test becomes operational, the materials prepared need to undergo a process of trialling to ensure their usefulness for the purpose intended. A pilot version of the test is administered to a trial population, which should resemble the target test population as closely as possible, for example with regard to factors such as sociodemographic data, learning history or general language capacity (McNamara 2000: 32). In the case of objectively scored tests, the scores obtained are then subjected to a number of statistical procedures referred to as item analysis[2] in order to establish the usefulness of the test items included in the pilot test. However, a caveat needs to be added here. Decisions on whether or not to include an item must, of course, not only be made on the strength of statistical evidence alone, but should "be based on multiple sources of information, both qualitative and quantitative" (Bachman 2004: 137). There may be good reasons, for example, for including an item even though it does not meet the statistical criteria laid down in the test specifications. In particular, content considerations may take precedence. Similarly, an 'easy' item might be included at the beginning of the test in order to ease people in and reduce anxiety (see Chapter 4.2.3.).

[1] Bachman and Palmer (1996: 90) distinguish between test task specifications and a blueprint, the latter dealing with how test tasks are to be organised into actual tests.

[2] In addition to classical item analysis, more recent approaches such as item response theory (e.g. Bachman 2004) may be useful. These are, however, not focused on in this thesis.

Typical indicators that are calculated are those measuring item facility (or item difficulty) and item discrimination (e.g. McNamara 2000: 60; Bachman 2004: 122). Item facility (also referred to as P value) simply calculates the percentage of test takers offering the correct solution and provides an estimate of how difficult the test is for the target population. It ranges from 0 (if no test taker provided the correct answer) to 1 (if everyone answered the item correctly).[1] On this basis test developers can decide whether the level of difficulty is appropriate for the purposes of the test. In the case of tests which are designed to differentiate between test takers (norm-referenced tests, see Chapter 2.2.4.), the ideal item facility is 0.5,[2] but a wider range is usually accepted. McNamara (2000: 61), for example, suggests an acceptable range from 0.33 to 0.67. Similarly, while Bachman (2004: 130) advocates selecting items "within a fairly narrow range, around 0.50", he later (2004: 137f) adds that in practice ranges between 0.25 and 0.75, or 0.20 and 0.80, are typically accepted.

Item discrimination, on the other hand, shows how well individual items distinguish between those who do well on the overall test and does who perform less well. It is thus concerned with "internal consistency reliability" (Bachman 2004: 130), i.e. whether candidates perform consistently across items. If consistency, typically measured by a discrimination index, is low, this means the scores will be misleading as "no clear picture of the candidates' abilities emerges from the test" (McNamara 2000: 61). As this would seriously challenge the reliability of the test scores, items with low discrimination indices[3] should be revised or removed from the test before it becomes operational. Finally, the reliability of the test scores can be estimated by calculating a reliability coefficient, such as coefficient alpha (also known as Cronbach's alpha) or Kuder-Richardson estimates (Bachman 2004: 163). Generally speaking, the reliability coefficient should be at least 0.9. This means that 80% of score variability can be traced back to differences in language ability, while 20% is due to measurement error (McNamara 2000: 62). An example of an item analysis involving facility, discrimination

[1] Strictly speaking, this definition only holds for items scored as either right or wrong (R-W). In the case of 'partial credit'(P-C) scoring, where possible scores range from 0 to whatever has been defined to be the maximum score, it refers to the average score achieved on an item (Bachman 2004: 122).

[2] This is due the fact that the distribution of items scores and that of test scores are inversely related. Therefore, the "more spread out the item difficulty indices are, the more closely bunched together the test scores will be" (Bachman 2004: 130). In norm-referenced tests, however, test developers want test scores to be spread out.

[3] Bachman (2004: 138) stipulates that only items with discrimination indices of at least 0.30 should be included.

and reliability indices is provided within the framework of the case study presented in this thesis (see Chapter 4.2.3.).

In those cases where scoring procedures involve the subjective rating of test performances by raters, the trialling period serves to test the usefulness of the rating scale and to train raters in order to make sure that the scores obtained by test takers "are affected as little as possible by the particular examiner who assesses them" (Weir 1990: 82). In other words the aim is to reduce, as much as possible, variability of test scores due to intra-rater and inter-rater inconsistency. Intra-rater inconsistency, i.e. a lack of self-consistency exhibited by individual raters, may be caused, for example, by factors such as the order in which performances are scored or the time of scoring (Bachman 2004: 169). Inter-rater inconsistency, i.e. differences in scoring between different raters, may be due to personal characteristics of the raters such as an overall tendency to be lenient or harsh, different patterns of relative leniency or severity depending on the aspects of language ability considered, their attitude towards different groups of test takers, or different interpretations of the rating criteria. (e.g. McNamara 1996: 123-5).

In order to estimate and control such effects certain procedures of quality control are required. Again, statistical analyses serve to calculate reliability coefficients estimating both intra-rater and inter-rater reliability. Moreover, rater training in the form of a number of moderation meetings, in which differences in scoring and in the interpretation of rating scales are to be discussed and thus reduced, should help improve reliability by "bringing about broad agreement on the relevant interpretation of level descriptors and rating categories" (McNamara 2000: 44). A certain amount of subjectivity will always remain, of course. However, raters who consistently provide scores that are at variance from those of their peers, or who prove to be highly inconsistent themselves, may have to be replaced.

It has already been pointed out that apart from statistical analyses, the trialling period should also be used to obtain qualitative information. Feedback from all parties involved may provide highly valuable information on how the test is seen from different perspectives. Test-taker feedback, for example, can provide useful information on how the test is perceived by candidates, e.g. in terms of level of difficulty, face validity or clarity of instructions (McNamara 2000: 32). On the basis of all the information obtained

during the trialling phase, decisions can be made about which items to keep, which ones to revise and which ones to eliminate.

2.5.4. Test administration

Once the trialling stage has been completed and the necessary revisions have been made, the test becomes operational and can be administered to the test target population. However, this does not mean that test development is completed once and for all. Rather, information gathered after the administration of the test may lead to further revisions and improvement. In this sense, unless a test is abandoned completely, the testing cycle is never complete.

3. Content and Language Integrated Learning (CLIL)

3.1. What is CLIL?

3.1.1. General aims and development

One of the aims of the European Union is to promote multilingualism, which is seen as an important and "desirable life-skill" (European Commission 2005: 3), among the population of its member states, the idea being that this will foster both integration and cohesion between member states and the continued existence of linguistic and cultural diversity.

Multilingualism can be defined as "the ability of societies, institutions, groups and individuals to engage, on a regular basis, with more than one language in their day-to-day lives" (High Level Group 2007: 6). On an individual level, it is seen as an important requirement to prepare European citizens for "life in a mixed global society" (Mehisto et al 2008: 10). Moreover, in view of the growing trend towards globalisation and internationalisation, it is considered an economic necessity in order to ensure the competitiveness of the European Union and the achievement of its economic goals. Without any doubt, educational systems need to respond to these trends and prepare future employees for a working environment which requires them to be "increasingly multi-skilled and mobile" (Marsh 2007: 6). One approach which is designed to help achieve these aims and may be seen as "a key instrument to create a multilingual

population in Europe" (Ruiz de Zarobe 2008: 61) is Content and Language Integrated Learning (CLIL), which can briefly be defined as a "dual-focussed educational approach in which an additional language is used for the learning and teaching of both content and language" (Maljers et al 2007: 8).

Bilingual education is nothing new, of course, and, on an international level, there have been a number of attempts to develop students' skills in languages other than their L1 through content teaching in these languages for quite some time. Cases in point are the French Immersion programme in Canada and Content-Based Instruction (CBI) in the US. The former was started in the 1960s and involved Anglophone children being taught content subjects exclusively in French, thus leading to 'total immersion' into the second language, which is the other official language of the country. The focus of CBI, on the other hand, is on helping immigrant children develop proficiency in their second language, which is the official language of instruction, through content-based teaching and learning (Dalton-Puffer 2008b: 140). In both cases, therefore, the language of instruction is one that has a clear presence outside the classroom. In contrast to this, CLIL, which is considered a "European approach" (Wolff 2007: 11), typically involves using a foreign language,[1] to which students are rarely exposed outside school, as the (additional) language of instruction. Of course, bilingual and multilingual education has existed in Europe for a long time, but was often considered 'elitist' and thus reserved for limited segments of the population. The CLIL movement, on the other hand, is a broader-based approach seeking to make bilingual education available in mainstream education and thus accessible to larger parts of the population (e.g. Maljers et al 2007: 9). Having been taken on board by the European Union as a means to foster integration, the basic idea started to gain momentum in the early 1990s. In 1994 the term 'Content and Language Integrated Learning' (CLIL) was coined (Mehisto et al 2008: 9) and in 1996 it was introduced as an 'umbrella' term covering a wide range of approaches involving the use of a language other than L1 as a medium of instruction.

[1] A foreign language can be defined as a language "which is not usually used in the surrounding social environment" (*CLIL Compendium*), while a second language is one that has a wide currency in the surrounding environment.

3.1.2. Defining CLIL

The term 'Content and Language Integrated Learning' is now the preferred term in the European context to refer to a situation where a language other than the L1 of the participants is used in the teaching and learning of a non-language subject.[1] In contrast to foreign language (FL) classes, the language concerned is not the subject of the lesson but instead serves as a means of communication in the construction of subject-specific knowledge. Here are some definitions of the term:

> CLIL is a dual-focussed educational approach in which an additional language[2] is used for the learning and teaching of both content and language [...] It's not 'language learning', and it's not 'subject learning'. It's a 'fusion' of both. CLIL always involves dual-focused aims, for in a CLIL class attention is simultaneously given to both topic and language. (Maljers et al 2007: 8)
>
> The term Content-and-Language-Integrated-Learning (CLIL) refers to educational settings where a language other than the students' mother tongue is used as medium of instruction (Dalton-Puffer 2007: 1).
>
> CLIL involves students learning subjects such as science or geography through the medium of a foreign language [...] CLIL is sometimes referred to as dual-focused education as lessons have two aims, one related to a particular subject or topic and one linked to language (British Council).
>
> CLIL is an educational approach which involves different types of language-supportive methods highly suitable for contexts where the medium of instruction is a second/third language (*CLILCOM*).

While each definition focuses on slightly different aspects, they all share the basic ingredients, i.e. non-language content being taught through a language other than the mother tongue of the participants. This, as well as the term CLIL itself, of course, raises questions about the relationship between the two main ingredients, ie. content and language, and how they are to be integrated.

---

[1] Typically, content subjects such as history, geography or social sciences are chosen. The most typical language used in CLIL is English, with French and German in second and third place respectively. (Wolff 2007: 12).

[2] The term additional language is used to "refer to any language other than the first language" (*CLIL Compendium*), thus replacing terms such as foreign, second or minority language.

3.1.3. The relationship between content and language

In spite of the harmonious relationship suggested by the term 'Content and Language Integrated Learning', reality is often somewhat different. In fact, the question which element should be considered primary has been answered in different ways so that three basic approaches to bilingual learning can be distinguished, i.e. a language-focused approach, a content-focused approach and an integrated approach (Vollmer 2008: 53-56). The Canadian Immersion programme and Content-Based Instruction, which have been mentioned above, fall into the first category, aiming mainly at developing students' second language proficiency and using content learning as a tool designed to help achieve this end. In this case the bilingual programme is basically seen as an "innovative form of language teaching [...] using other contents than the traditional foreign language teaching contents" (Wolff 2007: 13) to develop language competence. It is based on input-oriented theories of language acquisition and can be summed up as "language learning through language use" (Vollmer 2008: 54).

The European approach, on the other hand, has tended towards the second approach, seeing CLIL classes basically as variants of regular content classes, based on the same curricula as the latter (cf. Dalton-Puffer and Smit 2007: 12, Wolff 2007: 16). As a consequence, "learning content matter is primary and learning language a secondary goal" (Järvinen, forthcoming). In fact this 'secondary goal' is often left unspecified so that no explicit language learning goals are defined. Under this approach, language skills of participants are expected to improve, given the increased exposure, but this is seen as incidental to the acquisition of content knowledge and happens unsystematically and with little explicit support from the teacher. (Vollmer 2008: 54). Basically, learners are expected to "inductively pick up the foreign language while working with content" (Wolff 2007: 13). CLIL classes are typically taught by content teachers, who may, however, be supported by a native speaker assistant or a language teacher in cases of team teaching.[1]

There have been doubts, however, whether an approach that is predominantly focused on content and which basically consists in "changing the medium of instruction [...] without

[1] It should be noted in this context that in the Austrian system of secondary education many teachers have a dual qualification in two subjects so that the content teacher teaching CLIL often is a qualified foreign language teacher at the same time. The situation is somewhat different in vocational schools, however, where this situation only applies to general knowledge subjects such as geography or history.

adaptation of methods" (Maljers et al 2007: 6, see also Lasagabaster 2008: 40), does justice to the idea of CLIL and one may wonder if the full potential of CLIL can be realised under such an approach (cf. Vollmer 2008: 56). If we consider the definitions given above there seems to be a mismatch between the general principles of CLIL and how it is commonly implemented. Therefore, a third approach is called for which is "content-oriented but at the same time language-sensitive" (Wolff 2007: 17), involving true integration of content and language[1] and specifying both content-specific and language-specific goals. In other words "the 'how' and 'why' of integration of content and language" (Järvinen, forthcoming) needs to be further explored and specified. If true integration is achieved this should make it possible to exploit the potential of CLIL more fully.

In this context, it has been suggested that the development of CLIL has, in the meantime, entered a new phase of consolidation (Maljers et al 2007: 7). According to this view, the first decade, i.e. the period from 1994 to 2004, was characterised by rapidly growing interest on two levels, i.e. the European level, with a number of "landmark trans-national declarations, events, and a range of publications" (Maljers et al 2007:7) and, at the same time, the 'grassroots' level, with individual teachers and schools implementing CLIL programmes on their own initiative (cf Dalton-Puffer 2007:3). On the one hand, this has led to a large variety of approaches and activities which "cover a wide range of curricular realities" (Dalton-Puffer 2008b: 140) but are all considered to come under the CLIL umbrella.[2] Mehisto et al (2008: 13), for example, present 13 different approaches ranging from rather limited 'language showers' to total immersion in what is often referred to as a 'language bath'. On the other hand, as Dalton-Puffer (2007: 3f) has remarked, there is still a conspicuous gap between the transnational and the local, grassroots level that needs to be filled by both research and national education policies. It is to be hoped, therefore, that the second phase of CLIL development, which may be defined as the period from 2004 to 2014, will fill this gap. This should lead to a certain amount of consolidation, putting the emphasis on developing a more unified approach to CLIL and a language-

[1] Mehisto et al (2008: 11) define a third element in addition to language and content, i.e. learning skills, which "constitute the third driver in the CLIL triad".

[2] Two major dimensions along which CLIL programmes may vary are type of school and environmental factors such as, for example, interpretation of the concept, choice of subjects, and length of exposure (Wolff 2007: 13).

sensitive methodology which is geared toward fully exploiting the potential benefits of CLIL.[1]

3.2. The rationale for CLIL – What benefits are to be gained

3.2.1. Official aims

So far the development of CLIL has been sketched briefly, but the question still remains why we should engage in CLIL activities in the first place. Officially stated aims tend to be rather vague. The *CLIL Compendium*, for example, states the following aims, categorised according to issues relating to culture, environment, language, content and learning:

- Build intercultural knowledge & understanding
- Develop intercultural communication skills
- Learn about specific neighbouring countries/regions and/or minority groups
- Introduce the wider cultural context

- Prepare for internationalisation, specifically EU integration
- Access International Certification
- Enhance school profile

- Improve overall target language competence
- Develop oral communication skills
- Deepen awareness of both mother tongue and target language
- Develop plurilingual interests and attitudes
- Introduce a target language

- Provide opportunities to study content through different perspectives
- Access subject-specific target language terminology
- Prepare for future studies and/or working life

- Complement individual learning strategies
- Diversify methods & forms of classroom practice
- Increase learner motivation (cf. *CLIL Compendium*).

Similar texts provide similarly general aims. Dalton-Puffer (2007: 275), for example, sums up the language goals stated in a number of relevant meta-texts as "generally

[1] This would probably also involve putting more emphasis on the development of reading comprehension and writing skills (e.g. Wolff 2007: 16).

enhanced learning outcomes, in other words 'more of everything'." Nevertheless a number of arguments are typically made in favour of CLIL provision which have led to high face validity among practitioners as well as 'customers', i.e. parents and the students themselves.

The starting point might be seen as a general dissatisfaction with the outcome of traditional foreign language classes, which seem to produce limited outcomes after years of foreign language instruction. Wolff (2007: 10), for example, states that foreign language teaching seems ill-suited to achieve the aims of a plurilingual society as "only few students acquire the linguistic competence necessary of their future life in one foreign language, let alone two".[1] In this context Lasagabaster (2008: 31) refers to a Eurobarometer study published in 2006 which showed EU citizens being far from multilingual as "for a remarkable 44% communication in a language other than their mother tongue was highly implausible". CLIL is supposed to help improve this situation. Compared to 'traditional' foreign language classes CLIL lessons are typically held to be more effective as they offer the opportunity to use the target language in a more 'natural' way, focusing on the message rather than the form and thus increasing student motivation. This may explain why CLIL classes are "becoming commonplace throughout the continent" (Lasagabaster 2008: 31).

3.2.2. The benefits of CLIL

One of the most important arguments in favour of CLIL is the assumption that it creates more 'natural' conditions for the use of the target language than FL classes do. It is thus compared to 'learning in the street' and taking a 'language bath', where one can immerse oneself and 'soak up', or acquire, language skills without explicit instruction. (cf Dalton-Puffer and Smit 2007: 8). Instead of contrived activities thought up by the language teacher students engage in 'real' discourse about concepts and topics pertaining to the content subject. This is widely regarded as "something more real which interlinks with the world outside" (Wolff 2007: 15). Moreover, as the focus is on meaning rather than linguistic form, it is possible to create a "motivating, low-anxiety communicative atmosphere, thus emulating a non-instructional learning environment" (Dalton-Puffer

[1] This refers to the European Commission's "long-term objective [...]to increase individual multilingualism until every citizen has practical skills in at least two languages in addition to his or her mother tongue" (European Commission 2005: 4).

2007: 205).[1] Freed from the worry about linguistic correctness students are more likely to overcome inhibitions and become more self-confident in using the foreign language. In short, CLIL seems the epitome of the communicative classroom and it has thus been argued that the CLIL movement "can be viewed as the next phase of the 1970s' communicative revolution" (Maljers et al 2007: 9).

As far as theoretical models are called upon to underpin such rationales, it is usually Krashen's Monitor Model (e.g. Krashen 1982; Lightbown and Spada 2006: 36-38, Dalton-Puffer 2007: 258f) which most readily comes to mind. In simple terms, according to Krashen, a language is acquired if learners are exposed to sufficient comprehensible input, i.e. input that is just beyond their current level of competence,[2] but which they can understand due to contextual information or deliberate efforts on the part of interlocutors to make it comprehensible (e.g. in the form of 'foreigner talk' or 'teacher talk', which is intentionally simplified). In contrast to conscious learning, or attention to form, acquisition[3] is seen as a subconscious process which happens incidentally while learners are engaged in communication, focusing on meaning.[4] Spontaneous production of language is determined by what has been acquired in this way, and can only be polished a little by what has been learned through conscious attention to form provided the learner has enough time, is focused on producing 'correct' utterances and, of course, knows the relevant rule, conditions which are unlikely to be met in normal conversation (Krashen 1982: 16).

While the supply of sufficient comprehensible input is thus seen as the main factor determining acquisition, Krashen also holds that a learner may be prevented from benefiting from such exposure due to a barrier referred to as 'affective filter'. In other words, if learners are not motivated, lack self-confidence or experience high anxiety

[1] This view was also commonly expressed in the teacher interviews conducted by Dalton-Puffer (2007: 227). An interesting point in this context is that, when comparisons are made between CLIL and FL classes, the latter are often, at least implicitly, presented as being mainly concerned about linguistic correctness. In view of the currently dominant paradigm of communicative language teaching this may seem a little surprising.

[2] However, Krashen also emphasizes that in applying this model a teacher should not deliberately aim at providing input corresponding to what is referred to as *i*+1, with *i* representing current competence and *i*+1 the next stage. Rather, he holds that "if the acquirer understands the input, and there is enough of it, *i*+1 will automatically provided" (Krashen 1982: 21). It is enough to 'roughly tune' the input to the level of competence of the students. In fact, Krashen claims that deliberately focusing on *i*+1 might be counterproductive.

[3] Acquisition may also be referred to as implicit, informal or natural learning (Krashen 1982: 10).

[4] The ideal case would be for the input to be "so interesting and relevant that the acquirer may even 'forget' that the message is encoded in a foreign language" (Krashen 1982: 66).

levels, either individually or as a group, acquisition will not take place even if plenty of suitable input is provided (Krashen 1982: 31). Boredom, anxiety or other negative emotions may lead students to simply "'filter out' input, making it unavailable for acquisition" (Lightbown and Spada 2006: 37). To sum up the main ideas, according to Krashen's model, the ideal classroom is one fostering acquisition by providing ample exposure to comprehensible input in a highly motivating, low-anxiety environment that meets the needs of the students.

CLIL lessons seem to fit this model perfectly. As CLIL is offered in addition to foreign language classes, it clearly leads to increased exposure to the target language in quantitative terms.[1] Moreover, in contrast to foreign language classes, the focus is not on form and linguistic correctness, but on meaning as the lesson focuses on curricular concepts pertaining to the respective discipline. While this leads to more meaningful input, giving "the use of the foreign language a purpose over and beyond learning the language itself" (Dalton-Puffer and Smit 2007: 8), it also reduces anxiety levels and "is believed to boost the affective dimension" (Lasagabaster 2008: 32). In this way it increases motivation and confidence, putting learners into positive emotional states and thus creating good conditions for acquisition to take place.

Other theories that seem relevant are Long's Interaction Hypothesis, which emphasises the importance of negotiation of meaning in developing language competence (e.g. Long 1996: 451f), and Swain's Output Hypothesis. The latter holds that in addition to being exposed to comprehensible input, learners should also be encouraged to produce comprehensible output as this will require them to "move from the semantic, open-ended, non-deterministic, strategic processing prevalent in comprehension to the complete grammatical processing needed for accurate production" (Swain 1995: 128). Requiring students to produce comprehensible output would thus be instrumental in developing their competence with regard to syntax and morphology. By encouraging students to discuss subject-matter concepts, CLIL classes should be well-suited not only to provide increased exposure to the target language but also ample opportunity to negotiate meaning and produce output in the L2.

[1] Typically, CLIL teachers are also likely to make a deliberate effort to ensure this input is comprehensible, for example, by more frequent repetition or an increased use of pictures or graphs (see Chapter 3.3.1.).

Finally, it should not be overlooked that CLIL also offers practical advantages in terms of economy and efficiency. Basically the idea is that CLIL helps to increase exposure time to the target language without cutting time devoted to other, non-language subjects. If competences in language and content can be built up simultaneously, it follows that one can have "two for the price of one" (e.g.Vollmer 2008: 54) and it seems possible "to improve students' command of foreign languages without devoting too much time to their teaching" (Lasagabaster 2008: 31).

3.2.3. Some qualifications: CLIL and the development of communicative competence

As described above, CLIL seems almost too good to be true. It allows students to acquire the target language naturally and incidentally while they engage in meaningful discourse about subject matter concepts. The conditions seem ideal for acquisition through exposure to comprehensible input, negotiation of meaning and the production of comprehensible output. In short, it seems that a CLIL classroom is the ideal setting for a communicative approach to language teaching, which "sets out to involve learners in purposeful tasks which are embedded in meaningful contexts and which reflect and rehearse language as it is used authentically in the world outside the classroom" (Hedge 2000: 71). One might conclude that adopting CLIL is equivalent to implementing "the principles of the communicative approach on a grand scale" (Dalton-Puffer and Smit 2007: 8).

In order to determine whether this is an accurate picture of what goes on in the CLIL classroom and to gain more insight into the 'language bath' CLIL students are immersed in a closer look needs to be taken at actual classroom discourse. After all, the classroom is where teaching and learning takes place. In spite of the high expectations based on input and output oriented theories, however, research into actual CLIL classroom discourse has drawn a more differentiated, and somewhat disappointing, picture. Research has suggested, for example, that "[s]ustained negotiation [...] occurs rarely or not at all in these classrooms" (Musumeci 1996: 286) and that the output produced by the students is very often limited. As classroom discourse seems to be dominated by what is called IRF (Initiation-Response-Follow Up) cycle or Triadic Dialogue, students very often limit themselves to producing minimalist responses to the teacher's initiation, which are then followed up by the teacher. In addition, as teachers seem to shy away from expository teaching to avoid teacher monologue and 'lecturing' in favour of a distributed co-

construction of curricular content, opportunities may often be missed by the teacher to provide learners with clearly structured, syntactically complex input (cf. Dalton-Puffer 2007: 91).[1]

The question therefore arises whether CLIL classrooms are settings in which communicative competence can, in fact, be acquired. In this context, we need to recall that according to Hymes, who coined the term 'communicative competence' (see Chapter 2.4.3.1.), this competence is acquired "through participation in actual communicative events" (Dalton-Puffer and Smit: 9). In this way, children (or other language learners) learn "when to speak, when not [...] what to talk about with whom, when, where and in what manner" (Hymes 1972: 277), thus acquiring not only language, but also cultural knowledge. It has already been pointed out that, in theory, CLIL classrooms seem to provide opportunities for such 'communicative events'. However, what tends to be overlooked is what goes on in the classroom itself, i.e. the setting in which (institutional) learners' communicative competence is shaped and developed. In this context, we must not forget that, as Dalton-Puffer (2008b: 148) reminds us, "CLIL lessons are *lessons*" and can thus not be equated with acquiring a language 'in the street'. While there are, of course, differences between CLIL and foreign language classes with regard to the content which is taught, they are, at the same time, very similar in a number of respects such as participants, setting, purposes, participation structures and act-sequence (i.e. the order in which different phases of classroom teaching take place) (cf Dalton-Puffer 2008a: 15f).

On the basis of her own research into CLIL lessons held in Austria, Dalton-Puffer sets out to investigate empirically to what extent communicative competence can be fostered in CLIL classrooms. In carrying out this "reality check" (Dalton-Puffer 2008a: 14), she uses Canale and Swain's model of communicative competence (see Chapter 2.4.3.1.) as her point of reference. As stated before, this model defines communicative competence as consisting of linguistic competence, sociolinguistic competence, discourse competence and strategic competence. The question therefore is to what extent each of these competences can be developed in a CLIL classroom.

[1] Apart from the "considerable impoverishment of the linguistic input available to learners" (Dalton-Puffer 2007: 91) this can also be assumed to have negative effects on students' ability to understand and appropriate subject-matter concepts. However, as this thesis focuses on the language side of CLIL (rather than the acquisition of content), this issue cannot be explored here any further.

As has been mentioned before, in Canale and Swain's model linguistic competence[1] refers to knowledge about the formal aspects of a language. In practical terms it thus comprises "knowledge of spelling, pronunciation, vocabulary, word formation, grammatical structure, sentence structure and linguistic semantics" (Hedge 2000: 47). As this list shows, linguistic competence is usually taken to refer to aspects of the language system up to the sentence level and can be considered "the traditional realm of foreign language classes" (Dalton-Puffer 2007: 279). With regard to the CLIL classrooms investigated, Dalton-Puffer observes that they do, in fact, offer learning opportunities concerning vocabulary, but very few with regard to syntax. Lexical gaps are readily acknowledged by the students, who are even found to actively seek repair, a fact which may be regarded as "one of the chief distinguishing characteristics of CLIL classrooms over EFL lessons" (Dalton-Puffer 2007: 87). As for syntax, on the other hand, students encounter few (morpho-)syntactical problems due to the fact that classroom discourse is dominated by minimalist responses within the context of Triadic Dialogue.[2] Naturally, this also means that this kind of classroom interaction offers few opportunities for further development in this respect.[3] It is also worth noting in this context that teachers tend to avoid overt correction of grammatical mistakes when they do occur while by far the most frequent cases of repair and language-directed teaching concern lexical errors or gaps. At the same time, teachers typically mention vocabulary first when asked about the linguistic benefits of CLIL, which leads Dalton-Puffer (2007: 283) to conclude that "most learning seems to take place where most mistakes are made [...] [and that] students should maybe be given more 'chances' to make syntactic errors as well."

While learning opportunities in terms of grammar seem limited in a CLIL classroom due to the ways discourse is organised there, with regard to the other aspects of communicative competence, it is the classroom situation itself, and the conditions prevailing in it, which puts a limit on what can be acquired by learners. With regard to sociolinguistic competence, which is concerned with what is considered "appropriate

[1] As pointed out in Chapter 2.4.3.1., Canale and Swain (1980:29) use the term 'grammatical competence' to refer to this part of communicative competence.

[2] On the whole, CLIL teachers seem unlikely to actively encourage students to go beyond such minimalist responses. Several authors have come to the conclusion that teachers are inclined to accept student contributions as long as they understand them and are much less likely to encourage them to produce comprehensible output (cf Mariotti 2006).

[3] With regard to morphology, however, Dalton-Puffer (2007: 283) reports that the increased exposure to the target language in a CLIL classroom does have beneficial results, for example, by leading to automatisation in notorious problem cases such as third person –s.

within a given sociocultural context depending on contextual factors such as topic, role of participants, setting, and norms of interaction" (Canale and Swain 1980: 30), including issues of appropriate register, formality and interpersonal relationships, it must be acknowledged, for example, that the range of speech acts in a content classroom is necessarily limited (Dalton-Puffer 2008a: 17f). Moreover, the realizations of those speech acts which do form part of the discourse are often of limited variation due to the fixed role relationships in the classroom. Dalton-Puffer (2007: 286) concludes that in terms of sociolinguistic competence CLIL classrooms do not really offer any advantages over foreign language classes, as they share the same setting and roles and that neither of them can "be expected to 'prepare' learners for other situational contexts in any direct way".

The same holds true with regard to discourse competence, which deals with questions of "sequencing and arrangement" (Dalton-Puffer 2007:287) and is most typically associated with writing. With regard to CLIL, however, it refers almost exclusively to spoken discourse as there is hardly any writing done in a typical CLIL classroom apart from some note-taking. Obviously, the classroom has its own discourse rules, for example with regard to turn-taking or topic nomination, due to the asymmetric role distribution between teacher and students. As a consequence, it cannot be expected to help learners acquire discourse rules operating in other contexts. A similar situation is to be found with strategic competence, which refers to strategies employed to cope with problems in communication due to limited competence in the other areas. Again, the specific setting and role relationships of a classroom determine the use of strategies which may be undesirable in other settings.[1] All in all, it can be said, therefore that the conditions under which classroom discourse unfolds "necessarily impose restrictions on all aspects of communicative competence" (Dalton-Puffer 2007: 292).

While classroom conditions are, of course, basically the same in CLIL and EFL classes, the latter offer the possibility of explicit instructions and practice in those speech acts, discourse rules and communication strategies which will not naturally occur in the classroom and thus cannot be acquired. Moreover, one might try to adapt teaching arrangements in CLIL classrooms to include role plays or group work which could

[1] Two examples given by Dalton-Puffer (2008a:19) are avoidance strategies and message replacement. The former refers to the fact that, unless they are nominated personally, students find it easy to avoid responding altogether, while the latter denotes a situation where the learner's response does not really fit the question.

provide opportunities for more varied language use, provided students do not immediately switch to their L1 in such cases (as they invariably did in the classrooms studied by Dalton-Puffer). In addition, it would be helpful to conduct both the instructional and the regulative register in the target language, as the latter typically offers more variation in the language used (cf. Dalton-Puffer 2007: 202).

While the opportunities to acquire communicative competence in a typical CLIL classroom thus seem to be somewhat limited and not radically different from language classrooms, one should not overlook that, in general, the conditions of the classroom offer opportunities as well in so far as they are familiar to the students, giving them a feeling of security. Combined with the shift in focus away from language and linguistic correctness to subject matter discourse, this provides students with a kind of sheltered space in which language use can unfold. As Dalton-Puffer (2007: 294) observes, this may be one of the reasons for "the self-confident and self-evident use of the foreign language and its ultimate appropriation by many CLIL learners" which studies on the outcome of CLIL provision have regularly shown.

One area in which CLIL lessons would be particularly well suited to prepare students for 'real-life' use is the acquisition of knowledge and academic language functions. Wolff (2007: 16), for example, argues that

> the fact that content subject learning has a scientific character contributes to the learner's developing an academic competence in the foreign language [...] In CLIL the learner acquires linguistic competencies which do not only extend beyond communicative competence but also relate much more to the use the foreign language in professional life later on.

While this sounds convincing, it does not seem to be borne out by reality. In fact the development of 'cognitive academic language proficiency', to use the term coined by Cummins,[1] seems to be largely neglected in current CLIL lessons. The lessons observed by Dalton-Puffer, for example, provide very little evidence of the use of academic language functions such as defining, explaining and hypothesising. This seems quite surprising as it is an area where CLIL classrooms, given their focus on the acquisition of

[1] Cummins distinguishes between basic interpersonal communicative skills (BICS) and cognitive academic language proficiency (CALP) and claims that "academic aspects of L2 proficiency take considerably longer to develop" (1991: 75). This would suggest that the development of academic language proficiency will need to be started relatively early. Due to their focus on academic subjects CLIL classes would seem particularly well suited for this purpose.

subject-matter knowledge, could help develop students' language skills in fields other than oral fluency, which seems to be seen as the main benefit at the moment (cf. Dalton-Puffer 2007: 295).

3.3. Outcomes

3.3.1. Content knowledge

With regard to empirical studies on CLIL, Dalton-Puffer and Smit (2007: 12-14) distinguish between two dimensions. i.e. a macro-micro dimension and a product-process dimension. The study underlying most of the previous chapter, i.e. Dalton-Puffer's study on naturalistic discourse in CLIL classrooms, for example, adopts a micro-process perspective. This chapter, on the other hand, while retaining a micro-perspective, focuses on the outcome, or 'product' of CLIL provision.

In view of the two elements contained in the term 'Content and Language Integrated Learning', i.e. content and language, outcome studies typically focus either on the effects of CLIL provision on learners' subject-matter knowledge or on their language ability. With regard to the former, concern has often been voiced that students' content knowledge might suffer if they have difficulties following the lessons in another language. On the one hand, this may lead to slower progress so that less material can be covered, while, on the other, teachers might anticipate problems and thus simplify the content in order to facilitate understanding. As a consequence, it is argued, both coverage and depth of the content presented might suffer (cf Dalton-Puffer 2007: 5). However, most research undertaken so far suggests that this is not the case (e.g. Mehisto et al 2008: 20, Wolff 2007: 21). In fact, some studies have even shown CLIL students to outperform those taught in their L1. Van de Craen et al (2007: 270f), for example, compare the mathematical skills of CLIL and non-CLIL primary pupils and conclude that the provision of CLIL at an early age has a positive effect on cognitive abilities in general.

Generally speaking, positive effects on content knowledge may be due to students showing a higher degree of procedural competence and intensifying their efforts in the face of language difficulties (Dalton-Puffer 2008b: 142). On the other hand, teachers, aware of the additional challenge faced by their students when taught in a language other than their L1, seem to put more effort into preparing their lessons in such a way that

understanding is facilitated in a number of ways, such as using a greater variety of didactic approaches, more repetition, and increased use of visual support materials such as graphs and illustrations and proceeding more slowly (cf. Dalton-Puffer et al 2008: 8; Schindelegger 2008, Wilhelmer 2008). It could be that while there might be some quantitative reductions in the materials covered in CLIL classes as opposed to regular classes, the key concepts may be retained better due to the increased use of such didactic tools. As one teacher pointed out in a survey on CLIL in Austrian engineering schools (Wilhelmer 2008), in CLIL lessons teachers are more likely to deliberately build up curricular concepts in small steps to facilitate understanding while in regular classes using L1 they might be more tempted to start at a level that is conceptually beyond the level of the students. If this is the case, CLIL lessons might, to some extent, even serve as a model for content teaching in general.

3.3.2. Language competence

It is generally believed that CLIL helps "to improve overall language competence in the target language, in particular oral skills" (Lasagabaster 2008 32). In general, it can be said that this is borne out by research on the general language ability of CLIL students compared to their non-CLIL peers. Of course, this is hardly surprising as CLIL is offered in addition to foreign language classes and thus increases exposure to the foreign language.

When considering different aspects of language competence, Dalton-Puffer (2008b: 143) distinguishes between those that seem to be affected favourably by CLIL provision and those that seem unaffected or have not been sufficiently analysed (see Table 1).

Favourably affected	**Unaffected or Indefinite**
Receptive skills	Syntax
Vocabulary	Writing
Morphology	Informal/non-technical language
Creativity, risk-taking, fluency, quantity	Pronunciation
Emotive/affective outcomes	Pragmatics

Table 1: Language competencies favourably affected or unaffected by CLIL
(Source: Dalton-Puffer 2008b: 143)

It has already been mentioned that the field of vocabulary has been found to benefit particularly from the provision of CLIL classes, resulting in greater lexical richness and variety and a lower incidence of lexical transfer and direct borrowing from the students' L1 (cf Ruiz de Zarobe, forthcoming; see also Ackerl 2007: 9f). Given the fact that, as mentioned in Chaptr 3.2.3., this is the only language area where repair is typically offered in CLIL classes, and even sought actively by students, this corroborates the view that the greatest gains are made in the fields where most mistakes are made and dealt with in class.

Focusing on oral skills, Mewald (2007), for example, uses a number of communicative tasks in interviews designed to compare the foreign language ability of CLIL and non-CLIL learners at the lower secondary level within the context of a case study based on a school pilot project. The results of the test were analysed according to 21 criteria focusing on fluency as well as more 'traditional' categories such as grammar, lexis and pronunciation. The findings show that the CLIL learners outperformed their peers in terms of fluency, spoke more and used longer sentences, made fewer grammar mistakes, used more varied structures and produced a wider range of lexical items. Moreover, they were found to communicate more effectively and showed more creativity. An interesting point in this context is that the extent to which pupils benefited from CLIL seemed to vary with their level of proficiency, with low achievers partly even underperforming their mainstream peers. Other research has shown that while high achievers may also become highly proficient when taking foreign language classes only, average-proficiency students seem to benefit the most from CLIL provision (Dalton-Puffer 2008b: 142f).

That CLIL students show better oral skills does not come as a surprise as practitioners of CLIL typically see 'higher communicative competence' as one of the key benefits of CLIL. Typically, what they mean by this is, however, not the range of competences included in a model like Canale and Swain's. Instead a much narrower view is usually adopted which sees 'communicative competence' mostly in terms of "fluency and/or low anxiety in face-to-face interaction" (Dalton-Puffer 2007: 278). It has also been pointed out that, apart from note-taking, writing does not play any significant role in the CLIL lessons themselves (at least not in German-speaking countries) so that at first research efforts concentrated on oral skills. More recently, however, there has been increasing interest also in how improved language skills acquired through CLIL affect students' writing abilities. In a study comparing the writing skills of students from two different

CLIL programmes and mainstream students,[1] Ruiz de Zarobe (forthcoming) found that the CLIL students outperformed their non-CLIL peers in most of the categories analysed, showing statistically significant advantages with regard to content and vocabulary. Nevertheless, compared to a test on oral skills conducted with the same groups of students in which the CLIL students performed significantly better in every single category analysed (Ruiz de Zarobe: 2008),[2] the advantages were less clear.

A more comprehensive picture is provided by Lasagabaster (2008). His study, which was conducted in the Basque Country, involved administering a number of English tests to 198 secondary students, focusing on grammar, listening, speaking and writing skills. Grammar and listening were tested on the basis of the Oxford Placement test, the speaking test was based on the so-called 'frog story', a well-known research instrument based on a picture story, and the writing test consisted in asking the students to write a letter to a prospective host family (Lasagabaster 2008: 35). Results on the different sub-tests were added up and analysed, using four-level analytic rating scales for the productive skills of speaking and writing. The results show clear, and statistically significant, advantages for the CLIL students in all sub-tests and in every single scale used in the writing and speaking tests. This suggests that CLIL may lead to superior results even in areas which other research has found to be less affected by the CLIL experience, notably writing and pronunciation. It is also noteworthy that with regard to grammar and overall scores, third-year CLIL students performed better than fourth-year mainstream students. Lasagabaster (2008:40) concludes that "the CLIL approach is successful and helps to improve students' language competence even in contexts where English has little social presence and is hardly ever used outside the school setting".

So far only general language skills have been considered. In a study which in terms of sample size is comparable to that of Lasagabaster, Zydatiß (2007) compared CLIL and non-CLIL secondary students not only in terms of grammatical, lexical and communicative competences but also in terms of subject-matter literacy. While the CLIL students showed better results in all the areas analysed, it must not be overlooked that the advantage was much less clear in terms of academic discourse competencies. Zydatiß

[1] Students were asked to write a letter to a host family, a task which is also used by Lasagabaster and in the case study described in Chapter 4.

[2] Like Lagasabaster's more comprehensive study this study was based on the 'frog story'. The rating scales used in the speaking test referred to the categories of pronunciation, vocabulary, grammar, fluency and content (Ruiz de Zarobe 2008: 67). In the written test the categories analysed were content, organisation, vocabulary, language and mechanics (Ruiz de Zarobe, forthcoming).

(2007: 378) argues that the development of the latter may be hampered by low linguistic competences and that a certain minimum level of linguistic competence must be ensured before academic competences can be developed, for example by intensifying foreign language teaching before the onset of the bilingual programme.[1]

Generally speaking, empirical studies conducted so far suggest that CLIL students tend to outperform their non-CLIL peers in terms of general language ability, although it should not go unmentioned that there are also studies which have found no difference between CLIL and non-CLIL learners (cf Lasagabaster 2008: 33, Ruiz de Zarobe 2007). It should also be noted that there are differences depending on the type of language skills analysed and that in terms of academic skills there may be room for improvement for both CLIL and non-CLIL groups. Without any doubt, more empirical research will prove useful in providing further insights. Nevertheless the results available already point to certain strengths and weaknesses of current CLIL programmes and may thus serve to suggest possible improvements such as, for example, the development of academic skills or a greater focus on writing skills.

4. Empirical study

4.1. Introduction

4.1.1. Purpose of the study

The aim of the present two-case study is to assess the general level of language ability of 11th grade students (i.e. around 16 years of age) who, in addition to EFL classes, have undergone CLIL instruction in comparison to their EFL-only peers. All the students participating in the study attend one of two upper-secondary engineering colleges in Austria, hereinafter referred to as College A and College B. While the main focus is on general language ability, particular emphasis is also put on writing skills. Therefore, the main research hypotheses are:

[1] As will be shown in Chapter 4, the CLIL programme of one of the schools participating in the case study is based on the same idea, offering additional EFL classes before English is used as a medium of instruction in specialist engineering subjects.

1. CLIL students outperform their mainstream peers in terms of general language ability, measured both objectively and subjectively.
2. CLIL students outperform their mainstream peers in different aspects of writing ability.

The first hypothesis is in line with the general consensus that CLIL provision has a positive effect on students' general language ability (see Chapter 3.3.2.). As CLIL lessons are provided to the students in addition to their regular EFL classes, and, in fact, additional English classes are also offered to the CLIL students at least on an optional basis in both colleges,[1] CLIL students are clearly exposed to more language input and have more opportunity to use the language than those who are confronted with English in their regular EFL classes only.[2] The second hypothesis is perhaps less straightforward. As has been pointed out before, there is usually little writing done in Austrian CLIL classes, apart from students taking notes and filling in worksheets. The typical CLIL lesson therefore hardly offers any additional opportunity to practice writing. Nevertheless it is assumed that there is an indirect effect, i.e. that the improved general language ability will also translate into better writing. It should also be emphasised that the main focus in this case study is on general language ability and also general English writing skills. The study does not focus on subject-specific language or writing skills. It is mainly concerned with the question whether the provision of CLIL can help improve the general language skills of upper-secondary engineering students attending colleges referred to as 'HTLs' in Austria.

4.1.2. The situation at Austrian HTLs

'HTL' (which is short for 'Höhere Technische Lehranstalt') is the term used in Austria to describe an upper-secondary college of engineering, crafts and arts which takes five years to complete and ends with a school-leaving exam which qualifies graduates for university entrance. What is particularly noteworthy about these colleges is that they offer a wide range of specialisations which in many other countries are typically available at the tertiary level only. (Dalton-Puffer et al 2009: 20). Foreign languages, on the other hand, have traditionally played a rather marginal role in these colleges, which has also affected the status of the language teaching staff and the self-image of HTL students and

[1] At College B these additional EFL lessons have been included in the curriculum, while College A offers optional conversation classes.
[2] In the case of engineering colleges, these are regularly limited to two lessons per week.

graduates, who are typically expected neither to be particularly interested in languages nor to be particularly good at them. In fact, a common belief seems to be that quite a few (of the predominantly male) students choose this form of education in order to 'escape' what they see as a rather heavy load of foreign-language learning at other types of schools (Wilhelmer: 2008).

However, the trend towards increasing globalisation and internationalisation has also been highly visible in the engineering sector and there is little doubt that engineers will increasingly be expected to be able to use foreign languages, and English in particular, in the workplace. This was confirmed by a recent survey conducted among 1660 Austrian engineers, in which respondents were asked, among other things, how often they needed English at work or in their studies. 53.4% stated that they needed English frequently or very frequently, while 11.9 said they needed it very little and only 6.5% claimed they did not need English at all (Schneeberger et al 2008: 124). It was also observed that the need for English rose with age and position so that it can be concluded that knowing foreign languages, and English in particular, tends to become more important as a graduate moves up the career ladder. On this basis Schneeberger et al. (2008: 126) argue that, on the one hand, it is important to improve language teaching at the schools themselves but, on the other, it is also essential to make it clear to the students that they will have to continue to improve their language skills after graduation. Another interesting point is that foreign language skills are not only important to those working at large international groups but also those at small and medium-sized enterprises operating in global niche markets (Dalton-Puffer et al 2009: 19) as these are highly dependent on their export business. Obviously, engineering colleges need to respond to these trends and also the Austrian Ministry of Education has started an initiative to enhance the status of foreign languages, and English in particular, at engineering schools and to help engineering students achieve better performance levels (Dalton-Puffer et al 2008: 5). One of the instruments that might help achieve this aim is CLIL.

It has already been pointed out that the typical engineering student is believed to have little motivation to study languages for their own sake. An interesting point in this context is Lasagabaster's (2008: 33) claim that CLIL might help to overcome gender-based differences in language performance, the idea being that male students tend to be less intrinsically motivated to learn languages than females and thus need to be given additional reasons for improving their language skills. Lasagabaster feels that the focus

on the subject matter that characterises CLIL might provide this extrinsic motivation and may thus help to even out gender-based differences. Although this hypothesis is not borne out by Lasagabaster's own study, the idea that male students need to be motivated differently from female students may provide an additional reason for the provision of CLIL at schools which are (still) predominantly attended by men. Moreover, CLIL could be instrumental in improving attitudes towards foreign languages, which might not only have a beneficial effect while students still attend school but could also influence their willingness to continue improving their language skills after graduation, which, as has been pointed out above, might be important for their career development.

4.1.3. CLIL at the participating colleges

While CLIL activities[1] have been known in Austria since the early 1990s, also in the vocational sector, they have often tended to be fairly isolated and limited, with individual teachers deciding on their own to incorporate English activities in their teaching. The two colleges participating in this study, however, have opted for a more ambitious approach.

College B is the third largest engineering college in Austria.[2] In 1994 two teachers started an initiative to implement the concept of using English as a medium of instruction in content classes and quickly managed to gather a team of 12 teachers who shared this interest. What is particularly noteworthy is that from the beginning an effort was made to go beyond the usual practice and to implement CLIL as an integral part of one of the departments. Thus the PEA[3] project 'English as a Medium of Instruction in Science and Technology' was implemented at the Information Technology branch of the Electrical Engineering department. In addition to using English as a medium of instruction the curriculum was adapted specifically so that additional English lessons could also be offered to those opting for this programme. While in the first two years these are general EFL classes, a new subject, i.e. 'English for technical purposes' (ETP) is introduced in the third grade of college. English as a medium of instruction in content classes is introduced gradually and systematically, starting with general knowledge and science subjects and later widened to include the specialist subjects as well. In addition, every year every class undertakes a project involving hands-on tasks that students work on

[1] Formerly known under the label of 'Englisch als Arbeitssprache' or EaA (English as a working language).
[2] This description is based on Wilding et al 2008 as well as transcripts of interviews conducted with teachers at College B (Wilhelmer 2008).
[3] 'PEA' stands for 'Projekt Englisch als Arbeitssprache' (Project English as a working language).

independently for a week and a public presentation held in English at the end. Finally, all the students (rather than just the best performers) are prepared for the Cambridge First Certificate in English (FCE) when attending the fourth grade of college. As such international certification can provide a competitive advantage on the labour market (Schneeberger 2008: 126), the high pass rates (Wilding et al 2009: 33) are a remarkable achievement.

Moreover, internal training has been provided from the beginning to help improve the language skills of the CLIL teachers, who are typically subject specialists without any formal qualification in English. Team teaching, involving the specialist subject teacher and a language specialist, is used intermittently in the specialist subjects and in ETP. In most cases, CLIL lessons are used to revise topics already introduced in German rather than new material. Summing up, the main pillars on which the programme rests are organisational structure, internal staff training, team teaching and the annual project week (cf. Wilding et al 2009: 18-21). It should also be noted that the PEA programme is part of a general effort to provide engineering students not only with specialist skills and knowledge but also with soft skills such as project engineering, the ability to work both independently and as part of a team, presentation skills and language skills. As for the latter, languages other than English (French, Spanish, Italian, Chinese) are offered as optional subjects.

College A has opted for a different approach. While English as a medium of instruction has been used since 1993 in some subjects and on the initiative of individual teachers, in the winter term of 2005/6 a new project was started which involved implementing an ‘English-speaking’ programme at the Automotive Engineering department, the aim being to teach all theoretical subjects across the board (partly) in English (Reithuber 2006). Thus the programme is more comprehensive than the one implemented at College B, covering most of the subjects offered, the only exceptions being German (for obvious reasons) and Workshop. However, this does not mean that English is used exclusively in those subjects where CLIL has been implemented. In fact, there seems to be considerable variation, depending on the subjects and the teachers’ and students’ preferences. For example, students interviewed in 2008 (Schindelegger 2008) reported that, at their own request, very little English was used in Mechanics as this subject was perceived to be particularly difficult. Other subjects were taught (almost) exclusively in English and others were positioned somewhere in between. Nevertheless students reported that they

had CLIL classes every day, while College B students stated that there might be weeks without any CLIL at all (Schindelegger 2008; Wilhelmer 2008).

The two programmes are thus fairly different, which only confirms that CLIL does not stand for a uniform approach or programme, but is an umbrella term covering a wide range of activities (see Chapter 3.1.3.). Both programmes seem to have their strengths and weaknesses, with College B being in a pioneer position and offering a well-organised and well-supported programme which, apart from the project weeks, mainly involves reviewing some content in English. By contrast, the programme at College A is more comprehensive, but the interviews conducted in 2008 suggest that at that time it seemed less tightly planned and organised, with fewer efforts made to encourage cooperation among the teachers (Schindelegger 2008). In this context, it should be remembered, however, that at that time the project had not been running for very long and therefore was probably still suffering from some 'teething troubles'.

4.1.4. Test design and sample

To test the hypotheses specified above, a two-part written test was administered to a total sample of 88 students attending either College A or College B in May 2008. Overall, 41 students had received CLIL instruction in addition to their regular EFL classes, while 47 had not.

The first part consisted in a C-test, a variant of the cloze test, and was designed to provide a global statistic reflecting the overall language ability of the students. Like the cloze test (see Chapter 2.4.2.), the C-test offers the advantages of generally high reliability and objectivity. However, Carroll's (1980: 9f) claim that the cloze test "does not represent genuine interactive communication", of course, also applies to its relative, the C-test. In this context, it might be mentioned that Raatz and Klein-Braley (2002: 83) relate the construct of 'general language proficiency', as measured by the C-test, to Bachman's model of communicative language ability (see Chapter 2.4.3.1.), likening it to that part of language competence which Bachman terms 'organizational competence', involving lexical, morphosyntactical and graphological competence as well as textual competence. While they concede that C-tests are not suitable tools to specifically measure the pragmatic components of Bachman's model, they argue that "adequate or superior performance in the area of sociolinguistic competence can only be achieved if the basic

underlying organisational competence (= general language proficiency as assessed by the C-Test) is sufficient" (Raatz and Klein-Braley 2002: 83).

As the C-test does, therefore, not cover the whole spectrum of skills required for effective communication and is not regarded as a 'communicative' test, the second part, a writing task, was designed to meet the requirements of communicative language testing (see Chapter 2.4.3.2) in so far as students were put into a realistic communicative situation and asked to compose a piece of writing with a view to meeting a particular communicative purpose. On this basis, different aspects of their writing ability, including general language competence, were to be assessed.

4.2. The C-test

4.2.1. The C-test as a variant of cloze

Like the cloze test, which had its heyday in the 1970s and early 1980s, the C-test is an objectively scored, integrative test of general language proficiency based on the principle of reduced redundancy (see Chapter 2.4.2.). It was developed as a reaction to criticisms levelled at the cloze test which referred not to the underlying principle but technical factors such as length, possible test bias due to the choice of text, evidence suggesting that reliability and validity may vary with factors of test design and the frequent inability of native speakers to restore the original text (e.g. Raatz and Klein-Braley 2002: 77f). Cloze tests thus came to be seen as "unsatisfactory operationalizations of the concept of reduced redundancy" (Raatz 1985: 14). As a result, Raatz and Klein-Braley aimed to develop a test which would meet the following criteria:

- The new test should be much shorter, but at the same time it should have at least 100 items.
- The deletion rate and the starting point for deletions should be fixed.
- The words affected by the deletions should be a genuinely representative sample of the elements of the text.
- Examinees with special knowledge should not be favoured by specific texts, therefore the new test ought to consist of a number of different texts.
- Only exact scoring should be possible to ensure objectivity.
- Native speakers ought to be able to make virtually perfect scores on the test: 90% or higher. If native speakers cannot make scores higher than 90%, then the text should not be used for non-native speakers.
- The new test should be reliable, valid and easy to develop (Raatz and Klein-Braley 2002: 78).

The result of this quest to improve on the cloze test was the C-test, which was presented by Raatz and Klein-Braley in 1982. In contrast to the cloze test, which is based on one text, a C-test consists of five or six short texts, which ought to be authentic and, as Raatz and Klein-Braley (2002: 84) put it, "as normal as possible", which means that they should be neutral in content and contain no specialised vocabulary. Literary or humorous texts, for example, would be unsuitable. Naturally, they also need to be appropriate for the target group in terms of difficulty and should form a complete sense unit.

Another major difference between the cloze test and the C-test is that the latter involves deleting parts of words rather than whole words. As for the deletion system, the first sentence is left intact, after which the 'rule of two' applies.[1] Basically, this means that mutilation starts with the second word of the second sentence and consists in removing the second half of every second word ($n/2$ in the case of an even number (n) of letters, $(n+1)/2$ in the case of an odd number). In doing so, numbers and proper names are left undamaged and one-letter words are ignored. In the canonical C-test developed by Raatz and Klein-Braley the letters deleted are replaced by a simple line. Alternatively, the number of letters missing could be indicated (e.g. by dots or dashes), which is generally believed to make the test easier.[2] Once the necessary number of words have been mutilated (typically 20 or 25), the rest of the text is left intact (Raatz and Klein-Braley 2002: 75). It is claimed that this deletion system is likely to produce a random selection of words to be restored, which therefore form a representative sample of the word classes contained in the text (e.g. Raatz 1985: 16).

As for scoring procedures, a point is awarded for each word restored correctly and the final score is determined by adding up these points. The test thus yields a single score for each candidate which "represents the individual's global standing on the language proficiency continuum" (Raatz and Klein-Braley 2002: 78). In most cases it should be possible to use exact scoring, as recommended by Raatz and Klein-Braley, but alternative solutions which fit into the context syntactically and semantically can also be allowed. These are usually laid down after the test has been piloted with a group of native

[1] A number of researchers have also experimented with alternative deletion systems, producing tests of varying levels of difficulty (cf Grotjahn et al 2002: 108f, Sigott 2004:39-41). Moreover, it has been shown that the procedure needs to be adapted for different languages, depending on the morphological system of the language concerned (e.g. Grotjahn et al 2002: 96f).

[2] Grotjahn (1987: 227) opposes this practice as it may lead candidates to base their decision on how to fill in a blank on the number of letters contained in the alternatives considered. He argues that this "is surely not a component of the language processing competence intended to be measured by the C-test".

speakers. Another possibility to deal with ambiguities is to modify the deletion system for the items concerned, i.e. to remove fewer letters, in order to make the solutions more straightforward. In general, however, it is claimed that alternative solutions are comparatively rare (e.g. Raatz and Klein-Braley 2002: 80).

Generally speaking, the C-test has been well received and has generated a great deal of interest among researchers, making it "one of the most thoroughly studied approaches to the measurement of language proficiency" (Sigott 2004: 201). Various studies have attested it "extraordinarily good statistical qualities" (Grotjahn et al 2002: 98) with high levels of reliability and impressive concurrent validity in the form of high correlations with other measures of language proficiency such as other language tests, school grades or self-evaluation (Raatz and Klein-Braley 2002: 81). Nevertheless there is considerable controversy regarding the test's 'psycholinguistic validity', i.e. the question what exactly C-tests measure in terms of the psycholinguistic processes activated by learners when taking this type of test. Research in this field has involved methods such as 'think aloud' protocols, scrambled C-tests, computer tracking or error analysis but findings so far have been inconclusive. While some authors have provided evidence for both low-level and high-level processing,[1] others see the C-test as mainly tapping micro-level processing. To complicate matters further, some authors have suggested that the type of processing may vary with the proficiency level of the candidates (Grotjahn et al 2002: 104f; see also, for example, Sigott 2004: 203, Hastings 2002: 60).

Although these issues have not been resolved yet, the C-test is widely used in situations where a global measure of the candidates' general language ability is desired, such as, for example, in the case of placement tests. However, one problem that remains is the low face validity the C-test has among test users and test takers alike. They often "find it difficult to accept C-Tests as integrative language tests" (Raatz and Klein-Braley 2002: 82) and tend to see them as reading comprehension tests or even intelligence tests. An interesting point in this context is that teachers tend to overestimate the difficulty of C-tests before administering them, which may be due to a lack of familiarity with this test procedure (cf Sigott 2004: 54f). This was also the case in the present study, where teachers consulted in the design process of the C-test voiced considerable concern about

[1] High-level processing can be defined as "processing above sentence level" (Sigott 2002: 67). See also Sigott (2004: 92f) on the use of different terminology in this field. Sigott himself uses terms referring to syntactic structure such as processing on the word, sentence or text level. He thus sees high(er) and low(er)-level processing as relative terms.

the level of difficulty, but seemed somewhat less worried after they had tried out the test with some of their students (personal communication).

As mentioned above, another point of criticism is that the C-test is not a communicative test, which is why in the present case study it was complemented with a task aiming to test the communicative competence of the candidates, at least as far as their writing ability is concerned.

4.2.2. Development of the C-test

The C-test for this case study was developed in accordance with the instructions provided by Raatz and Klein-Braley (1985). The first step consisted in selecting appropriate texts. According to Raatz and Klein-Braley, these should be about 60 to 70 words long and should be of a general nature, not requiring any specialist knowledge of content or vocabulary. Moreover, the texts should be authentic.[1] In the present case texts were sourced from the internet (e.g. the BBC website), the print media and non-fictional books.

To decide whether the texts were suitable for inclusion in the test, their level of difficulty had to be established. On the one hand, the opinion of teachers working at an engineering college was sought and some texts were discarded as they had been judged as too difficult (personal communication). On the other hand, this subjective assessment was complemented with a few ratios designed to provide an objective measure of readability. Klein-Braley (1985), for example, identifies the type-token ratio, which measures lexical variation, as well as mean sentence length, an indicator of syntactic complexity, as the best predictors of C-test difficulty. For the present case study the following indicators were selected:

- syntactic complexity: sentence length in words, Flesch Reading Ease formula and the Flesch-Kincaid readability index
- lexical variation (type-token ratio) and frequency.

To measure syntactic complexity, three indicators were chosen, with the Flesch Reading Ease formula (‘Flesch RE’) and the Flesch-Kincaid grade level index (‘FKGL’) based on

[1] Exceptions are possible with learners whose level of proficiency does not permit the use of authentic texts. In this case the texts could be sourced from text books designed for the same level. Alternatively it is possible “to ‘doctor’ the texts slightly” (Raatz and Klein-Braley 1985: 20) to reduce their level of difficulty.

both word length and sentence length.[1] They are both implemented in MS Word. With the Flesch RE formula, the score is inversely related to text difficulty, i.e. the higher the score the easier the text is estimated to be. The theoretical maximum is 120 but scores up to 100 can normally be achieved. In simple terms, scores between 60 and 70 are considered 'standard', while texts with scores between 30 and 49 are labelled 'difficult' (*Readability formulas*). The Flesch-Kincaid Grade Level index was developed on the basis of the Flesch Reading Ease formula and is said to indicate the grade level for which a text is considered suitable.[2]

As for the lexical aspect, the level of difficulty of a text is taken to vary with lexical variation (as measured by the relationship between types and tokens, known as type-token ratio, TTR[3]) and the frequency of the lexical items used. As for the former, the general rule is that greater lexical richness is associated with a higher type-token ratio, the theoretical maximum being one, which would mean that not a single word is repeated in a text (e.g.Ellis and Barkhuizen 2005: 155). With regard to the latter, the proportion of the 1000 most frequent words (K1), the 2000 most frequent words (K1+K2), academic words and other words was analysed.[4] Table 2 shows the readability indices for the eight texts that were selected for the next step, i.e. the first pilot test with native speakers of English. The texts themselves are reproduced in Appendix 1.

[1] The specific formula for the Flesch Reading Ease test is RE = 206.835 – (1.015 x ASL) – (84.6 x ASW), with ASL standing for Average Sentence Length (in words) and ASW for Average number of Syllables per Word. The formula for the Flesch-Kincaid Grade Level index is FKGL= (0.39 x ASL) + (11.8 x ASW) – 15.59 (*Readability formulas*).

[2] It should be noted, however, that in MS word the maximum grade level indicated is 12. This means that a grade level reported as grade 12 can be grade level 12 or higher.

[3] The type-token ratio is defined as "the total number of different words used (types) divided by the total number of words in the text (tokens)". (Ellis and Barkhuizen 2005: 154).

[4] This was done with the help of a tool found on *Lextutor* (www.lextutor.ca/vp/eng).

	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8
Title	Internet	China	PowerPoint	Broadband	Talent wars	Happy Planet Index	Mobile phones	Energy
Source	BBC News[1]	BBC News[2]	Reilly[3]	BBC News[4]	Business Spotlight[5]	FoE website[6]	BBC News[7]	Miller[8]
Sentence length	16.20	13	20	17	17.30	21.20	18.60	22.20
Flesch RE	58.70	56.90	60.60	64.10	50.10	51.80	42.20	30.20
FKGL	9.00	8.50	9.70	8.50	10.50	11.20	11.90	12.00
TTR	0.72	0.72	0.72	0.72	0.79	0.75	0.76	0.79
K1 in %	83.13	82.05	87.50	80.46	86.67	84.09	74.47	78.89
K1+K2 in%	90.36	87.18	90.00	86.21	89.53	88.64	82.98	86.67
Academic in %	3.61	7.69	10	2.30	7.62	4.55	8.51	8.89
Other in %	6.02	5.13	0	11.49	2.86	6.82	8.51	4.44

Table 2 Readability ratios for texts selected for first pilot test

Generally speaking, the ratios suggest that Texts 1-4 can be characterised by greater readability in terms of lower syntactic complexity and lower lexical variation than Texts 5-8 and can thus be considered less difficult. With regard to lexical frequency, a somewhat different picture emerges, with Texts 5 and 6 apparently 'easier' than Texts 2 and 4. In this context, one must not forget, however, that these ratios can only cover certain aspects. When interpreting the information included in Table 2, one must also take the topics of the texts into consideration. For example, Text 4 ('Broadband') shows a comparatively large share of words outside the range of the 2000 most frequent words. It should be added, however, that most of these words belong to the semantic field of 'computing and internet' and can therefore be assumed to be known to students of engineering. To some extent, this is also true of Texts 7 and 8, which deal with mobile phone use and forms of energy respectively.

[1] Thompson 2002.
[2] "Quick Guide: China" 2006.
[3] Reilly 1997: xxii.
[4] "Quick Guide: Broadband" 2006
[5] "The war for talent" 2007:9
[6] Thompson et al 2007: 1
[7] "Mobile phone use" 2007
[8] Miller 1991: 408

It should also be added that with 80 to 94 words the texts exceeded the length recommended by Raatz and Klein-Braley. However, in the interest of preserving the texts as sense units and providing sufficient context, they were left uncut, meaning that there was a slightly longer 'run-out' after the last blank.

Having been selected for the first pilot test, the texts were then subjected to mutilation according to the C-principle, as described above. The 'rule of two' was applied and the blanks marked by one line, the length of which depended, however, on the number of letters deleted. The reason for this procedure was that in this way a hint could be given as to how long the original word was without inviting a 'letter-counting' strategy.[1] In each text 20 words were mutilated.

4.2.3. The trialling phase

The trialling phase consisted of two types of pilot tests, one administered to a group of native speakers and the second one to representatives of the test target group, students attending a college of engineering, arts and crafts in Vienna. Piloting the test with native speakers is recommended for two reasons (e.g. Grotjahn et al 2002: 98):

1. To establish the level of difficulty among native speakers
2. To identify acceptable alternatives (unless an exact scoring system is to be used).

Regarding the first point, it should be remembered that one of the advantages the C-test is said to have over the cloze procedure is that "native speakers ought to be able to make virtually perfect scores on the test" (Raatz and Klein-Braley 2002: 78). In practical terms, this means that when administering the pilot test to "a control group of adult educated native speakers or teachers of the language" success rates should be at least 95% (i.e. facility indices *p* should be 0.95) and that texts with rates of less than 90% ($p<0.9$) are unacceptable and should be discarded (Raatz and Klein-Braley 1985: 21).

Table 3 shows the results of the first pilot test, which was administered to 18 native speakers, 14 of which were involved in EFL teaching:

[1] Moreover, to make the deletion principle clear to test takers, they were to be instructed by invigilators that 'about half' of the words had been deleted and the test paper itself included instructions and an example that was to be explained at the beginning of the test (see appendix).

	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8
Title	Internet	China	PowerPoint	Broadband	Talent wars	Happy Planet Index	Mobile phones	Energy
Success rate	96.7%	96.4%	96.7%	93.9%	95.3%	93.6%	96.9%	97.8%

Table 3 Success rates in Pilot test 1 (native speakers)

As Table 3 shows, two texts ('Broadband' and 'Happy Planet Index') were below the 95% threshold. As oral feedback from the native speakers taking the test also suggested that individual passages of these texts were perceived to be rather challenging for students, these texts were discarded. The remaining texts were used for the second phase of the trialling process, which consisted in administering the second pilot test to a trial population of 27 students attending an engineering college in Vienna who could be considered representatives of the test target group.

As usual, the scoring procedure consisted in awarding a point for each correct solution. Orthographic mistakes were marked as incorrect and in a few cases alternatives, which had been laid down in the first phase of trialling, were accepted. Subsequently, a statistical analysis was carried out, calculating facility, discrimination and reliability indices (see Chapter 2.5.3.) in order to establish the suitability of the test for the purposes of the case study. Table 4 summarises the results of the analysis:

	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6
	Internet	China	PowerPoint	Mobile phones	Talent wars	Energy
Mean score (max=20P)	15.26	12.78	13.22	14.22	9.85	11.00
Facility index (P value)	0.76	0.64	0.66	0.71	0.49	0.55
Discrimination index	0.28	0.32	0.35	0.29	0.40	0.44
Correlation with total score (Pearson)	0.7742	0.8463	0.7120	0.7503	0.8147	0.7726
Reliability (Cronbach's Alpha)	0.86					

Table 4: Item analysis of texts in Pilot Test 2

As discussed in Chapter 2.5.3., the facility index (P value) corresponds to the percentage of test takers providing the correct (or an acceptable) solution. It estimates how difficult or easy a test is for a particular target group. As the test is supposed to help distinguish between students of different ability levels, texts with very high or very low facility indices are considered unsuitable, as they can be expected to be solved by virtually every or hardly any candidate respectively. The range of values considered acceptable varies between 0.2 to 0.8 (Raatz and Klein-Braley 2002: 86) and 0.3 to 0.7 (Raatz and Klein-Braley 1985: 22). Overall, a P value of 0.5 is considered ideal, although a value of up to 0.6 is seen as acceptable (Klein-Braley 1985: 23).

As Table 4 shows, the texts seem to be relatively easy. Text 1, in particular, has a high facility value, which could be considered too high. Nevertheless, it was decided to retain it as an 'ice-breaker item' designed to 'ease test takers in'. In this way, anxiety can be reduced and it can be ensured that "every test subject understands exactly what the C-Principle demands from him or her" (Klein-Braley 1985: 23). At this stage, it was left open whether or not Text 1 would be included in the calculation of overall scores in the case study itself. It was decided that if it proved to have an even higher facility index then, total scores could be calculated on the basis of the other texts only in order to ensure meaningful results. Finally, overall facility across all texts was 0.64, and 0.61 if only Texts 2-6 were included.

In order to ensure that individual test items discriminate sufficiently between test takers of different levels of language proficiency, the discrimination index should not be lower than 0.3 (Bachman 2004: 130; see Chapter 2.5.3.). Again, the data in Table 4 show that the items do not all meet this criterion. Nevertheless they were considered close enough to be retained. In addition, the correlations of the individual items with the total score were determined, the idea being that items that do not correlate positively might distort the results and should be discarded. Finally the reliability of the test was estimated using Cronbach's Alpha and found to be 0.86. While Raatz and Klein-Braley (1985: 22) see a value of 0.9 as ideal, they consider values between 0.8 and 0.9 as acceptable so that the test can be considered sufficiently reliable.

On the basis of the statistical data obtained in the trialling phase and additional considerations, such as the desirability of an 'ice-breaker', it was decided to retain all six

texts and use them in the main study. The next step was therefore to administer the test to students of College A and College B, which was done in May 2008.

4.2.4. The case study

As pointed out before, the purpose of including a C-test in the case study was to obtain an objective and global measure of the language ability of the test takers in order to determine whether there were significant differences between those students who had undergone CLIL instruction and those who had not.

First, the same statistical analysis was carried out as for the second pilot test. Table 5 summarises the results:

	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6
	Internet	China	PowerPoint	Mobile phones	Talent wars	Energy
Mean score (max=20P)	14.97	12.88	13.64	14.75	11.46	12.82
Facility index (P value)	75%	64%	68%	74%	57%	64%
Discrimination index	0.28	0.29	0.34	0.29	0.35	0.29
Correlation with total score (Pearson)	0.776	0.760	0.802	0.803	0.761	0.749
Reliability (Cronbach's Alpha)	0.865					

Table 5: Item analysis of texts in the case study

On the whole, the results were similar to those obtained in the trialling phase. The better results achieved on the last two texts may be due to the fact that in the pilot study several students had hardly attempted to tackle those texts, presumably for reasons of inefficient time management. In the operational phase invigilators were therefore instructed to indicate when five minutes had passed and test takers were supposed to move on to the next text.

As can be seen from Table 5, Text 1 still has a high facility value. Nevertheless it was not excluded from further analysis due to the fact that a closer analysis of discrimination indices showed that with 0.3 this text had the second-highest discrimination index in one

of the participating colleges (College B). Table 6 shows the average scores obtained by the different groups of test takers on the individual texts and on the test as a whole.

	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	TOTAL SCORE
	Internet	China	PowerPoint	Mobile phones	Talent wars	Energy	
College A CLIL (AC)	16.38	14.86	15.52	16.90	14.57	15.33	93.57
College A non-CLIL (ANC)	14.61	12.52	13.30	14.04	10.48	12.61	77.57
College B CLIL (BC)	15.85	13.55	13.50	14.70	11.10	12.25	80.95
College B non-CLIL (BNC)	13.33	10.92	12.42	13.58	9.96	11.29	71.50

Table 6: Average scores C-test

The scores already suggest that the AC group clearly outperformed the other groups, while the BNC group seemed considerably weaker. To establish the statistical significance of these differences, the data were analysed using the t-test procedure. Table 7 shows the results for the total sample, combining both schools participating in the study, while Table 8 breaks them down further by distinguishing between the schools. On this basis Figure 5 illustrates the main information graphically.

	No of test persons	Range (total score)	Average total score	Standard Deviation (SD)	T value
Total CLIL	41	73-107	87.415	11.032	5.622**
Total non-CLIL	47	54-101	74.468	10.550	
	**Difference highly significant (level of significance 99%)				

Table 7 Statistical analysis of C-test scores, total sample

	No of test persons	Range (total score)	Average scores	Standard Deviation (SD)	T value
College A CLIL (AC)	21	77-103	93.571	7.521	6.024**
College A non-CLIL (ANC)	23	60-101	77.565	9.825	
College B CLIL (BC)	20	73-107	80.950	10.526	2.960**
College B non-CLIL (BNC)	24	54-91	71.500	10.558	
	** Difference highly significant (level of significance 99%)				

Table 8 Statistical analysis of C-test scores, by school

As can be seen from Tables 7 and 8, the differences between the CLIL and the EFL-only students are highly significant for the sample as a whole as well as for each of the schools participating in the study. This clearly supports the hypothesis that CLIL students outperform their EFL-only peers with regard to general language ability.

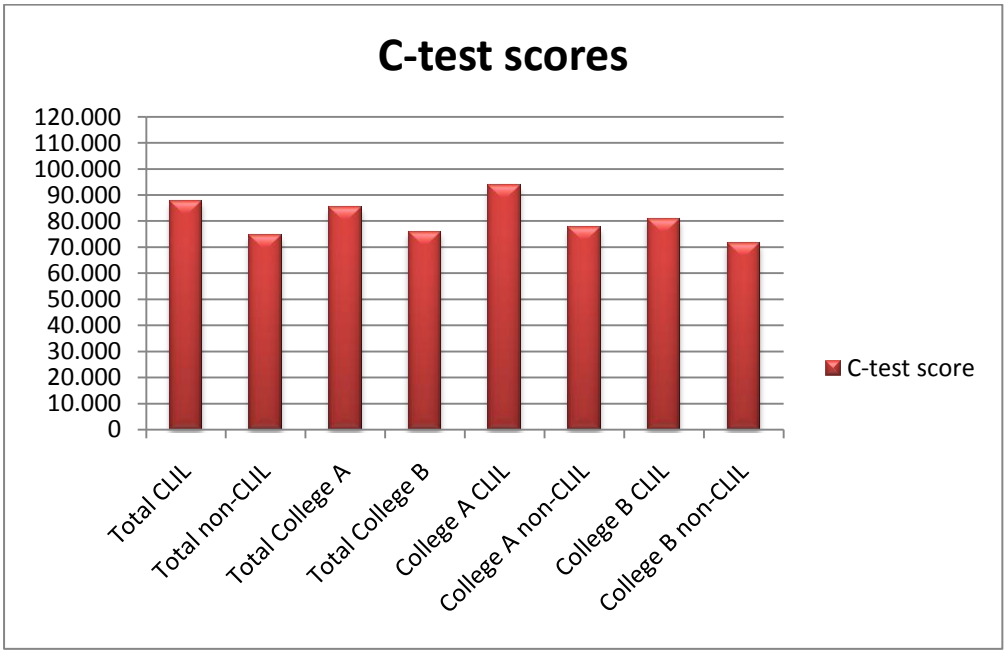

Figure 5 C-test scores total sample and individual groups

It should also be noticed, however, that there is rather high variability of the scores obtained, with the AC group the most homogeneous. The highest score was obtained by a student in the BC group, while average scores were significantly lower in this group than in the AC group. Apart from the BNC group, scores above 100 were achieved in all groups. The scores also suggest that there is a clear difference in performance between the two colleges. Therefore they were also analysed in this respect. Tables 9 and 10 show the findings of this analysis:

	No of test persons	Range (total score)	Average total score	Standard Deviation (SD)	T value
Total College A	44	60-103	85.205	11.878	3.782**
Total College B	44	54-107	75.796	11.456	
	**Difference highly significant (level of significance 99%)				

Table 9 Statistical analysis of C-test scores by school

	No of cases	Range (total score)	Average scores	Standard Deviation (SD)	T value
College A CLIL (AC)	21	77-103	93.57	7.521	4.435**
College B CLIL (BC)	20	73-107	80.95	10.526	
College A non-CLIL (ANC)	23	60-101	77.57	9.825	2.037*
College B non-CLIL (BNC)	24	54-91	71.50	10.558	
	* Difference significant (level of significance 95%) ** Difference highly significant (level of significance 99%)				

Table 10: Statistical analysis of C-test scores CLIL vs non-CLIL groups, by school

As Tables 9 and 10 show, there is a statistically significant difference between the results achieved by students of the two schools, with the overall scores and those of the CLIL students showing highly significant differences. The difference in scores between the two CLIL groups is particularly striking. While there may be other influencing factors, which cannot be investigated here, it seems plausible to assume that, to a large extent, this is due to the different approaches and implementations of CLIL in the two schools. It has already been pointed out that the two programmes are rather different, with the CLIL programme at College A being conceived as a comprehensive, across-the-board programme and thus providing more and more regular input of the foreign language while at College B CLIL instruction is provided at irregular intervals (see Chapter 4.1.3.). In this context, we must not forget that CLIL is a label that stands for a large variety of approaches and that this fact naturally affects the comparability of test results. Another problem that limits the interpretability of the results is that no data were available on the level of language competence of the students at the time they entered the respective engineering colleges. Due to the self-selection of students undergoing CLIL instruction it may well be the case that the CLIL students already showed higher competence before they joined the CLIL programmes. However, the influence of this factor cannot be estimated given the lack of data available.

4.3. Text production

As mentioned before, the second part of the test consisted in a writing task. On the one hand the idea was to provide another (subjective) measure of the candidates' general language ability, thus complementing the objective method of the C-test, while on the other hand the question was to what extent the students' communicative writing ability was affected by CLIL provision.

4.3.1. Task design and assessment

The students were asked to write a letter or e-mail to a (hypothetical) host family living in New York with whom they were going to spend two weeks as part of an intensive language week offered by the school. As such language weeks are offered at both schools it was assumed that the students should be able to identify with their role in this communicative situation and that it could thus be regarded as an 'authentic' task from their point of view. Basically, it was a free-writing task as the students were not required to use any particular structures or vocabulary. Nevertheless, in the interest of comparability some guidance was provided in the form of several suggestions as to the content of the letter or e-mail. These included providing personal information, reference to previous stays abroad but also questions to be put to the host family and a positive conclusion (see Appendix 2). Apart from making the texts easier to compare, these suggestions were also designed to ensure that the task did not make too many demands on the students' creativity and imagination, qualities which the test was not supposed to measure and which could potentially have distorted the results, reducing the test's validity (cf. Hughes 2003: 90).

Assessment was to be carried out by means of an analytic rating scale (see Appendix 3). In spite of the obvious advantages of holistic rating scales in terms of time and effort to be invested in the marking process (see Chapter 2.5.1.2), an analytic scale seemed preferable in the context of the present case study as such scales generally provide a more differentiated picture of a learner's writing ability and it was felt that valuable insights might be gained from analysing different sub-skills. As for the categories to be included, the main requirement was that they ought to reflect what was seen as key components of writing ability. These can be derived from the general models on communicative language ability (see Chapter 2.4.3.1.), which suggest that a good writer needs to have

good language skills (i.e. a good command of grammar and vocabulary), textual competence (i.e. the ability to compose coherent and cohesive texts) as well as sociolinguistic or pragmatic competence (i.e. the ability to use language appropriately in any given context). Therefore these were the categories that needed to be covered by the rating scale.

Different existing rating scales were considered and finally the scale developed for the new common school-leaving exam in Austrian grammar schools (Friedl and Auer 2007: 93) was chosen as a basis and slightly adapted to fit the task. The scale consists of four equally-weighted categories reflecting different aspects of writing ability, i.e. task fulfilment, organisation, grammar and vocabulary, with potential scores ranging from 0 to 5 for every category (see appendix). In the field of lexico-grammar, this qualitative assessment was complemented by a few quantitative ratios.

4.3.1.1. Task fulfilment

It is in this field that some adaptations had to be made to the original rating scale in order to fit the task. In particular, those descriptors that refer to factual knowledge or the quality of the arguments used had to be eliminated as they were geared to the assessment of argumentative essays and did not suit the text type chosen for the present case study. The remaining aspects thus were:

- The degree of task fulfilment
- Relevance
- Appropriateness in terms of text format, length and register.

Considering these aspects the scripts produced were assessed according to the degree to which the content points listed in the instructions had been covered and elaborated. Some flexibility was allowed here as these points had been worded as ‘suggestions’ and not all of the points were always applicable (e.g. previous stays abroad). Another important aspect was the degree to which the texts produced were relevant and the format appropriate. For example, the register was considered as well as structural elements such as an opening and closing that was appropriate to the text type (letter or e-mail). Finally, an important aspect in evaluating the texts was the ability of the writer to build rapport with the host family. In terms of communicative purpose, this seems an essential requirement as the main rationale for writing to a prospective host family is to start

building a harmonious and positive relationship at an early stage. The assessment thus took into account to what extent the test persons were aware of the communicative purpose of the text and whether this purpose was achieved.

To obtain a high score in this category, a test taker thus had to produce a text that covered and elaborated (most of) the points mentioned in the instructions, was relevant and appropriate and was likely to achieve the communicative purpose of the task by building rapport with the recipients. Finally, it should be added that the category of task fulfilment was defined as a 'veto category' (cf. Friedl and Auer 2007: 92). In fact, two texts which completely failed to fulfil the task were rated 0 and, as a consequence, eliminated from the sample.

4.3.1.2. Organisation

No changes were made to the scale in this category, which was designed to measure the textual competence of the writer, i.e. his or her ability to create a text rather than "a random collection of sentences" (McCarthy 1991: 35). The aspects to be considered were:

- Overall structure
- Paragraphing
- Use of connectives
- Editing mistakes and punctuation

Basically, this category concerns the extent to which a script exhibits a clear arrangement and progression of ideas and can thus be considered coherent. Put simply, coherence refers to "the feeling that a text hangs together, that it makes sense" (McCarthy 1991:26) and a coherent script is one in which ideas are clearly and logically arranged so that one paragraph leads to another and transitions are smooth. In practical terms this means that if a script is coherent "the reader does not have to stop and reread it in order to understand the connection between its sentences or paragraphs" (Tankó 2005: 175). Coherence is also helped by good and meaningful paragraphing, where each paragraph of a text develops an aspect of the main topic and each sentence in a paragraph helps to develop the relevant subtopic. If irrelevant sentences or paragraphs are included, this will break the flow and reduce coherence.

While the coherence of a text "is something created by the reader in the act of reading" (McCarthy 1991: 26), the relationships between sentences or paragraphs can be made visible by using cohesive devices such as 'connectives'.[1] It should also be considered that cohesion is not a necessary precondition for coherence, but serves to facilitate understanding by clearly marking structural relationships. While a well-organised script can be coherent without exhibiting any cohesive devices, the use of such devices alone will not make an otherwise unstructured piece of writing coherent (cf Tankó 2005: 176, Widdowson 2007: 49f). Finally, the scale also refers to editing mistakes, which can affect the flow of the script, and punctuation, with the latter helping readers understand by "showing where one set of ideas ends and where the next begins" (Tankó 2005: 180), in the same way as prosodic features facilitate understanding in spoken discourse.

To achieve a high score in this category, a script had to exhibit a logical and coherent overall structure with smooth transitions, good use of paragraphing as well as a range of appropriate connectives, going beyond the most basic ones such as *and*, *but* and *because*. Moreover, it had to be (virtually) free of editing and punctuation errors.

4.3.1.3. Grammar

The third category aimed to assess grammatical competence. Generally speaking, this involves "checking whether candidates are familiar with the form, meaning and use of a range of grammatical structures that can be expected to occur in a communication situation" (Tankó 2005: 219). This category thus addresses the following aspects:

- Accurate use of grammar and structures, frequency of morphological and syntactic errors (e.g. agreement, tense, word order, articles, pronouns)
- Variety of structures and frequency of complex structures
- Effect on communication

It seems essential to consider the first two aspects, i.e. the accuracy with which the student applies grammatical rules and uses grammatical forms and the complexity and variety of the structures used, together as a trade-off can often be identified between accuracy and structural complexity (cf. Ellis and Barkhuizen 2005: 144). Obviously, it is

[1] In general, cohesion can be established by a number of instruments such as reference, conjunction, ellipsis and substitution as well as the creation of lexical chains (Tankó 2005: 177-180), cf McCarthy 1991, Halliday and Hasan 1976), which were all considered in awarding a score in this category. The analysis of differences between the groups, however, focused on connectives.

easier for test takers to achieve high accuracy if they restrict themselves to simple structures they can manage easily while those who take greater risks and attempt to use a wider range of more complex structures are more likely to make mistakes. Moreover, in assessing the seriousness of an error, it is essential to consider its effect on communication. While some minor inaccuracies are tolerated even in the top band, errors that interfere with communication or even lead to communication breakdown weigh much more heavily and lead to scripts being downgraded.

To be awarded a high score in this category a text thus had to be characterised by accurate use of a variety of structures, including complex ones, as well as high accuracy in terms of categories such as verb forms, tenses, plural, word order and prepositions. In addition, such a script had to be free of errors that interfered with understanding, while occasional and minor inaccuracies were tolerated.

4.3.1.4. Vocabulary

The last category assessed the lexical competence of test takers, focusing on the following aspects:

- Range of vocabulary and choice of words
- Accurate form and usage
- Orthographic control
- Effect on communication

As with the previous category, there tends to be a trade-off between range of vocabulary and the accuracy with which it is used. Range of vocabulary refers to the writer's "ability to use an adequately broad vocabulary within a script" (Tankó 2005: 258). What is considered 'adequate', of course, depends on the task set. In the case of the present study the task did not require any specialised vocabulary but mainly asked test takers to write about personal and familiar topics. To be given a high score, students had to prove that they possessed a large enough repertoire to avoid undesirable repetition and to express ideas clearly and precisely through appropriate choice of words. Extensive lifting from the prompt and substantial L1 interference, on the other hand, were seen as signs of poor lexical range and choice.

Accuracy, on the other hand, refers to choosing the correct form of a lexical item and its correct usage. Obviously, this is easier to achieve if test takers limit themselves to very basic vocabulary, so that these two aspects need to be weighed against each other. In awarding a score, attempts to use a wider range were thus acknowledged even if test takers did not always use less frequent words accurately or did not exhibit full orthographic control of them. Finally, as with the previous category, errors made were assessed with regard to their effect on communication. In other words, errors that obscured meaning and rendered an expression incomprehensible were considered more serious than those that did not interfere with understanding.

To achieve a high score in this category, a script thus had to exhibit accurate and appropriate use of a wide range of relevant vocabulary, expressing clear ideas and featuring few, if any, orthographic errors.

4.3.2. Test results

To first consider the global test results, Tables 11 and 12 provide an overview of the average scores achieved by different groups of test takers out of a total of 5 for each category. While Table 11 refers to the total sample (i.e. both schools combined), Table 12 breaks the results down further by distinguishing between the participating schools. Figures 6 and 7 represent these results graphically. Following this global overview, the individual results will be discussed by category of assessment.

	No. of cases	Task fulfilment	Organisation and structure	Grammar	Vocabulary	TOTAL
Total CLIL	39	3.897	2.949	3.564	3.718	14.128
Range CLIL		1-5	2-5	2-5	2-5	
Total non-CLIL	47	3.021	2.404	2.511	2.936	10.872
Range non-CLIL		1-5	1-4	1-5	1-5	

Table 11 Average scores on writing task, total sample

The overall impression at this point is that the results of the writing task confirm those of the C-test as the CLIL students achieved higher averages throughout. However, the extent to which these differences are significant depends on which competence one focuses on, as will be discussed below. Moreover, the rather wide range suggests considerable

variation in student performance, with the CLIL students slightly more homogeneous than the non-CLIL ones on this global level.

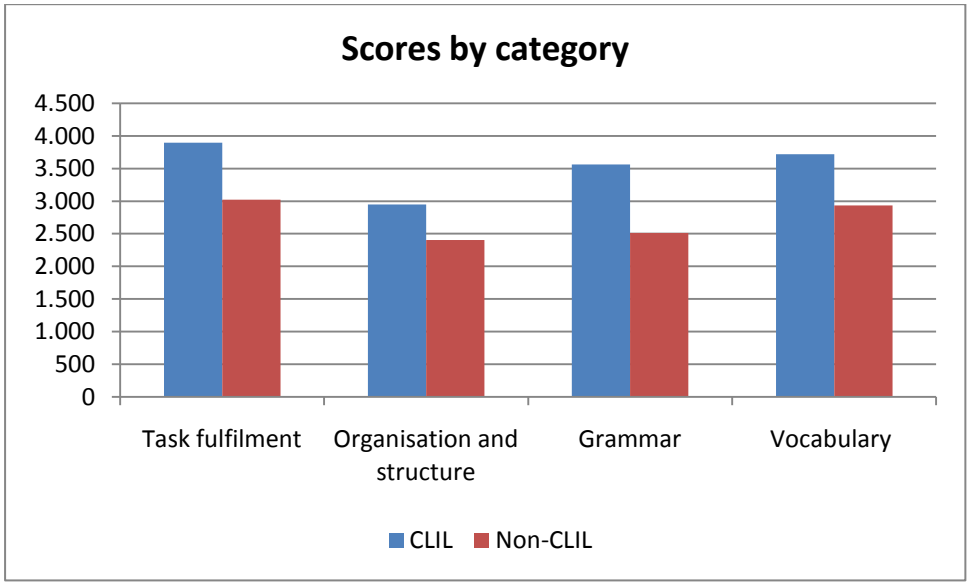

Figure 6 Average scores by category, total sample

If we further break down the results by considering the school attended, the following results are obtained:

	No. of cases	Task fulfilment	Organisation and structure	Grammar	Vocabulary	TOTAL
College A CLIL (AC)	21	4.238	3.143	3.810	3.952	15.143
Range AC		2-5	2-5	2-5	3-5	9-18
College A non-CLIL (ANC)	23	3.000	2.609	2.739	3.130	11.478
Range ANC		1-5	2-4	2-5	2-5	7-18
College B CLIL (BC)	18[1]	3.500	2.722	3.278	3.444	12.944
Range BC		1-5	2-5	2-4	2-5	9-18
College B non-CLIL (BNC)	24	3.042	2.208	2.292	2.750	10.292
Range BNC		1-5	1-3	1-4	1-4	4-16

Table 12 Average scores on writing task, by school

[1] As mentioned above, two texts had to be eliminated from the sample due to the test takers' complete failure or refusal to fulfil the task.

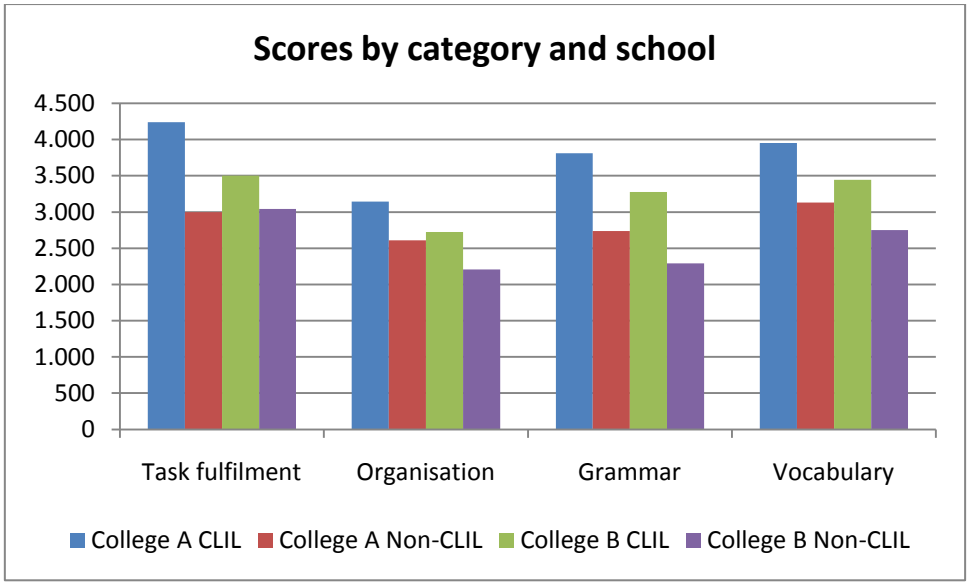

Figure 7 Average scores by category and school attended

Again, the findings resemble those obtained from the C-test. If we consider the results by category and school, we see that the CLIL students at College A considerably outperform the other groups, the only exception being the organisational aspect, where their advantage, while still there, is less obvious. There seems to be a clear difference between CLIL and non-CLIL students at both schools but overall the results of College B students are on a somewhat lower level. Overall, the highest levels of performance were reached in the categories of task fulfilment and vocabulary, and the lowest in the field of organisation, while the scores suggest that the dimension showing the greatest differences between the CLIL and non-CLIL students is grammar, not, as one might perhaps have expected, vocabulary.

As is the case with the overall sample, the individual groups show a wide range of performance. If we consider the total scores, for example, we see that the highest score (18 out of 20) was reached by students in all groups except for the BNC group.[1] While it is true that this score was obtained by three AC students (14%) but only one each in the ANC and BC groups (about 5% each), it still goes to show that high performers are to be found in all of these groups. At the other end of the spectrum, a score of less than 10 was achieved by 5 ANC (22%) and 8 BNC (33%) students, but also in the CLIL groups one candidate each (i.e. around 5%) fell into this category. Thus, while certain overall trends

[1] As mentioned in Chapter 4.2.4., the analysis of the C-test scores showed a similar pattern.

can clearly be observed, both the CLIL and the non-CLIL groups seem, at the same time, to be fairly heterogeneous.

4.3.2.1. Task fulfilment

Tables 13 and 14 summarise the results in the field of task fulfilment:

	No of test persons	Range	Average scores	Standard Deviation (SD)	T value
Total CLIL	39	1-5	3.897	1.119	3.697**
Total non-CLIL	47	1-5	3.021	1.073	
	**Difference highly significant (level of significance 99%)				

Table 13 'Task fulfilment', total sample

	No of test persons	Range	Average scores	Standard Deviation (SD)	T value
College A CLIL (AC)	21	2-5	4.238	0.831	4.213**
College A non-CLIL (ANC)	23	1-5	3.000	1.087	
College B CLIL (BC)	18	1-5	3.500	1.295	1.248
College B non-CLIL (BNC)	24	1-5	3.042	1.083	
	**Difference highly significant (level of significance 99%)				

Table 14 'Task fulfilment', by school

As can be seen from Table 14, the AC group clearly outperformed not only the non-CLIL group from the same school but also both groups at College B. The difference in average scores between the CLIL and the non-CLIL students is highly significant for the total sample as well as for the College A groups but not significant for the College B groups.

A more detailed analysis shows that the main differences between the groups lie in the appropriateness of text format, length and register as well as in the ability to build rapport with the host family. With regard to opening the text in an appropriate manner, there were clear differences between the groups at College A, with 90% of the AC group, but only 40% of the ANC group using an appropriate salutation.[1] At College B the groups were much more homogeneous in this respect, with the majority choosing an appropriate opening in both the BC and the BNC groups. With regard to closing the text it can be said that at both schools the non-CLIL groups were less likely to end the text appropriately

[1] Salutations considered inappropriate include, for example, "Hello guys!"(ANC4) or "Hello Family Ferguson!" (ANC11).

than their CLIL counterparts. It is difficult to say, however, whether this reflects a lack of awareness of text type requirements or can be traced back to time management problems as the non-CLIL students in general seemed to have more problems writing a text of the required length (150-200 words) in the time given. The texts produced by about a quarter of the ANC and about one third of the BNC groups were below the required length and ended abruptly.

While the content points included in the instructions were mostly covered by all the groups, the questions to the host family were the most likely to be neglected, presumably because this required more initiative from the students as they had to decide themselves what kind of questions might be appropriate in the given context. No clear picture emerged, however, with regard to any systematic differences between CLIL and non-CLIL students in this respect. Whereas fewer than 10% in the BNC and AC groups did not include any question at all, more than 20% of the ANC and almost 40% of the BC students neglected this point.

In terms of register, the differences were particularly obvious between the College B groups, with about 50% of the BC texts, but only about 8% of the BNC texts showing adequate register throughout. The College A groups were more homogeneous in this respect. With 57% the AC group was about 10 percentage points ahead of their non-CLIL peers. It can thus be said that the BNC group clearly underperformed the other groups in this respect, showing little awareness of pragmatic requirements in this field.

A major difference can also be identified in the extent to which the communicative task of establishing rapport was fulfilled. As pointed out above, this is an important requirement in terms of the communicative purpose of the given text type. Typical examples of strategies contributing to positive rapport management would be, for example, expressing gratitude to the host family for the opportunity to stay with them (examples 1-5), saying how much you look forward to seeing their country (examples 6-8) or showing interest in them as a family (examples 9-10). A few examples may serve to illustrate how the students employed these strategies:

> (1) At first I want to thank you for providing a place to stay and I'm really looking forward to seeing NY. I've never been to the USA or any other English-speaking countries so this is going to be new to me. (AC3)

(2) All in all I think that it will be quite interesting in New York and so there is nothing else to say than thank you for hosting me for 2 weeks. (AC13)

(3) I'm happy of being able to travel to NY and I'm looking forward to stay at your place. So at first I want to thank you for your permission to have you as my host family (AC21)

(4) I'm very happy about the stay in New York and that you have said I can life with you. (ANC18)

(5) So when I'm thinking of the weeks in New York I think of a great time! So I'm very honoured, that I can stay at your place (BNC11)

(6) I was never to the USA or to the UK so I'm really happy to visit you for two weeks. It must be really great to watch the skyline of New York at night. (BNC12)

(7) I have never been to NY before, so I'm looking forward to stay there. It was a dream of mine since I was a young boy to go to the USA to see all the sights and landscapes (ANC5).

(8) But, enough about me, let's talk about your city, New York, the "Big Apple". I'm very excited to visit Manhattan and its busy crowd and I'd love to see the Liberty Statue from close (BC 14).

(9) I hope we will have two wonderful weeks together. It would be very nice if you tell me something about your family (ANC5).

(10) Have you ever been to Austria? If not, I will tell you a bit about it, if you want to. Have you had many guests from other countries, or is this the first time you going to be a host-family? I really hope I will have a good time in N.Y. and that we all enjoy the time we will spend together (BC3)

While half of the examples given above are taken from the non-CLIL groups, this does not reflect the overall trend. If we look at the overall performance of the groups a similar picture emerges for both schools. While 57% of the AC group and about 50% of the BC group make a clear effort to establish a good relationship with their future hosts, the percentages for the non-CLIL groups are considerably lower at 21% and about 25% respectively. Even the positive conclusion, which is explicitly mentioned in the instructions and can be seen as the bare minimum in terms of positive rapport management, was missing in almost half of the ANC and about a third of the BNC groups (while it was included in practically all (95%) of the AC texts and 83% of the BC texts).

In general, it seemed that the CLIL students were much more likely to identify the communicative purpose of the task and to come up with appropriate strategies to fulfil this purpose whereas the non-CLIL students showed a tendency to simply go through the instructions point by point and to neglect the communicative goals to be achieved. Example (11) is a case in point.

(11) Hello my name is XXX.
I'm 14 years old and I live in Austria.
I have one mother and one father and two brothers. I'm attending the third year at the XXX. I want to finish school and want to gain wealth. My spare time activities are running, driving and eating. I'm happy about going to NY. Because this happns in school time. So I have not to learn for school. A question which I will ask is why so many people had choosed Bush. And I want to know if Amanda is a beautiful girl (ANC 16)

While the student has covered all the points, even including two questions, and has thus formally fulfilled the requirements, the text still seems inappropriate. While there is one positive statement ("I'm happy about going to NY"), this simply mirrors the instructions ("...why you are happy to go to NY") and any positive effect it might have is destroyed by the reason given ("because this happns in school time. So I have not to learn for school"). The questions chosen are also unlikely to help establish rapport and even the positive ending that is explicitly mentioned in the instructions is missing.

It should also be noted that in all groups about one third of the students included points (often in rather rude language) which must be regarded as totally inappropriate for the task given. Cases in point are associations of New York with danger (examples 12 &13), writers' self-characterisations as heavy drinkers of alcohol and other references to alcohol abuse (examples 14-16), inappropriate references to the teenage daughter of the host family (examples 16 & 17), references to the problems of obesity and insinuations about American eating habits (examples 18 & 19) and rude remarks about Americans in general or the US president (examples 20 & 21). Obviously, such passages would only serve to alienate the American hosts, thus jeopardising relationships from the start. As a result, some of the texts were graded down quite significantly, which, to some extent, might explain the wide range of scores on this dimension.

(12) Is life realy so dangerous in NY? (AC5)

(13) I hope I don't have to travel far, to come to the city and I hope I don't get ill from the bad air. (BNC14)

(14) In my spare time I often drink alcohol with my friends and I play football, basketball and Baseball (ANC19)

(15) I once went to Malta in the middle school and last year I went to England with my XXX class, where we drank a lot of alcohol (BC13)

(16) Two questions I would you ask. Do the two children often drink alcohol. And another furder question. Is your daughter very pretty and sexy? (ANC19)

(17) Question1: Is your daughter Amanda single? Because I love 15 years old girls (BC11)

(18) What I like to ask is: Do you cook on your own or do you just visit fast food restaurants ? (AC7)

(19) It rumours that in NY nearly every person is fat because of eating only food from McDonalds?! Is it really true? Because I don't want to eat only Fastfood and therefore I hope you are a good cooker?! (AC19)

(20) For me it will be the first time, that I'm visiting the USA. I'm really happy to go to New York, because I want to see how the Americans are and if they are as stupid as everybody says in my country. (BNC 9)

(21) One question I really want to ask you is: have you voted for president Bush? If yes, how is it possible that a country votes a man who is as stupid as a handy-caped person. A leader of a country is, like a symbol for the whole state (BNC 9)

To sum up the findings in the field of task fulfilment, it can be said that overall the CLIL students clearly outperformed their non-CLIL peers in terms of text format, length and register and, in particular, with regard to their being aware of and meeting the communicative task of building a relationship with the host family. There were differences between the CLIL groups as well, however, with the AC group seeming more prepared to go beyond the instructions and to bring in their own ideas in addition to what was explicitly mentioned.[1] As has been observed above, the BC students were also the

[1] For example, the BC students were the least likely to refer to their forthcoming stay in New York, which seems a typical opening strategy but was not explicitly mentioned in the instructions. Only 22% of this group referred to their stay, while this percentage is about twice as high in all other groups.

most likely not to include questions to the host family. The reasons for these differences cannot be explored in this thesis, however.[1]

4.3.2.2. Organisation

The second category refers to the field of organisation and structure, or the textual competence of test takers. Tables 15 and 16 summarise the results in this field:

	No of test persons	Range	Average scores	Standard Deviation (SD)	T value
Total CLIL	39	2-5	2.949	0.887	3.158**
Total non-CLIL	47	1-4	2.404	0.712	
	**Difference highly significant (level of significance 99%)				

Table 15 'Organisation', total sample

	No of test persons	Range	Average scores	Standard Deviation (SD)	T value
College A CLIL (AC)	21	2-5	3.143	0.854	2.339*
College A non-CLIL (ANC)	23	2-4	2.609	0.656	
College B CLIL (BC)	18	2-5	2.722	0.895	2.061
College B non-CLIL (BNC)	24	1-3	2.208	0.721	
	* Difference significant (level of significance 95%)				

Table 16 'Organisation', by school

As can be seen from Tables 15 and 16, the differences were less marked in this category. While they were highly significant for the total sample, only the College A groups showed significant differences when the data were analysed separately. What is perhaps even more striking is the generally low level of achievement in this field, with only the AC group reaching an average score above 3. The groups were also somewhat less heterogeneous on this dimension than on the first one.

The two schools show similar patterns in this category, the level of performance being generally lower in the College B groups. While there were deficiencies in overall structure in all groups involved in the study, the non-CLIL groups showed even greater shortcomings. A coherent structure throughout the text was found in only 25% of the AC

[1] It should perhaps also be added that with regard to the communicative purpose it was not necessarily important whether questions were included or not. While in some cases students managed to build rapport with the host family without including questions, in other cases no rapport-building was achieved even though questions were included as the questions were largely unsuitable.

texts and 11% of the BC texts, closely followed by the ANC texts at a rate of 9%. In the BNC group none of the texts could be characterised as coherent throughout. At the other end of the competence spectrum, 25% of the BNC texts can be described as lacking completely in structure and coherence. Transitions between the different parts of the text were often not clearly marked and abrupt. In particular, this concerned the questions addressed to the host family, with a fairly similar performance shown by the College B groups, where 75% of the BC students and 77% of the BNC students failed to integrate questions properly. The percentages of the College A students were 63% and 83% respectively, showing a much clearer difference between these groups.

Paragraphing was another weakness in all the groups participating in the study, with few consistent patterns emerging that would point to a systematic difference between the CLIL and non-CLIL students. Interestingly, the BC group was, at a rate of 33%, the most likely not to use this structuring device at all, while the percentage was around 20% for all the other groups. If we include those who used paragraphs sparingly or not effectively, the College B groups show a rather similar performance at rates of 83% (BC) and 88% (BNC), while the relevant percentages are 61% for the AC students and 48% for the ANC group. It is thus difficult to draw clear conclusions from the data in this respect.

A much clearer picture emerges with regard to connectives, where the CLIL groups clearly outperform their non-CLIL peers. While the majority of the AC students mostly used simple connectives (*and, but, because*), about half of them made a clear attempt to go beyond the most basic connectives and tried to use a wider range. 24% of the AC group, but 52% of the ANC group, used simple connectives only, with the ANC students partly limiting themselves to one or two connectives only. In particular the conjunction *and* was often overused. The situation is similar for the College B groups, but at a lower level. 44% of the BC students mostly used simple connectives, with 38% of these showing a clear tendency to use other linkers as well. 56% of the BC group and the overwhelming majority of the BNC group (96%) used simple connectives only, with the conjunction *and* being clearly overused in about half these cases.

To illustrate the differences, examples are provided below of (a) a text written by a student who makes an effort to go beyond the most basic connectives (example 22), (b) two texts where the use of connectives is limited to the most basic ones (examples 23 & 24) and (c) a text which hardly uses any connectives at all (example 25):

(22) Dear host family Fergusen,
My name is XXX **and** I am supposed to stay with your family during our two week stay in New York.
Let me first introduce myself. I am 17 years old **and** I live with my parents and two sisters in a house in XXX. My favourite hobbies are biking and having a good time with friends. **Further on** I am very interested in all kinds of stuff that have something to do with cars.
I am happily looking forward to my stay in New York **because** I have never been abroad before. The only information that I have got of American cities was transferred by movies. **And therefore** I am looking happily forward to our stay in New York.
To come to an end, I hope that all expectations will be fullfilled within this two weeks.
Yours scincerly
XXX (AC9)

(23) Dear Host family.
My name is XXX **and** I'm 17 years old. I have one Brother which is 12 years old.
My Dad is 45 years old and a Manager **and** my mum is 47 and teacher.
I attend to an technical college in XXX **and** when I finished that I want to go to the universaty of Vienna.
In my spare time I like to play football with my friends **and** I also like to spend time with my girl friend.
But I really don't like to learn for school and do my homework.
Also I have to say that I really love big citys like New York **and so** I'm happy to be there soon. To see the „Big Apple"! I hope you like sport like me **and** Amanda is a nice girl?
I'm looking forward to see you and the nice city
I will stay!
Best wishes,
XXX (BC10)

(24) Hello!
My name is XXX. I am 17 years old **and** live in the beautiful country Austria.
I am **also** excited to see you and our family. At this time I atend the XXX in XXX. I am in the third class **and** want to improve my english. In the future I will be a konstructor, **and** I learn hard for it.
My Hoppies are skiing and swimming. I **also** spend a much of time with my friends.
I hope that I will find friends in New York. That would very nice.
In New York I will see the life in the big city and (ANC3)

(25) Hy
I'm XXX, I'm 18 years old. I'm from Austria **and** I live in XXX. My Familie and I live in a nice village. The name is XXX. I go to school at the HTL XXX. At my free time I place ice-hockey and football in a team.
I'm happy to go to NY. I will see the NY-Rangers **and** I will go shopping in NY.
Have your son a boyfriend? (BNC5)

Overall, it can be said that the level of performance in the category of organisation and structure was generally the lowest for all the groups involved in the study. With regard to the overall structure and the use of connectives the performance of the CLIL students was comparatively stronger, while no clear picture emerged with regard to the use of paragraphs.

4.3.2.3. Grammar

Having dealt with aspects of pragmatic and textual competence we now turn to one of the main aspects of language skills, i.e. grammar. Tables 17 and 18 summarise the results in this field:

	No of test persons	Range	Average scores	Standard Deviation (SD)	T value
Total CLIL	39	2-5	3.564	0.882	5.353**
Total non-CLIL	47	1-5	2.511	0.930	
	**Difference highly significant (level of significance 99%)				

Table 17 'Grammar', total sample

	No of test persons	Range	Average scores	Standard Deviation (SD)	T value
College A CLIL (AC)	21	2-5	3.810	0.873	3.961**
College A non-CLIL (ANC)	23	2-5	2.739	0.915	
College B CLIL (BC)	18	2-4	3.278	0.826	3.618**
College B non-CLIL (BNC)	24	1-4	2.292	0.908	
	**Difference highly significant (level of significance 99%)				

Table 18 'Grammar', by school

As can be seen from Tables 17 and 18, in the field of grammar the differences in performance between the CLIL and non-CLIL groups were highly significant throughout, showing marked differences with regard to accuracy in particular. Moreover, while both groups seemed quite willing to use a variety of structures, the CLIL students tended to use them more accurately. Table 18 also suggests that on this dimension the College B groups are slightly more homogeneous, with no test taker reaching the top score. Figure 8 illustrates the scores obtained by the different groups:

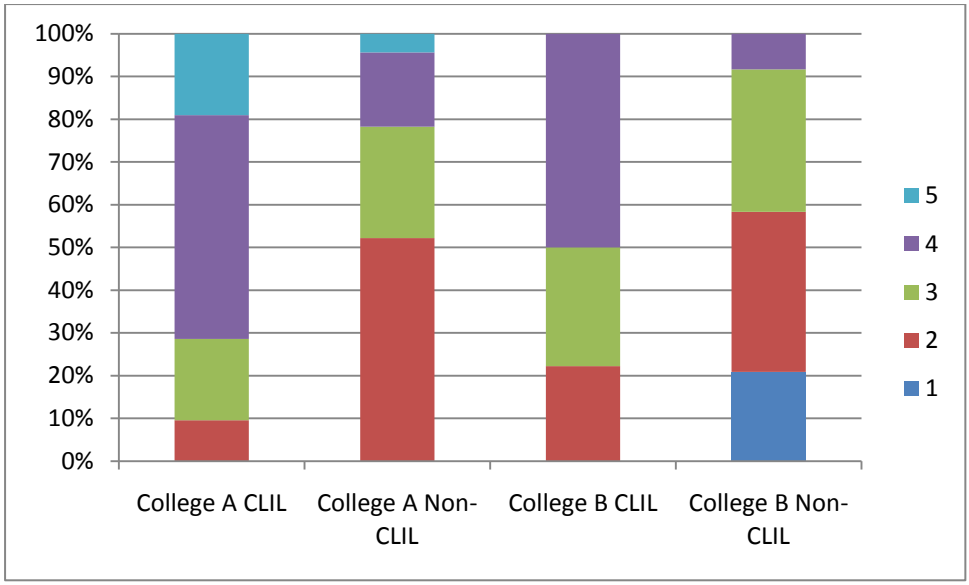

Figure 8 Distribution of scores, grammar

Overall, about 71% of the AC group and 50% of the BC group reached the highest or second highest category, thus using grammatical forms and structures accurately or mostly accurately and using a great or good variety of structures, including complex ones. The difference between these groups is that while none of the BC students obtained the highest score, 19% of the AC students did. In the non-CLIL groups, on the other hand, the majority (52% of the ANC group and 59% of the BNC group) did not go beyond the second lowest score, their texts being highly inaccurate with frequent morphosyntactic errors. Again, it should be added that none of the ANC group fell into the lowest category, while 21% of the BNC group did, using grammar and structures poorly and showing poor variety of structures. Thus there seems to be a clear advantage for the CLIL students in general, with the College B groups generally achieving a lower level than their College A counterparts.[1]

As for the types of mistakes made, one of the most frequent problems concerned the use of the present tense (present simple vs present progressive), with students typically overusing the progressive aspect:

[1] If only accuracy is considered, the overall differences between the CLIL and non-CLIL students show a similar pattern on a slightly lower level. Interestingly, however, in this case the CLIL groups seem more homogeneous in the top range, as 11% of the BC group achieved the top score and the percentages of those achieving the highest or second highest score are more similar, with 57% for the AC and 50% for the BC group. On the other hand, a rather large percentage, i.e.39% of the BC students only achieved a score below 3, compared to 9.5% of the AC group. Thus differences seem larger among the low achievers. These issues can, however, not be explored further in the context of this thesis. As a similar pattern emerged for the non-CLIL students, it does not seem to be linked to differences between the two CLIL programmes.

(26) **I'm living** in family with 5 members**;** mum, dad, one little brother and one little sister (AC15)

(27) My name is XXX and **I'm living** near XXX, which is the second biggest city in Austria (BC1)

(28) **I'm visiting** the HTL-XXX in XXX, It's a school for electronix and managment (BNC9)

(29) My name is Philipp and I am 17 years old. I **am attending** the HTL in XXX and I'm in the third class [...] I **am living** in Garsten, Austria. When **I'm finishing** the HTL, I want to go to work and gain wealth. In my spare time **I am playing** soccer the most time. (ANC8).

19% of the AC group and 33% of the BC group had problems choosing the right form, while the corresponding figures in the non-CLIL groups are 61% for the ANC students and 46% for the BNC group. Other mistakes that occurred in all groups were the use of the present simple instead of a future form, especially following *hope* (examples 30-32) and problems in using the present perfect correctly (examples 33 & 34). The non-CLIL groups tended to overuse the present simple, substituting it for the present perfect and past tenses (examples 35 &36), while some BNC students also substituted the past tense for the present (example 37). In this group, some students showed a rather erratic use of tenses, while others limited themselves to using present tense and future simple only.

(30) My English is not the best, therefore I hope that this two weeks **help** me to get better in English. (AC19)

(31) I hope **I find** a company, where I can use this. (BC4)

(32) And I hope Paul **plays** with me football.(BC17)

(33) Since 2005 **I visit** the HTL XXX. (ANC 15)

(34) I also **have** a boyfriend since two years. (BNC 16)

(35) For years ago I was in London and it was very good and **I enjoy** it very much. (BNC7)

(36) **I join** the XXX for nearly three years (BNC12).

(37) My family contains a father, Mother and a good friend, **I had** no sisters or brothers. **I had now** no future career plans (BNC6)

Problems with word order, often involving L1 interference, were rather infrequent in the AC group (4.8%), but relatively prominent in the BC group (39%), the corresponding figures for the non-CLIL groups being 26% and 29% respectively (examples 38-41). The need for do-support in questions seemed to pose a particular problem (examples 42-44). Finally, the percentage of students using incorrect verb forms (examples 45-50) was considerably higher in the BNC group (25%), while it was below 10% in all other groups.

(38) In our family it's normal to **play seven days in the week soccer**.(ANC9)

(39) Two questions I would **you ask**.(ANC19)

(40) And I hope Paul **plays with me football**.(BC17)

(41) We stayed seven days at a host family in Bexhill and afterwards **we went by bus to London** and checked in a awful hotel. (BC18)

(42) How look your doughter? (BNC4)

(43) Have your son a boyfriend? (BNC5)

(44) What room I get to live then? (BNC6)

(45) Hello, my name is XXX and I will spend the holidays at your home **too learning** English (BNC2)

(46) **Can we looking** some sightseeings. (BNC2)

(47) **I want to knew** how are the Americans (BNC7)

(48) Our class **will made** a two-week stay in New York and my school selected You as my host family for this two weeks.(BNC13)

(49) A question which I will ask is why so much people **had choosed** Bush.(ANC16)

(50) Before the HTL-XXX **I've visit** the secondary modern school in XXX (AC17)

In the field of grammar, the qualitative assessment carried out on the basis of rating scales was supplemented with quantitative measures. The following ratios were calculated:

- Errors per 100 words
- Number of different verb forms.
- Sentence length in words as well as number of subordinate clauses per 100 words as measures of syntactic complexity.

In view of the fact that error frequency was supposed to measure grammatical accuracy, only morphosyntacic errors were counted. As for the number of different verb forms, all tenses, passive forms and modal verbs were taken into account, in accordance with Yuan and Ellis (2003:13). In addition, infinitives, gerunds and participles were included. Finally, when establishing the number of subordinate clauses, basically only finite clauses were counted. However, they were supplemented with non-finite constructions serving to shorten subordinate clauses. In view of the fact that these seem to indicate a higher stage of development than finite clauses (Wolfe-Quintero et al 1998:73) they were included as otherwise students using these structures would have been assigned a lower ratio, which was hardly desirable. Table 19 summarises the results:

	Number of cases	Errors per 100 words	Verb forms	Sentence length (in words)	Subordinate clauses per 100 words
Total CLIL	39	3.175**	6.949**	15.518**	3.270
Total non-CLIL	47	4.977**	5.830**	12.630**	2.751

** highly significant (level of significance 99%)

Table 19 Accuracy and complexity ratios, total sample

As shown by Table 20, the analysis of the total sample shows a highly significant difference in all ratios except for the number of subordinate clauses. If we consider the two schools separately a more differentiated picture emerges:

	Number of cases	Errors per 100 words	Verb forms	Sentence length (in words)	Subordinate clauses per 100 words
College A CLIL (AC)	21	2.742**	7.000	16.319**	3.560
College A non-CLIL (ANC)	23	4.394**	6.217	13.326**	2.712
College B CLIL (BC)	18	3.681*	6.889**	14.583*	2.933
College B non-CLIL (BNC)	24	5.535*	5.458**	11.963*	2.789

** highly significant (level of significance 99%) *significant (level of significance 95%)

Table 20 Accuracy and complexity ratios by school attended

As Table 20 shows, the greatest differences can be observed in the number of errors committed as well as average sentence length, with the CLIL students outperforming their

non-CLIL peers at both colleges and the difference being statistically significant for the College B and highly significant for the College A students. With regard to the range of verb forms used, only the College B groups showed significant differences. No significant differences could be identified with regard to the number of subordinate clauses.

To sum up, we can conclude that on the grammar dimension the groups analysed showed considerable differences with regard to the accurate use of grammatical forms and structures, with the CLIL students clearly outperforming their EFL-only peers. While the non-CLIL groups were prepared to use more complex structures such as subordinate clauses to a similar extent, they tended to have greater difficulties using them correctly.

4.3.2.4. Vocabulary

The last category to be analysed, dealing with vocabulary and expression, also concerns language competence, Table 21 summarises the overall result in this field:

	No of test persons	Range	Average scores	Standard Deviation (SD)	T value
Total CLIL	39	2-5	3.718	0.857	4.561**
Total non-CLIL	47	1-5	2.936	0.704	
	**Difference highly significant (level of significance 99%)				

Table 21 'Vocabulary', total sample

As was the case with the other three categories, the differences identified between the CLIL and non-CLIL students are also highly significant with regard to lexical competence. Again, however, the results suggest a high degree of variability. If we consider the different subgroups, the following picture emerges:

	No of test persons	Range	Average scores	Standard Deviation (SD)	T value
College A CLIL (AC)	21	3-5	3.952	0.740	3.801**
College A non-CLIL (ANC)	23	2-5	3.130	0.694	
College B CLIL (BC)	18	2-5	3.444	0.922	2.820**
College B non-CLIL (BNC)	24	1-4	2.750	0.676	
	* Difference significant (level of significance 95%) **Difference highly significant (level of significance 99%)				

Table 22 'Vocabulary', by school attended

Figure 9 illustrates the distribution of the scores in these groups:

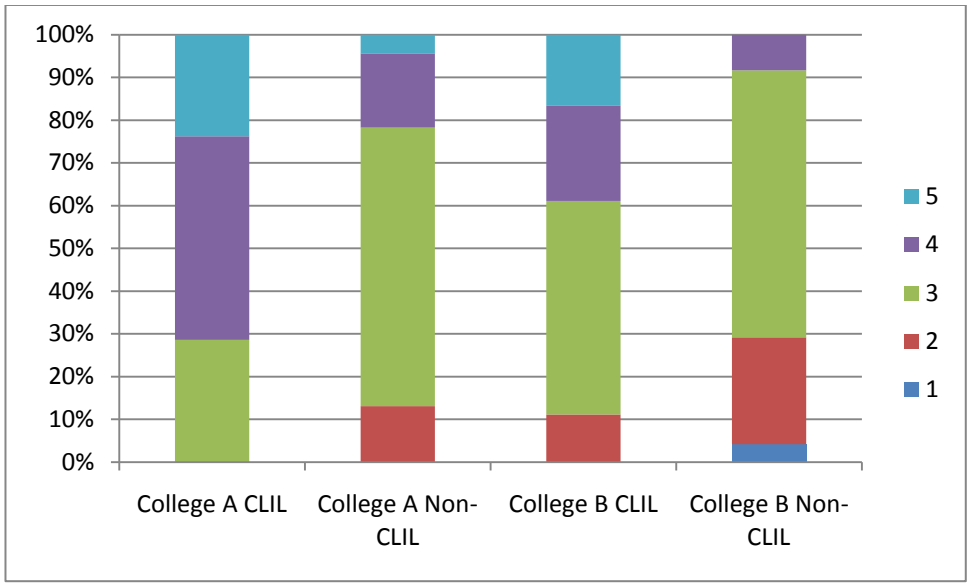

Figure 9 Distribution of scores, vocabulary

As Table 22 shows, the differences between CLIL and non-CLIL students are highly significant for both groups. As was the case with the other three categories, the AC group again showed the best performance. What is remarkable is that the range of vocabulary and the accuracy of its use were rated at least 'adequate' for all members of this group (as shown by the range of scores). Almost half the AC students achieved the second highest score and thus their texts exhibited a good range of vocabulary, which was mostly used accurately, and spelling mistakes were few. Almost a quarter even reached the highest score, and the lexical competence of the rest was judged at least adequate, as mentioned above. There were few problems due to L1 transfer in this group, the most frequent one by far being the expression *to visit a school* substituted by 43% for *attend a school*. Moreover, these students rarely resorted to simple expressions such as *other things* or did not use them at all.[1] Communicative problems due to wrong use of lexis were unlikely to occur. Compared to this group, the BC group showed a generally lower, but also more varied performance, ranging from the highest score to the second lowest. While half the group achieved an adequate performance (i.e. a score of 3), 11% of the text were judged less than adequate, showing deficiencies such as a limited range of vocabulary, frequent

[1] If only lexical range and choice of words are considered, the scores of all the groups are lower than for the overall scores on lexical performance. Nevertheless even there more than half the AC students (57%) reach one of the top two scores, with the remaining 43% obtaining a score of 3. In this case there are clearer differences between the AC and BC groups as only one third of the latter reach a score of 4 or 5.

repetitions and frequent spelling mistakes. The rest, however, had a good or even wide range of vocabulary and few orthographic problems, and thus achieved the highest or second highest score. With about 17% in the top range, again the differences between the two CLIL groups are not so great at the top, but there are more students in the middle range than in the AC group, where the majority achieved a score above 3.

With regard to the non-CLIL groups, we can observe that they both performed on a lower level than their CLIL peers. In both cases about two thirds of the test persons achieved an adequate performance (i.e. a score of 3), but while about a fifth of the ANC group ranked higher and only about 13% fell below a score of 3 ('adequate'), the situation is more or less reversed in the BNC group. Only 8% of the texts produced by this group can be described as exhibiting a good range of vocabulary and few mistakes. On the other hand, with 29% quite a sizable percentage of the BNC group fail to achieve an adequate performance, with one person (4%) even falling into the lowest category as he or she showed such a limited range of vocabulary that, for example, even the word *Familie* ('family') was borrowed directly from the L1.[1] On the whole, the vocabulary used by this group was fairly basic and major orthographic deficiencies were clearly more frequent than in the CLIL group (and in fact, any other group).

As for the types of problems that occurred in the lexical field, they affected the range and choice of words as well as orthography. In general, compared to the CLIL groups, the non-CLIL students were more likely to (over)use simple words such as the adjective *big* (examples 51-53) and had more orthographic problems, which in some cases could lead to problems in understanding the intended meaning or at least put a considerable burden on the recipient trying to make out the writer's meaning (examples 54-57). In some cases spelling problems even concerned basic words such as *apple* or *ask* (see examples 58 & 59). Moreover, there were more cases of L1 interference (examples 60-62) and in some cases students simply borrowed words from their L1 (see examples 63-65). A few examples are listed below:

(51) I'm happy to go to NY, because **I want to see the big buildings there** and the Central Park and other sighseeings. (BNC14)

[1] In fact, if only lexical range and choice of words are considered almost half of the BNC group (46%) do not reach an adequate performance (i.e. a score of 3). With roughly one quarter this percentage is considerably lower in the ANC group.

(52) Do you life **in a big house with many families** or do you have your own house?(BNC17)

(53) I am looking forward to see **the big towers** and the yellow cabs.(BNC2)

(54) My **Hoppies** are skiing and swimming (ANC3)

(55) I'm **ettendig** the 3 class of the HTL-XXX (ANC14)

(56) My reason for writing this letter is that I have some **quescens abaut** staying in your house. **Cane** you tell my something about your flat? Or can you send me some **foto's** about ? (BNC18)

(57) Is there hot **wether**? Isit very **expenciev** (?) or is it **cheep**? (BNC7)

(58) I always wanted to see the big **appel** all the skyscrapers and the statue of liberty (ANC20)

(59) At the and I have 2 questions to **aske** (ANC20)

(60) [...] my future plans are to finish this school and **to go working** (ANC 10)

(61) I'd like to **become a job** in the racing szene, like the Formular 1 or DTM. (ANC13)

(62) At the moment I'm **visiting** the HTL-XXX and it is very **funny** there.(ANC22)

(63) When I finish the HTL I will study **Maschinenbau** in XXX. (ANC11)

(64) When I pass the **matura** I want to have an own company (ANC22)

(65) My **Familie** and I live in a nice village. (BNC5)

While to some extent such problems also occurred in the texts of the BC students, they were virtually non-existent in the AC group. Overall, this confirms the superior performance of the CLIL students in the field of lexical performance.

As in the field of grammar, a number of quantitative ratios were calculated in addition to the qualitative assessment based on subjective rating:

- Percentage of the 1000 and 2000 most frequent words (K1 and K1+K2)
- Type-token ratio (per 50-word segment, TTR)
- Average word length (characters per word)

Using the type-token ratio to assess range of vocabulary poses the problem that this ratio is not independent of text length, but inversely related to it. As the texts analysed in this

study varied in length, they were split into 50-word segments and the average TTR per text was used (cf. Ellis and Barkhuizen 2005:155).

Tables 23 and 24 show the values of these ratios for the total sample and for the two schools analysed separately:

	Number of cases	K1 in %	K1+K2 in %	Type-Token Ratio	Average word length
Total CLIL	39	91.686*	96.862	0.777**	3.923**
Total non-CLIL	47	92.685*	97.145	0.747**	3.809**

** highly significant (level of significance 99%) *significant (level of significance 95%)

Table 23: Lexical ratios, total sample

	Number of cases	K1 in %	K1+K2 in %	Type-Token Ratio	Average word length
College A CLIL (AC)	21	92.530	97.082	0.792**	3.938
College A non-CLIL (ANC)	23	92.486	97.302	0.753**	3.891
College B CLIL (BC)	18	90.702**	96.605	0.759	3.906*
College B non-CLIL (BNC)	24	92.875**	96.994	0.742	3.730*

** highly significant (level of significance 99%) *significant (level of significance 95%)

Table 24: Lexical ratios, by school attended

As can be seen from Tables 23 and 24, there are significant or highly significant differences in the percentage of K1, the type-token ratio and word length if the two groups are considered together. Analysed by school attended, there are still significant differences but for different ratios. At College A the CLIL group used less lexical repetition, as indicated by a higher type-token ratio. At College B, on the other hand, the CLIL group used the most frequent 1000 words less often than the non-CLIL group and tended to use longer words.

5. Summary and conclusion

The purpose of the present study was to assess and compare the general language ability as well as the writing skills of CLIL students and non-CLIL students at two Austrian upper-secondary engineering colleges. For this purpose different types of tests as well as the most important properties of tests were investigated and on this basis a test was developed which aimed to provide two different measures of general language ability. For the first part, which was to assess test takers' general language competence in an objective manner, a C-test, i.e. an integrative test based on the cloze principle, was chosen. The second part consisted in a writing task, which was designed, on the one hand, to provide a subjectively measured assessment of general language ability and, more specifically, to assess the writing skills of the students. To meet the requirements of communicative language testing, students were put into a specific communicative situation which was assumed to be familiar to them and were asked to write a letter or e-mail to a prospective host family in New York. The basic research question was to find out whether the hypothesis that CLIL students outperform their EFL-only peers with regard to both general language ability and writing skills could be supported on the basis of the data collected.

Generally speaking, the results support this hypothesis, with the scores obtained on both the C-test and the writing part showing clear, statistically significant advantages for the CLIL students. When analysing different aspects of the students' writing ability, however, we see that the extent to which CLIL students outperform their non-CLIL peers depends on which aspect of writing ability we focus on. The results suggest that those areas which involve purely linguistic skills (i.e. grammar and vocabulary) seem to benefit the most from CLIL instruction. In the field of grammar, the difference was found to be highly significant, in particular with regard to the accuracy of language use. Moreover, the CLIL group was found to have a wider range of vocabulary at their disposal and better orthographic skills. The two other categories analysed, i.e. task fulfilment and organisation, however, demand pragmatic and textual skills. As for task fulfilment, the CLIL group clearly outperformed their non-CLIL peers with regard to awareness and fulfilment of the communicative purpose of the text to be written, thus showing greater pragmatic awareness. In the field of organisation and structure, however, the differences

were smaller and it must be noted that, on the whole, these skills were not very well developed. At one college no significant difference could be identified in this respect.

The results thus seem to suggest that the most significant advantages of CLIL students in terms of their writing skills result from a more advanced general language ability as well as a greater awareness of the pragmatic requirements of the task. The effects on textual competence, on the other hand, seem limited. This is hardly surprising in view of the fact that, apart from some note-taking and completion of worksheets, very little writing is typically demanded of Austrian CLIL students. While this thesis did not focus specifically on the development of subject-specific competence, one should not disregard the fact that writing may also help students to come to terms with, and ultimately appropriate, the curricular concepts discussed (cf Lemke 1990: 168). Therefore an opportunity to promote both an understanding of the content and the development of language skills may be lost if writing is neglected in CLIL classrooms, as seems to be the case at the moment.

With regard to general language ability, a caveat should be added. As there were no data available on the students' level of language competence at the start of the CLIL programme, it cannot be ruled out that the performance of the CLIL and non-CLIL groups was already significantly different at that time. It should also be mentioned that there were statistically significant differences not only between the CLIL and non-CLIL groups but also between the two participating schools. On the one hand, as has been pointed out, the CLIL programmes implemented at the two colleges differ with regard to organisational aspects and amount of exposure and it seems that quantity of exposure is a key factor determining the effectiveness of a CLIL programme. On the other hand, the fact that differences were also observed between the non-CLIL groups suggests that there may be other significant differences between the schools, the analysis of which would, however, go beyond the scope of this thesis.

Due to these considerations, as well as the very limited scope of the case study, it seems difficult to draw general conclusions about CLIL provisions on the basis of the findings presented here. The results do suggest a positive influence of CLIL provision on the language output of the students and, given the traditionally low profile of foreign language teaching at Austrian engineering colleges, the use of CLIL in this environment seems to be helpful if the level of language proficiency is to be improved. However, to

obtain greater insights into the effects of CLIL a more broadly based longitudinal study would probably be required. Moreover, as the CLIL umbrella covers such a large variety of approaches, it would be interesting to investigate the effects of different types of programmes on the development of language competence in order to identify those factors which are most beneficial.

It should also be mentioned that the present case study formed part of a larger study commissioned by the Austrian Ministry of Education. A variety of approaches were taken in order to assess the present state of CLIL provision at Austrian engineering colleges and to make recommendations concerning the future development and implementation of CLIL. The final report identifies three main pillars determining the success of CLIL programmes, i.e. commitment, structure and support, and concludes that while there is plenty of commitment on the part of the teachers involved, there are, at the same time, considerable shortcomings with regard to organisational structure and the support teachers receive. (Dalton-Puffer et al 2008: 10). Recommendations include certain minimum standards that should form the basis of all CLIL provision in this context. For example, at least 25% of the lessons should, on average, be conducted in English in order to ensure a certain minimum of exposure to the foreign language and to achieve measurable differences in language output. Furthermore, CLIL should be implemented in a systematic and structured way, involving a mix of different subjects introduced step by step, and support ought to be provided in the form of team teaching and additional English classes.[1] Other suggestions concern a minimum level of language proficiency for all CLIL teachers corresponding, at least, to level B2 of the Common European Framework of Reference and the provision of a common framework for CLIL programmes to be adapted by individual schools to meet local requirements. Moreover, it is recommended that schools should lay down the principles of their CLIL programmes and communicate them actively to stakeholders (Dalton-Puffer et al 2008: 11f). All in all, the authors conclude that it is time for CLIL to move from the experimental stage to being fully integrated into the educational system, albeit on an optional basis (cf. Dalton-Puffer et al 2009: 25).

[1] With regard to the findings of the case study presented in this paper, we might say that ideally a CLIL programme would combine the advantages of the programmes implemented at the two colleges participating in the study, i.e. substantial exposure to the foreign language and a structured and systematic approach to implementing the CLIL principle.

Apart from such framework conditions, changes might also be necessary with regard to how CLIL classes are taught. As mentioned in Chapter 3 of this thesis, it seems desirable to put more emphasis on certain elements that, at the moment, seem underrepresented in CLIL classrooms. In particular, this concerns writing and the development of academic language functions. While there is very little writing done in CLIL classes (or, in fact, any non-language class) in Austria, both language competence and curricular competence could benefit if students were required to engage in written tasks on a regular basis. With regard to academic language functions, CLIL seems to be particular well-suited to equip students with the tools for further studying and learning, a fact which so far seems to have been neglected in favour of an almost exclusive focus on developing fluency and reducing inhibitions. Undoubtedly, these are important benefits of CLIL, and the achievements made so far are not to be belittled. Nevertheless one might argue that the full potential of CLIL has so far not been exploited. More research will certainly be helpful in this field.

6. Bibliography

Ackerl, Christina. 2007. "Lexico-grammar in the essays of CLIL and non-CLIL students: error analysis of written production" *VIEWS*, Vol. 16 (3), 6-11.

Alderson, J.Charles. 1978. *A study of the cloze procedure with native and non-native speakers of Englis*h. PhD thesis, University of Edinburgh.

Alderson, J. Charles. 1991. "Bands and scores". In Alderson, J. Charles; North; Brian (eds.). *Language testing in the 1990s: the communicative legacy*. London: Macmillan, 71-86.

Bachman, Lyle F. 1990. *Fundamental considerations in language testing*. Oxford: OUP.

Bachman, Lyle F. 2000. "Modern language testing at the turn of the century: assuring that what we count counts". *Language Testing* 17 (1), 1-42.

Bachman, Lyle F. 2004. *Statistical Analyses for Language Assessment*. Cambridge: CUP.

Bachman, Lyle F.; Palmer Adrian S. 1996. *Language Testing in Practice*. Oxford: OUP.

British Council. Teaching English: *CLIL (Content and Language Integrated Learning) – Introduction*, www.teachingenglish.org.uk/transform/teachers/specialist-areas/clil, (accessed on 15 August 2009).

Canale, Michael; Swain, Merrill. 1980. "Theoretical bases of communicative approaches to second language teaching and testing". *Applied Linguistics*, Vol.1 (1), 1-47.

Carroll, Brendan J 1980. *Testing communicative performance*. Oxford: Pergamon Press

CLILCOM, http://clilcom.stadia.fi (accessed on 16 August 2009).

CLIL Compendium. www.clilcompendium.com.

Cummins, Jim. 1991. "Conversational and academic language proficiency in bilingual contexts". *AILA Review 8*, 75-89.

Ellis, Rod; Barkhuizen, Gary. 2005. *Analysing Learner Language*. Oxford: OUP.

European Commission. 2005. *A new framework strategy for multilingualism*, http://ec.europa.eu/education/languages/archive/doc/com596_en.pdf (accessed on 15 August 2009).

Dalton-Puffer, Christiane. 2007 *Discourse in content and language integrated (CLIL) classrooms.* Amsterdam: John Benjamins Publishing Company.

Dalton-Puffer, Christiane. 2008a. "Communicative competence in ELT and CLIL classrooms: same or different?" *VIEWS*, Volume 17 (3), 14-21.

Dalton-Puffer, Christiane. 2008b. "Outcomes and Processes in Content and Language Integrated Learning (CLIL): Current Research from Europe". In Delanoy, Werner; Volkmann, Laurenz. *Future Perspectives for English Language Teaching*. Heidelberg: Universitätsverlag Winter, 139-157.

Dalton-Puffer, Christiane; Smit, Ute. 2007. "Introduction". In Dalton-Puffer, Christiane; Smit, Ute (eds.). *Empirical perspectives on CLIL classroom discourse.* Frankfurt: Peter Lang, 7-23.

Dalton-Puffer, Christiane; Hüttner, Julia; Jexenflicker, Silvia; Schindelegger, Veronika; Smit Ute. 2008.*CLIL an österreichischen HTLs: Kurzfassung des Forschungsprojekts- Executive Summary*. Wien: Bundesministerium für Unterricht, Kunst und Kultur.

Dalton-Puffer, Christiane; Hüttner Julia; Schindelegger, Veronika; Smit, Ute. 2009. "Technology-geeks speak out: what students think about vocational CLIL", *International CLIL Research Journal* Vol 1 (2), 18-25.

Davies, Alan. 1997. "The construction of language tests". In Allen, J.P.B.; Davies, Allen (eds). *The Edinburgh Course in Applied Linguistics. Volume Four. Testing and experimental methods.* Oxford: OUP. 38-104.

Davies, Alan; Brown, Annie; Elder, Cathie; Hill, Kathryn; Lumley, Tom; McNamara, Tim. 1999. *Dictionary of language testing*. Cambridge: CUP.

Fulcher, Glenn; Davidson, Fred. 2007. *Language Testing and Assessment*. London: Routledge.

Friedl, Gabriele; Auer, Margit. 2007. *Erläuterungen zur Novellierung der Reifeprüfungsverordnung für AHS, lebende Fremdsprachen.* Wien-St.Pölten.

Grotjahn, Rüdiger 1987. "How to construct and evaluate a C-test: a discussion of some problems and some statistical analyses". In Grotjahn et al. *Taking their measure: the validity and validation of language tests.* Bochum: Brockmeyer, 219-253.

Grotjahn, Rüdiger (ed.). 2002. *Der C-Test – theoretische Grundlagen und praktische Anwendungen*, Band 4. Bochum: AKS-Verlag.

Grotjahn, Rüdiger; Klein-Braley, Christine; Raatz, Ulrich. 2002. "C-Tests: an overview". In Coleman et al. *University language testing and the C-test*. Bocum: AKS-Verlag, 93-114.

Halliday, Michael A.K.; Hasan, Ruqaiya. 1976. *Cohesion in English*. London: Longman.

Hamp-Lyons, Liz. 1991. " Scoring procedures for ESL contexts". In Hamp-Lyons, Liz (ed.). *Assessing second language writing in academic contexts*. Norwood, NJ: Ablex, 241-276.

Hastings, Ashley J 2002. "Error analysis of an English C-test: evidence for integrated processing". In Grotjahn, Rüdiger (ed.). *Der C-Test – theoretische Grundlagen und praktische Anwendungen*, Band 4. Bochum: AKS-Verlag. 53-66.

Hedge, Tricia. 2000. *Teaching and learning in the language classroom*. Oxford: OUP.

High Level Group on Multilingualism. 2007. *Final report*. Luxembourg: Office for Official Publications of the European Communities. http://ec.europa.eu/education/languages/pdf/doc1664_en.pdf (accessed on 15 August 2009)

Hughes, Arthur. 2003. *Testing for language teachers*. (2nd edition). Cambridge: CUP.

Hymes, D.H. 1972. "On communicative competence". In Pride, J.B; Holmes, Janet. *Sociolinguistics*. Harmondsworth: Penguin, 269-293.

Järvinen, Heini-Marja. Forthcoming. "Language as a meaning making resource in learning and teaching content: Analysing writing in content and language integrated learning In Dalton-Puffer, Christiane; Nikula, Tarja; Smit; Ute. *Language use in Content-and-language integrated learning (CLIL).* Amsterdam: John Benjamins Publishing Company.

Klein-Braley, Christine. 1985a. "Advance prediction of test difficulty". *Fremdsprachen und Hochschule* 13/14, 23-41.

Klein-Braley, Christine. 1985b. "Reduced redundancy as an approach to language testing". *Fremdsprachen und Hochschule* 13/14, 1-13.

Krashen, Stephen D. 1982. *Principles and practice in second language acquisition.* Oxford: Pergamon Press.

Lasagabaster, David. 2008. "Foreign language competence in content and language integrated courses". *The Open Applied Linguistics Journal* 1, 31-42.

Lemke, Jay L. 1990. *Talking science – language, learning, and values.* Norwood NJ: Ablex.

Lextutor (www.lextutor.ca/vp/eng).

Lightbown, Patsy M; Spada, Nina. 2006. *How languages are learned.* (3rd edition). Oxford: OUP.

Long, Michael H. 1996. "The role of the linguistic environment in second language acquisition." In Ritchie, William C. and Bhatia, Tej K. (eds.) *Handbook of Second Language Acquisition.* San Diego: Academic Press, 413-468.

Maljers, Anne; Marsh, David; Wolff, Dieter (eds). 2007. *Windows on CLIL – Content and language integrated learning in the European spotlight.* The Hague: European Platform for Dutch Education

Mariotti, Cristina. "Negotiated interactions and repair patterns in CLIL settings". *VIEWS* 15 (3), 33-39.

Marsh, David (ed.). 2007. *The CLIL quality matrix*. Graz: European Centre for Modern Languages.

McCarthy, Michael. 1991. *Discourse analysis for language teachers*. Cambridge: CUP.

McNamara, Tim. 1996. *Measuring second language performance*. London: Longman.

McNamara, Tim. 2000. *Language testing*. Oxford: OUP.

Mehisto, Peeter; Marsh, David; Frigols, Maria Jesús. 2008. *Uncovering CLIL – Content and language integrated learning in bilingual and multilingual education*. Oxford: Macmillan.

Messick, S.A. 1989. "Validity". In Linn, R.L. (ed.). *Educational Measurement*. (3rd edition). New York: American Council on Education/ Macmillan, 13-103.

Mewald, Claudia. 2007. "A comparison of oral foreign language performance of learners in CLIL and mainstream classes at lower secondary level in Lower Austria". In Dalton-Puffer, Christiane; Smit, Ute (eds.). *Empirical perspectives on CLIL classroom discourse*. Frankfurt: Peter Lang, 139-177.

Miller, George Tyler. 1991. *Environmental Science: sustaining the earth*. 3rd edition, Belmont: Wadsworth.

"Mobile phone use backed on planes", *BBC News*, 18 October 2007 http://news.bbc.co.uk/1/hi/technology/7050576.stm (accessed on 21 October 2007)

Morrow, Keith. 1979. "Communicative language testing: revolution or evolution?". In Brumfit, CJ; Johnson, K. *The communicative approach to language teaching*. Oxford: OUP, 143-157.

Musumeci, Diane. 1996. "Teacher-learner negotiation in content-based instruction: communication at cross-purposes?" *Applied Linguistics 17(3)*, 286-325.

Oller, John W. 1979. *Language tests at school: a pragmatic approach*. London: Longman.

Pilliner, A.E.G. 1970. "Subjective and objective testing". In Davies, Alan (ed.). *Language testing symposium: a psycholinguistic approach.* London: OUP, 19-35.

"Quick guide: broadband", *BBC News*, 29 June 2006, http://news.bbc.co.uk/go/pr/fr/-/2/hi/technology/5098866.stm (accessed 10 January 2008)

"Quick guide: China's economic reform", *BBC News*, 3 November 2006, http://news.bbc.co.uk/go/pr/fr/-/2/hi/asia-pacific/5237748.stm (accessed 10 January 2008).

Raatz, Ulrich. 1985. "Tests of reduced redundancy – the C-Test, a practical example", *Fremdsprachen und Hochschule* 13/14, 14-19.

Raatz, Ulrich; Klein-Braley, Christine. 1985. "How to develop a C-test", *Fremdsprachen und Hochschule* 13/14, 20-22.

Raatz, Ulrich; Klein-Braley, Christine. 2002."Introduction to language testing and to C-tests", in Coleman et al. University language testing and the C-test. Bochum: AKS-Verlag. 75-91.

Readability formulas (www.readabilityformulas.com).

Reilly, Brian. 1997. *Create PowerPoint Presentations in a weekend*. Rocklin: Prima Publishing

Reithuber, Franz. 2006. "HTL Steyr goes global: Erste 'Englischklasse' der österreichischen HTL's", http://www.htl.at/fileadmin/content/downloads/erfahrungsberichte_-_fallbeispiele/Fallbeispiel10_Steyr.pdf (accessed 10 July 2008)

Ruiz de Zarobe, Yolanda. 2007. "CLIL in a bilingual community: similarities and differences with the learning of English as a Foreign Language, *VIEWS* Vol 16 (3), 47-52.

Ruiz de Zarobe, Yolanda. 2008. "CLIL and foreign language learning: a longitudinal study in the Basque country" *International CLIL Research Journal*, Vol 1 (1), 60-73.

Ruiz de Zarobe, Yolanda. Forthcoming. "Written production and CLIL: an empirical study". In Dalton-Puffer, Christiane; Nikula, Tarja; Smit; Ute. *Language use in Content-and-language integrated learning (CLIL).* Amsterdam: John Benjamins Publishing Company

Schindelegger Veronika. 2008. Transcripts of interviews conducted with teachers and students, unpublished.

Schneeberger, Arthur; Petanovitsch, Alexander, Nowak, Sabine; Gruber Angelika. 2008. *Mittelfristige Perspektiven der HTL: Erhebungen und Analysen zur Sicherung und Weiterentwicklung der Ausbildungsqualität.* ibw-Schriftenreihe Nr. 138. Wien: ibw.

Sigott, Günther 2002. "High-level processes in C-Test taking?" In Grotjahn, Rüdiger (ed.). *Der C-Test – theoretische Grundlagen und praktische Anwendungen*, Band 4. Bochum: AKS-Verlag, 67-82.

Sigott, Günther 2004. *Towards identifying the C-test construct*. Frankfurt am Main: Peter Lang.

Spolsky, Bernard. 1995. *Measured words*. Oxford: OUP.

Spolsky, Bernard. 1973. "What does it mean to know a language; or how do you get someone to perform his competence?" In Oller, John W. Jr.; Richards, Jack C. (eds.). *Focus on the learner*. Rowley, Mass.: Newbury House, 164-176.

Swain, Merrill. 1995. "Three functions of output in second language learning". In Cook, Guy and Barbara Seidlhofer. *Principle & Practice in Applied Linguistics*. Oxford: OUP. 125-144.

Tankó, Gyula. 2005. *Into Europe: The writing handbook*. Budapest: Teleki László Foundation.

"The war for talent", *Business Spotlight* July-August 2007, 9.

Thompson, Bill. 2002. "Why the poor need technology". *BBC News*, 6 October 2002, http://news.bbc.co.uk/2/hi/technology/2295447.stm (accessed on 24 November 2007)

Thompson, Sam; Abdallah, Saamah; Marks, Nic; Simms, Andrew; Johnson, Victoria. 2007. *The European (un)Happy planet index – an index of carbon efficiency and well-being in the EU*, nef (the new economics foundation) and Friends of the Earch, http://www.foe.co.uk/resource/reports/euro_happy_planet_index.pdf (accessed 10 January 2008)

Van de Craen, Piet; Ceulers, Evy; Lochtman, Katja; Allain, Laure; Mondt, Katrien. 2007. "An interdisciplinary research approach to CLIL learning in primary schools in Brussels". In Dalton-Puffer, Christiane; Smit, Ute (eds.). *Empirical perspectives on CLIL classroom discourse.* Frankfurt: Peter Lang, 253-274.

Vollmer, Helmut 2008 "Bilingualer Sachfachunterricht als Inhalts- und als Sprachlernen". In Bach, Gerhard; Niedermeier, Susanne (Hrsg.). *Bilingualer Unterricht: Grundlagen, Methoden, Praxis, Perspektiven*. (4. Auflage). Frankfurt am Main: Peter Lang, 47-70..

Weir, Cyril J. 1990. *Communicative language testing*. New York: Prentice Hall.

Widdowson, H.G. 1978. *Teaching language as communication*. Oxford: OUP.

Widdowson, H.G. 2007. *Discourse analysis*, Oxford: OUP.

Wilding, Günter; Plösch, Margit; Kupplent, Wolfgang. 2009. *PEA – das Projekt Englisch als Arbeitssprache* ÖSZ Themenreihe 3. Graz: Österreichisches Sprachen-Kompetenz-Zentrum.

Wilhelmer Nadia. 2008. Transcripts of interviews conducted with teachers and students , unpublished.

Wolfe-Quintero, Kate; Inagaki, Shunji; Kim, Hae-Young. 1998. *Second language development in writing: measures of fluency, accuracy & complexity.* Honolulu, Hawaii: Second Language Teaching and Curriculum Center, University of Hawaii at Manoa.

Wolff, Dieter. 2007. "Approaching CLIL", in Marsh, David (ed.). *The CLIL quality matrix*. Graz: European Centre for Modern Languages, 10-25.

Yuan, Fangyuan; Ellis, Rod. 2003. "The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production". *Applied Linguistics* 24 (1), 1-27.

Zydatiß, Wolfgang. 2007. *Deutsch-Englische Züge in Berlin (DEZIBEL).* Frankfurt am Main: Peter Lang.

List of Appendices

Appendix 1: Texts used for the C-test

Internet

More than 600 million people worldwide have some sort of access to the internet. That is an astonishing number, and reflects the rapid growth of the network since it was invented in the 1970s. However, that still leaves about 5.5 billion people who do not use the net and who have no access. Most of these people live outside the developed Western countries. While over half of UK households are online, only 0.1% of homes in Bangladesh can claim the same.

Source: Thompson 2002.

China

China consumes more steel, coal, meat and grain than any other nation. It is also the world's fifth largest exporter, trading extensively with the EU, Japan and the US. In 2006, 80% of the world's consumer electronics were made in China. The rapid economic progress has transformed cities and coastal areas. But for those in China's underdeveloped rural interior life remains difficult. The gap between rich and poor in China is one of the biggest in the world.

Source: “Quick Guide: China” 2006.

PowerPoint

Public speaking has always ranked very high on the list of the most fearful experiences. Those good speakers who are not afraid of speaking in public are able to speak without fear because of a few simple things. Primarily, they know the subject matter, and more importantly, they know how their visual aids will help communicate the subject matter. Knowing how to create simple but effective visual aids in PowerPoint will help you with the second part of this equation.

Source: Reilly 1997: xxii.

Broadband

The two big advantages of broadband over dial-up are that it frees up the telephone line for voice calls and can be left on permanently. It still requires a modem, which is paid for when you sign up. Broadband allows you to browse the internet much faster as the pages download more quickly. It can be used for downloading files, such as music or films. In the first quarter of 2005, 4.6 million tracks were downloaded, nearly as many as in the whole of 2004.

Source: “Quick Guide: Broadband” 2006

Talent wars

Firms all over the world are reporting a shortage of employees, and not just in the highly qualified areas where the "talent wars" are being fought. According to a new survey by staffing company Manpower, the ten jobs that companies are having most difficulty filling range from top management to factory workers (see table). The biggest demand is for sales representatives with enough technical knowledge to sell today's complex products. "They need to understand innovations, logistics and the global picture. Selling is very different from what it was," Manpower head Jeff Joerres told BusinessWeek.

Source: "The war for talent 200: p.9.

Happy Planet Index

The Happy Planet index is a measure of the ecological efficiency with which human well-being is delivered. In an age of climate change, it gives a better picture of the true health and wealth of nations. Using new data this report reveals that Europe is less carbon efficient now than it was 40 years ago at delivering human well-being in terms of relatively happy, long lives to its citizens. The Index explores why some European countries produce well-being at a much higher cost than others.

Source: Thompson et al 2007: 1

Mobile phones

Passengers could soon be using their mobile phones on planes flying through European airspace. Plans have been developed across EU countries to introduce technology which permits mobile calls without risk of interference with aircraft systems. Regulators around Europe are calling for consultation on the potential introduction of the technology. If given the go ahead, the service would allow calls to be made when a plane is more than 3,000 metres high. Individual airlines would need to decide if they wanted to introduce the technology, if the green light is given by national regulators.

Source: "Mobile phone use" 2007.

Energy

Since 1950 oil, coal, and natural gas have supported most of the world's economic growth. Their use is also responsible for much of the world's pollution and environmental degradation. Nuclear energy was supposed to be providing much of the world's electricity by the year 2000. But high costs (even with massive government subsidies), safety concerns, and failure to find an economically and politically acceptable solution for storing its long-lived radioactive wastes have led many countries to sharply scale back or eliminate their plans to build new nuclear power plants.

Source: Miller 1991: 408.

Appendix 2: Test instructions

School:

Class:

Name:

Dear student!

My name is Silvia Jexenflicker and I am a student of English at the University of Vienna. By taking part in this test you are helping me to collect valuable data for my thesis ("Diplomarbeit") on language skills and language testing and I am very grateful for your support. The test consists of two parts. First I would ask you to complete six short texts from various fields of interest, which should take about 30 minutes. The second part consists in writing an e-mail according to the instructions given. All in all the test should take about 50 minutes.

For my analysis I would also ask you to fill in a short questionnaire about yourself as a learner of English. Of course, I will not publish any of your personal data (including your name) in any form. The results of the study will be presented in anonymised form only.

I would like to thank you again for your participation in this study. If you have any questions please feel free to contact me on silvia.jexenflicker@aon.at.

Thank you very much for your support!

Silvia Jexenflicker

Part 1: Gap-filling

This is a test of how well you comprehend written English. In the following texts **the second half** of some of the words has been deleted, i.e. half of the letters (plus one) are missing. (This means that the length of the gap roughly indicates the length of the word.)

First study each text. Then, try to reconstruct the words. Do not fill in any extra words and do not change the letters left standing. **You should not spend more than 5 minutes on any of the texts.**

Example: Is t____ Austrian nati_____ team go_____ to b__ the Euro_____ champion?

Solution: Is **the** Austrian **national** team **going** to **be** the **European** champion?

Text 1

More than 600 million people worldwide have some sort of access to the internet. That i_ an aston_____ number, a____ reflects t____ rapid gro_____ of t____ network si_____ it w____ invented i__ the 1970s. How_____ , that st_____ leaves ab_____ 5.5 billion peo_____ who d__ not u____ the n____ and w____ have n__ access. Mo____ of th_____ people live outside the developed Western countries. While over half of UK households are online, only 0.1% of homes in Bangladesh can claim the same.

Text 2

China consumes more steel, coal, meat and grain than any other nation. It i__ also t____ world's fifth lar_____ exporter, tra_____ extensively wi____ the EU, Japan a____ the US. I__ 2006, 80% o__ the wor_____ consumer elect_____ were ma____ in Ch_____ . The ra_____ economic prog_____ has trans_____ cities a____ coastal ar_____ . But f____ those i__ China's underde_____ rural interior life remains difficult. The gap between rich and poor in China is one of the biggest in the world.

Text 3

Public speaking has always ranked very high on the list of the most fearful experiences. Those go____ speakers w____ are n____ afraid o__ speaking i__ public a____ able t__ speak wit_____ fear bec_____ of a f____ simple thi_____ . Primarily, th____ know t____ subject mat_____ , and mo____ importantly, th____ know h____ their vis_____ aids wi____ help commu_____ the subject matter. Knowing how to create simple but effective visual aids in PowerPoint will help you with the second part of this equation.

Text 4

Passengers could soon be using their mobile phones on planes flying through European airspace. Plans ha____ been deve_____ across EU coun_____ to intr_____ technology wh_____ permits mob_____ calls wit_____ risk o__ interference wi____ aircraft sys_____. Regulators aro_____ Europe a____ calling f____ consultation o__ the pote_____ introduction o__ the techn_____ . If gi_____ the g__ ahead, t____ service would allow calls to be made when a plane is more than 3,000 metres high. Individual airlines would need to decide if they wanted to introduce the technology, if the green light is given by national regulators.

Text 5

Firms all over the world are reporting a shortage of employees, and not just in the highly qualified areas where the "talent wars" are being fought. According t__ a n____ survey b__ staffing com_____ Manpower, t____ ten jo____ that comp_____ are hav_____ most diffi_____ filling ra_____ from t____ management t__ factory wor_____ . The big_____ demand i__ for sa_____ representatives wi____ enough tech_____ knowledge t__ sell tod_____ complex products. "They need to understand innovations, logistics and the global picture. Selling is very different from what it was," Manpower head Jeff Joerres told BusinessWeek.

Text 6

Since 1950 oil, coal, and natural gas have supported most of the world's economic growth. Their u____ is al____ responsible f____ much o__ the wor_____ pollution a____ environmental degra_____ Nuclear ene_____ was supp_____ to b__ providing mu____ of t____ world's elect_____ by t____ year 2000. B____ high co_____ (even wi____ massive gover_____ subsidies), saf_____ concerns, a_____ failure to find an economically and politically acceptable solution for storing its long-lived radioactive wastes have led many countries to sharply scale back or eliminate their plans to build new nuclear power plants.

Part 2: Free writing (about 150-200 words):

Your school has organised a two-week stay in New York for you and your classmates. You will be attending language classes in the morning and you will be staying with a host family that your school has selected for you. Your teacher has encouraged you to establish first contact with this family (Mr and Mrs Ferguson, 2 children: Paul, aged 17, and Amanda, aged 15) by sending an e-mail or letter to them and has suggested that you include the following points:

- Personal information (name, age, family)
- Your school & future career plans
- Spare time activities and special interests
- Previous stays abroad, especially in English-speaking countries
- Why you are happy to go to NY and what you expect
- Two (or more) questions you would like to ask them
- A positive conclusion

You may, of course, add any other point that seems relevant or interesting to you.

Appendix 3: Rating scale used for assessment of the writing task (adapted from Friedl/Auer 2007)

	Task fulfilment:content and relevance; text format, length and register
5	Task fully achieved, content entirely relevant; appropriate format, length and register
4	Task almost fully achieved, content mostly relevant; mostly appropriate format, length and register
3	Task adequately achieved, some gaps or redundant information, acceptable format, length and register
2	Task achieved only in a limited sense, frequent gaps or redundant information, often inadequate format, length and register
1	Task poorly achieved, major gaps or pointless repetition; inadequate format, length and register
0	Not enough to evaluate
	Organisation: Structure, paragraphing, cohesion and coherence, editing and punctuation
5	Clear overall structure, meaningful paragraphing; very good use of connectives, no editing mistakes, conventions of punctuation observed
4	Overall structure mostly clear, good paragraphing, good use of connectives, hardly any editing mistakes, conventions of punctuation mostly observed
3	Adequately structured, paragraphing misleading at times, adequate use of connectives; some editing and punctuating errors
2	Limited overall structuring, frequent mistakes in paragraphing, limited use of connectives; frequent editing and punctuation errors
1	Poor overall structuring, no meaningful paragraphing, poor use of connectives; numerous editing and punctuation errors
0	Not enough to evaluate
	Grammar: Accuracy/ errors, variety of structures, readiness to use complex structures
5	Accurate use of grammar and structures, hardly any errors of agreement, tense, word order, articles, pronouns, etc.; meaning clear, great variety of structures, frequent use of complex structures
4	Mostly accurate use of grammar and structures, few errors of agreement etc.; meaning mostly clear; good variety of structures, readiness to use complex structures
3	Adequate use of grammar and structures; some errors of agreement etc.; meaning sometimes not clear; adequate variety of structures; some readiness to use complex structures
2	Limited use of grammar and structures; frequent errors of agreement etc.; meaning often not clear; limited variety of structures; limited readiness to use complex structures
1	Poor use of grammar and structures; numerous errors of agreement etc.; meaning very often not clear; poor variety of structures
0	Not enough to evaluate
	Vocabulary: Range and choice of words, accuracy, spelling, comprehensibility
5	Wide range of vocabulary; very good choice of words; accurate form and usage; hardly any spelling mistakes; meaning clear.
4	Good range of vocabulary; good choice of words; mostly accurate form and usage, few spelling mistakes; meaning mostly clear.
3	Adequate range of vocabulary and choice of words; some repetitions; some errors of form and usage; some spelling mistakes; meaning sometimes not clear; some translation from mother tongue
2	Limited range of vocabulary and choice of words; frequent repetitions; frequent errors of form and usage; frequent spelling mistakes; meaning often not clear; frequent translation from mother tongue
1	Poor range of vocabulary and choice of words; highly repetitive; numerous errors of form and usage; numerous spelling mistakes; meaning very often not clear; mainly translation from mother tongue.
0	Not enough to evaluate

Appendix 4: Abstracts and Curriculum vitae

Abstract

Globalisation and internationalisation are making increasing demands on the foreign language competence of European citizens. In reaction to this, a trend has emerged throughout Europe involving the use of Content and Language Integrated Learning (CLIL) in mainstream education, the basic idea being to use a language other than the students' L1 in teaching non-language subjects. This paper aims to investigate the effects of CLIL provision on the language output of students attending two Austrian colleges of engineering, crafts and arts ('HTLs') on the basis of a two-part test. The first part consists in a C-test, a variant of the cloze test, and is designed to provide a global statistic reflecting test takers' general language proficiency. While this test has the advantage of providing an objective assessment instrument, allowing the use of well-established statistical procedures, it does not meet the requirements of communicative language testing. Therefore it was complemented with a second task requiring the students to compose an e-mail or letter to a prospective host family. This part was designed, on the one hand, to provide a subjective measure of general language ability and, on the other, to assess test takers' writing skills. Generally speaking, the findings of the study support the hypothesis that CLIL students outperform their non-CLIL peers in terms of their general language competence as well as with regard to their writing skills. The evidence also suggests, however, that the extent to which this is the case depends on the subskill analysed, the advantage of the CLIL group being the least obvious in the field of organisation and structure, an area in which the level of performance was generally low. It can be concluded that, on the whole, the provision of CLIL in engineering colleges seems to have a positive effect on language output, with quantity of exposure as a key factor. Recommendations include a minimum level of exposure to the foreign language and a systematic approach to the implementation of CLIL if it is to move from the experimental stage to forming an integral part of the Austrian education system.

Zusammenfassung

Die zunehmende Globalisierung und Internationalisierung bringt unter anderem auch höhere Anforderungen an die Fremdsprachenkompetenz der Europäer mit sich. Ein möglicher Weg, sich dieser Herausforderung zu stellen, ist der verstärkte Einsatz von CLIL (Content and Language Integrated Learning), wobei eine Fremdsprache als Unterrichtsmedium im Sachfachunterricht eingesetzt wird. Ziel der vorliegenden Arbeit ist es, die Auswirkungen des CLIL-Unterrichts auf den allgemeinen Sprachstand und die schriftliche Kommunikationsfähigkeit von Schülern zweier österreichischer HTLs zu untersuchen. Zu diesem Zweck wurde ein zweiteiliger Test entwickelt. Den ersten Teil bildet ein C-Test, eine Variante des Cloze-Tests, der den allgemeinen Sprachstand auf objektive Weise misst. Um auch den Anforderungen kommunikativer Sprachtests zu genügen, besteht der zweite Teil im Verfassen einer e-mail bzw. eines Briefes an eine Gastfamilie in New York. Aufgrund der Ergebnisse der Studie kann die zugrundeliegende Hypothese, dass CLIL-Schüler jene ohne CLIL-Unterricht hinsichtlich ihrer allgemeinen Sprachkompetenz und ihrer schriftlichen Kommunikationsfähigkeit übertreffen, beibehalten werden. Allerdings ergeben sich Unterschiede in verschiedenen Teilbereichen, wobei der Vorsprung der CLIL-Schüler im Bereich der Textkompetenz am wenigsten ausgeprägt zu sein scheint. Allerdings erscheint das Niveau in diesem Bereich in allen untersuchten Gruppen niedrig. Zusammenfassend spricht das Ergebnis für die verstärkte Einführung von CLIL an höheren technischen Schulen, wobei das Ausmaß des Fremdspracheninputs von entscheidender Bedeutung zu sein scheint. Allgemeine Empfehlungen zielen unter anderem auf ein Mindestmaß für den Einsatz der Fremdsprache im CLIL-Unterricht und eine strukturierte und systematische Umsetzung von CLIL ab. Im Allgemeinen scheint der Zeitpunkt gekommen zu sein, CLIL vom experimentellen Stadium in einen fixen Bestandteil des Bildungsangebotes in Österreich zu überführen.

Europass Curriculum Vitae

Personal information

Surname(s) / First name(s)	**Jexenflicker Silvia Anna**
Nationality	Austrian
Date of birth	3 June 1967

Education and training

Dates	1996-2000
Title of qualification awarded	**1st diploma exam, English studies**
Principal subjects/occupational skills covered	English language skills, English literature, linguistics, culture of English-speaking countries
Name and type of organisation	University of Vienna
Dates	July 1998
Title of qualification awarded	**LCCI Certificate in Teaching English for Business** (passed with distinction)
Principal subjects/occupational skills covered	Teaching business English skills such as presentations, meetings & negotiations, business-related writing (e.g.correspondence, reports) etc.
Name and type of organisation	London Guildhall University
Dates	Since 1996
Principal subjects/occupational skills covered	Various **teacher training courses** on aspects such as teaching methods and strategies, stress and pronunciation, creative teaching, discourse, computer assisted learning, methodology and language for vocational schools, portfolio assessment, presentations; attended conferences for teachers of English
Name and type of organisation	Various institutions
Dates	1989-2004
Title of qualification awarded	**Doctor (of social and economic sciences)**
Principal subjects/occupational skills covered	English, economics, statistics; doctoral thesis written on “The green consumer – an agent of change? The possibilities and limitations of this concept, with special reference to the UK experience” (in English)
Name and type of organisation	Wirtschaftsuniversität Wien (Vienna University of Economics and Business Administration)
Dates	1985-1989
Title of qualification awarded	Magister der Sozial-und Wirtschaftswissenschaften (**Master of social and economic sciences**)
Principal subjects/occupational skills covered	Business studies (specialisation: market research and advertising), languages (English, Spanish, French), economics, commercial and public law, mathematics and statistics; master’s thesis on “Nonprofit marketing” (in English)
Name and type of organisation	Wirtschaftsuniversität Wien (Vienna University of Economics and Business Administration)

Work experience

Dates	Since October 2000
Occupation or position held	**Full-time lecturer for English and Business English, English Department**
Main activities and responsibilities	• Teaching General and Business English in the fields of business consultancy (bachelor and master's programmes) and health studies (bachelor programme) • Designing courses and contributing to curriculum development • Liaising with External Lecturers
Name and address of employer	Fachhochschule Wiener Neustadt für Wirtschaft, Technik, Gesundheit, Sicherheit und Sportmanagement (University of Applied Sciences Wiener Neustadt for Business, Engineering, Health Studies, Security and Sport)
Type of business or sector	Tertiary education
Dates	1990-2003
Occupation or position held	**Lecturer at the Department of English**
Main activities and responsibilities	Planning and teaching courses in the field of business English (Proseminars I + II)
Name and address of employer	Wirtschaftsuniversität Wien (Vienna University of Economics and Business Administration)
Type of business or sector	Tertiary education
Dates	October 1996-October 2000
Occupation or position held	**Free-lance trainer for general and business English**
Main activities and responsibilities	Planning and teaching courses in the fields of business English, English grammar , preparation for Cambridge Certificate for International Business and Trade (CEIBT)
Name and address of employer	Various institutions
Type of business or sector	Adult education, tertiary education
Dates	1989-1995 and 1997-98
Occupation or position held	**Teaching and Research assistant at Department of English**
Main activities and responsibilities	• Teaching business English & supporting departmental research • Organising and running seminars (2nd part of studies) • Co-tutoring students writing their master's thesis at the Department of English
Name and address of employer	Wirtschaftsuniversität Wien (Vienna University of Economics and Business Administration)
Type of business or sector	Tertiary Education

Language skills

Mother tongue(s)	German
Other language(s)	English, basic knowledge of Spanish and French