ISSN 1798-4769
Journal of Language Teaching and Research, Vol. 5, No. 3, pp. 487-497, May 2014

doi:10.4304/jltr.5.3.487-497

Vocabulary in CLIL and in Mainstream Education

Riika Merikivi
Department of English, University of Turku, Turku, Finland

Päivi Pietilä
Department of English, University of Turku, Turku, Finland

***Abstract*—The focus of the study reported in this article was vocabulary size attained in two learning environments, i.e. in regular mainstream instruction and in CLIL (Content and Language Integrated Learning). Receptive and productive vocabulary sizes of sixth-graders from both environments were compared with the respective vocabulary sizes of corresponding ninth-graders using the Vocabulary Levels Test and the Productive Vocabulary Levels Test. It was hypothesized that CLIL would produce larger vocabularies, as it offers learners more extensive and versatile exposure to the target language. This turned out to be the case, as did the previously attested phenomenon that receptive vocabularies are larger than productive vocabularies. However, the development of the productive-receptive ratio was not uniform across the frequency levels, even though it was at its highest at the third frequency band (3000 most common English words).**

***Index Terms*—CLIL, L2 vocabulary, vocabulary size, lexical frequency levels, language learning**

I. Introduction

The main objective of the study reported in this article was to compare vocabulary acquisition in two different learning environments, i.e. in regular mainstream classes and in CLIL (Content and Language Integrated Learning) instruction. The study investigated the development of English vocabulary of pupils in the sixth and ninth grades of the Finnish comprehensive school, using both receptive (the Vocabulary Levels Test (VLT); Nation, 1983, 1990) and productive (the Productive Vocabulary Levels Test (PVLT); Laufer & Nation, 1999) tests to measure the size of the learners' vocabularies. The study was, thus, concerned with the breadth of the learners' vocabulary knowledge and not the depth of it, as the tests that were used only addressed the form-meaning connections of words. We hypothesized that students in CLIL classrooms would have bigger receptive and productive vocabularies than those in mainstream education because of the larger amount of foreign language input that is available in CLIL. The benefits of CLIL education have been documented in previous studies (e.g. Dalton-Puffer, Nikula & Smit, 2010), although a recent review by Sylvén (2013) demonstrates that CLIL does not always lead to better learning results compared with traditional foreign language learning methods. It is important to take into consideration a number of nation-specific contextual factors which influence learning in different countries. According to Sylvén, the not so encouraging results of CLIL in Sweden can at least partly be attributed to the fact that CLIL is not recognized in the national curriculum and, therefore, also the amount of research into CLIL is relatively scarce in Sweden. On the other hand, the generally high level of English proficiency of young people in Sweden seems to be largely due to the abundance of English in Swedish society today. In her 2004 study on vocabulary development of Swedish learners of English, Sylvén had already concluded that the importance of exposure to English outside school was greater than that of CLIL (Sylvén, 2004).

The present article focuses on the situation in Finland, which in some ways is different compared with Sweden. First of all, immersion and CLIL programmes are recognized and encouraged in the national curriculum, which gives schools a great deal of freedom to plan their teaching. This is probably one reason why there has been a considerable amount of research interest in the functioning and learning results of CLIL (e.g. Järvinen, 1999, 2005; Nikula, 2005). Secondly, Finland requires CLIL teachers to have a specific level of foreign language competence (at least C1 on the CEFR proficiency scales) and offers them both pre- and in-service training in CLIL teaching (Järvinen, 2012). Moreover, CLIL programmes are implemented in all school levels, unlike in many other countries.

In addition to investigating the entire vocabulary sizes of the participants in CLIL and in mainstream education, the relative sizes of their receptive and productive vocabularies were also examined, particularly as they bear on word frequency.

II. Vocabulary Size

How many words can L2 learners of English be expected to learn? Nation and Waring (1997) estimate that native speakers typically learn an average of 1000 word families[1] a year in their early life. According to them, this goes on up to the vocabulary size of around 20000 word families; thus, a child beginning school at the age of five has a vocabulary of about 4000 to 5000 words and a 20-year-old knows approximately 17000 to 20000 words. The vocabularies of L2 learners are generally substantially smaller than those of native speakers. Table I summarises the results of studies by Schmitt and Meara (1997) and Laufer (1998) on L2 vocabulary sizes (in word families).

TABLE I.
PREVIOUS RESEARCH ON L2 VOCABULARY SIZES

	Schmitt & Meara (1997)	Laufer (1998)
Years of acquisition	5-6	6-7
Receptive vocabulary	3900	3500
Productive vocabulary		2550

The results of the present study might be expected to resemble those depicted in Table I, at least to some extent, as the number of years that the subjects had learned English in those studies (in Japan and in Israel, respectively) were rather similar to the situation in the present study, and the vocabulary tests that were used were the same (VLT and PVLT). The subjects in the present study, however, came from two different learning environments, so we expected there to be a difference in their learning outcomes.

III. WORD FREQUENCY

A good measure of word usefulness is *frequency*, i.e. how often a word occurs in normal use of a language. A small number of the words in English, for instance, occur very frequently and thus comprise large proportions of both written and spoken texts. It makes sense that language learners are taught words which belong to the most frequent lexemes of the language, as those frequent words are likely to be the most useful. According to Laufer and Nation (1999), teachers of English should focus on the 2000 most frequent words, i.e. on *high-frequency words*, and instead of teaching individual words which are less frequent, they should introduce strategies for coping with unfamiliar vocabulary. Several studies have also found, unsurprisingly, that high-frequency words are mastered better than low-frequency words (e.g. Laufer & Nation, 1999), and that learners with larger vocabularies use more low-frequency words than learners with smaller vocabularies (Laufer & Nation, 1995).

In a study involving students at the final stages of lower secondary school (i.e. from the same age group as the 9th graders of the present study) in Denmark, Stæhr (2008) found that the 2000 most frequent English words constituted an important watershed and a sensible learning goal, as those learners who knew those frequent words (the minority of the pupils) also performed relatively well in listening, reading and writing. The vocabulary size test used by Stæhr was the same as the one used in the present study to measure receptive vocabulary (the VLT devised by Schmitt, Schmitt and Clapham, 2001, based on Nation, 1983), except that he had excluded the academic word level from the test, claiming that it was not relevant for low-proficiency learners. We decided to keep the academic words in our test, as they occur quite commonly, for example, in textbooks (for more information, see Section VII C).

IV. RECEPTIVE AND PRODUCTIVE VOCABULARY KNOWLEDGE

An L2 learner's lexical competence consists of *receptive* (*passive*) and *productive* (*active*) vocabulary knowledge. Generally, receptive vocabulary comprises the words a person is able to understand, whereas the words in productive vocabulary are not merely understood but they can also be produced.

Laufer (1998) differentiates between *passive knowledge*, *controlled active knowledge* and *free active knowledge*. Passive knowledge includes comprehending the core meaning of a word. Active knowledge is divided between words one can produce when required to do so (controlled active knowledge) and words one would use without prompts to specific items, as is the case when writing a composition (free active knowledge). She considers the division necessary because people tend to provide different lexical items when the situation necessitates particular words as opposed to when they are left to their own devices. According to Laufer and Nation (1999), this usually has to do with word frequency, as people may hesitate to use some infrequent words of their controlled active vocabularies in free production and, consequently, end up choosing simpler, more frequent synonyms instead.

It is commonly acknowledged that reception precedes production and that production is more demanding than comprehension (cf. Waring, 1997). Furthermore, passive vocabulary is typically regarded as wider than active vocabulary. Laufer (1998) found that passive vocabulary size was larger than controlled active vocabulary, the mean ratio being as high as 80%, and the difference in the sizes was bigger in the more advanced learner group (11th graders) than in the less advanced group (10th graders), indicating that passive vocabulary grows considerably faster than active vocabulary. Nemati (2010), on the other hand, compared passive and controlled active vocabulary competence and discovered that the productive vocabularies of her participants were around 20% of their receptive vocabularies. According to her results, the ratio was not fixed but mounted along development, implying a narrower gap between the two knowledge

[1] A *word family* consists of a head word, its inflected forms and its transparent derivations (cf. Nation & Waring, 1997).

systems at higher levels of learning. Laufer and Paribakht (1998) noticed both tendencies: higher proficiency led to an increase in the ratio in their EFL group and to a decrease in their ESL group. It is worth noticing that the results of these studies vary substantially despite the fact that they exploited the same instruments (i.e. the Levels Tests), revealing the complexity of the phenomenon. Acknowledging the widely documented supremacy of reception over production in language learning, our study took the matter one step further and examined the reception-production relationship in the various frequency bands of the English lexicon.

V. CLIL as a Learning Environment

The study reported here examined the vocabulary sizes attained by learners of English in two different learning environments, regular foreign language education involving two or three hours of English instruction per week, and content and language integrated learning (CLIL), where most school subjects were taught in English (L2), within the Finnish educational system.

As a learning environment, CLIL has its roots in the immersion education which first began in Canada in the 1960s (e.g. Swain & Johnson, 1997). It also relies on Krashen's (e.g. 1983) notions of comprehensible input and on the ideas of the so-called communicative approach. Krashen's work has been reappraised on a number of occasions, but the core ideas still constitute some of the fundamental principles of CLIL. The notions that a degree of learning is automatic (i.e. learning can occur even when attention is not consciously paid to vocabulary or structures), and that learning is partly related to exposure (i.e. high exposure to a second language is seen to enhance language mastery) are still of marked importance for CLIL (cf. Wode, 1999). On the other hand, as also output is claimed to be of crucial importance if learning is to be successful (Swain, 1996; see also Swain & Lapkin, 1995), CLIL typically provides input that learners can understand as well as gives them plenty of opportunities for meaningful interaction using the L2.

However, N. Ellis (1994) and Schmidt (2001) strongly oppose Krashen's conception of the minimal role of consciousness and attention, and state that acquisition never occurs in a totally incidental manner but is a by-product of attention to relevant, rule-governed structures. As far as lexical learning is concerned, N. Ellis (1994) claims that both implicit and explicit operations are crucial in L2 word acquisition. Verspoor and Lowie (2003) and Carter (2001) argue that sometimes looking up words in dictionaries or using explicit mnemonic strategies to commit words to memory may be more beneficial than simply meeting words in various contexts over time.

VI. Learning in CLIL

Research concentrating on the development of L2 skills in CLIL classes has mostly produced promising results. Järvinen (1999) tracked the syntactic development of relativization and found that the CLIL group could produce sentences significantly longer, more complex and more accurate than the mainstream control group, in an elicited imitation task. Valtanen (2001) investigated overall English proficiency at the end of lower secondary school. He exploited a national language test and measured English skills in reading, listening, writing, speaking, and grammar. CLIL learners scored on average higher than their peers, attending the monolingual stream, in each section of the test, which, in line with Järvinen, implies that CLIL has a clear positive effect on the development of English competence as a whole.

Studies on vocabulary learning in CLIL are still fairly sparse. Karonen (2003) explored lexical organisation by conducting a word association test on both CLIL students and students following the mainstream curriculum. She hypothesised that CLIL learners' lexicons would be more organised and that they would give more paradigmatic responses typical of native speaker adults. On the other hand, she expected the learners in formal language instruction to react to the stimulus differently by giving more syntagmatic responses, typical of native speaker children. Contrary to her expectations, however, no developmental syntagmatic-paradigmatic shift was found in either of the groups. As far as we can see, one possible explanation for this may be that her subjects' (5^{th}, 7^{th} and 9^{th} graders) lexicons were still so narrow that the organisation dimension simply had not developed sufficiently (cf. Meara, 1996).

Nevertheless as mentioned in Section I, some studies conducted in Sweden have produced contradictory results. Sylvén (2004) concluded that exposure to English outside school was more influential in the development of the learners' vocabulary skills than participation in CLIL. Similarly, Lim Falk (2008) discovered that there was less interaction (and hence L2 use) in a CLIL classroom than in a mainstream class. Results of this kind are in the minority but they certainly warrant a careful consideration of individual CLIL situations. Learning outcomes are the result of a combination of factors, and, therefore, generalizations should be made with caution.

In conclusion, research has indicated that, by and large, learning in a CLIL environment seems to have a more favourable effect on pupils' L2 skills than the monolingual stream, although many questions still remain open in this area. As for word acquisition, optimal learning may be best attained by combining explicit and implicit learning conditions which is exactly what a CLIL context normally does. Besides receiving formal instruction, the pupils are exposed to authentic input and have a multitude of opportunities to communicate and practise. The matter is far from established, however, and relatively little is known about the effects of CLIL on second language vocabulary development, which is precisely the focus of the present study.

VII. The Study

The main objective of the present study was to examine how much English vocabulary pupils learn in a general (GEN) and in a CLIL classroom by the end of the lower and upper levels of the Finnish comprehensive school, i.e. by the end of the sixth and ninth grade. The vocabularies were studied from the perspectives of receptive and productive knowledge and word frequency.

A. Research Questions

The study at hand sought to answer the following three questions:

1. How large are sixth-graders' passive and active vocabularies in GEN and in CLIL education?
2. How large are ninth-graders' passive and active vocabularies in GEN and in CLIL education?
3. How does word frequency affect vocabulary learning, especially the active-passive ratio?

B. Subjects

Altogether 367 Finnish comprehensive school pupils took the vocabulary tests, but the sample was pruned to comprise 330 pupils according to the information gathered based on a background information questionnaire. Only the learners who had Finnish as their mother tongue, who did not have English as the language spoken at home and who had not lived in an English speaking country for any significant period of time (here defined as more than two weeks) were chosen. The sample was, thus, as homogenous as possible, and therefore, the educational context (CLIL or mainstream instruction) could conceivably be assumed to account for the test results.

A total of 149 of the 330 subjects were finishing primary school and 181 lower secondary school, which means that the participants were aged around 13 and 16 years, respectively. The subjects in the experimental groups were CLIL pupils in grades six (CLIL6) (N=75) and nine (CLIL9) (N=88). Altogether eight CLIL classes participated, four of them primary and four lower secondary classes. As is common in Finland, their teaching had followed the national content curriculum, but they had been exposed to English from the first grade onwards (cf. Nikula & Marsh, 1999), the proportion of English having been largest in the lower grades (even close to 80% in the first and second grades), after which it had started to decline to reach approximately 40% of class time in the ninth grade. Their formal English instruction had begun in the third grade, similarly to mainstream education.

The participants in the control groups were pupils attending the general monolingual education in grades six (GEN6) (N=74) and nine (GEN9) (N=93). In total, nine mainstream classes were tested, four of them at primary level and five at lower secondary level. As stated above, the content syllabus had been congruent with CLIL education and the formal English classes had begun in the third grade, entailing that the mainstream participants had studied English for four and seven years, respectively. Table 2 shows the ages of the subject groups and the amount of English instruction that they had received.

TABLE II.
SUBJECTS OF THE STUDY

Groups	N (total = 330)	Age	Instruction in English (years)	Instruction in English (h)
CLIL 6	75	13	6	2600
CLIL 9	88	16	9	3400
GEN 6	74	13	4	330
GEN 9	93	16	7	600

The figures indicating the number of hours are estimates based on the general guidelines of CLIL teaching and formal foreign language teaching in Finland. The considerably lower figures of the GEN groups refer to formal English classes which amount to about 2.3 hours per week throughout the school year.

C. Methods and Procedures

The data were collected in a total of seventeen classes in eleven comprehensive schools situated around southwest Finland. The collection took place in the spring term in 2011. All the schools were state schools of typical, if different, sizes and thus in all likelihood representative of the Finnish educational context.

As the present study aimed to explore both receptive and productive vocabularies, two distinct tests were included in the test pattern. The Vocabulary Levels Test (Nation, 1983, 1990) was administered in order to measure receptive vocabulary knowledge whereas the Productive Vocabulary Levels Test (Laufer & Nation, 1999) was used to gauge productive vocabulary knowledge. These tests were chosen because, at present, they embody the nearest-to-standard instruments in the field of vocabulary testing, they are widely used and, most importantly, their validity has been assessed by distinct scholars (see e.g. Beglar & Hunt, 1999; Laufer & Nation, 1999; Schmitt, Schmitt & Clapham, 2001).

The validity and reliability of the study were enhanced by diminishing the effect of extraneous variables. In order for the testing situation to be as familiar and neutral as possible, the testing was implemented under normal classroom circumstances during schooldays. To provide enough time, the tests were conducted on two different occasions, the duration of each being 45 minutes, and to counteract lassitude, there was a break of 15 minutes between the sessions. This time frame proved to be sufficient for finishing the tests without haste, as the fastest pupils completed them in approximately 25 minutes and even the slowest ones did not use more than 40 minutes. On the first testing occasion, the par-

ticipants were asked to fill in the background information form and to take the receptive test and on the second, they were asked to take the productive test.

1. The Vocabulary Levels Test

The Vocabulary Levels Test (VLT) was originally devised by Nation (1983). It is described and is available in Nation (1983, 1990). Recently, the initial test has been replaced by improved versions 1 and 2 by Schmitt, Schmitt and Clapham (2001), found in Schmitt (2000) and in Nation (2001), respectively. The present study exploited Version 1. The VLT was in fact originally not designed to estimate learners' vocabulary size but as a diagnostic test to be used for pedagogic purposes. However, it has been used to measure vocabulary size in a number of studies (Stæhr, 2008). The version used in this study consists of five parts, representing the following five levels of word frequency in English: the levels of 2000, 3000, 5000 and 10000 words and academic words.

The 2000 and 3000 word levels contain high-frequency words. Knowledge of the 2000 most common words provides the resources required for rudimentary everyday spoken discourse, whereas the next 1000 words provide additional material for oral uses as well as enable learners to begin to read at least some unsimplified materials. The 5000 word level represents the ultimate boundary of high and low-frequency items. The words below this threshold are central if one wishes to read authentic texts fairly fluently. The 10000 word level contains low-frequency items. An L2 learner with a vocabulary of the 10000 most common words can be considered notably proficient as he can read practically any texts, excluding specialised materials, without major difficulty. Finally, the academic word level is based on Coxhead's (2000) Academic Word List. This level of formal words contains specialised vocabulary important for learners who engage in an English-medium learning environment. The items in it occur widely in textbooks and in other academic materials. It is not a separate frequency level but it contains items from the fourth to sixth levels.

As for the format, the test involves matching a word with a suitable definition. At each level, there are 60 words and 30 definitions, in groups of six and three respectively, as in example (1) from the 2000 word level:

(1) 1 cream
2 factory _____ part of milk
3 nail _____ a lot of money
4 pupil _____ person who is studying
5 sacrifice
6 wealth

According to Nation (1990), this format allows test-takers to exploit whatever knowledge they have of the meanings of the words; the option words in each cluster are selected so that they are not related in their meanings nor do they have similar orthographic forms. Thus, an examinee should be able to make the appropriate match even if s/he had only a rough impression of the meaning of a word. Nation (1990) also reports that all the words in each section were selected so that they would be representative of the words at that frequency level. In this way the results of the test are presumed to provide an estimate of the proportion of the words at each level that a learner knows.

2. The Productive Vocabulary Levels Test

The Productive Vocabulary Levels Test (PVLT) seeks to measure the ability to provide a word when required to do so by the given context. In other words, it is a test of controlled productive knowledge. The PVLT has been modelled on the original VLT by Laufer and Nation (1999). There are two equivalent versions available, both of which can be found in Laufer and Nation (1999). The present study exploited Version 2. Instead of meaning-definition matching, as in the VLT, the examinees are presented with sentences including a missing word and required to fill in the blanks with appropriate target words. Each frequency level section consists of eighteen sentences. Examples (2) and (3) from the 2000 word level elicit the words *hungry*, and *usual*.

(2) They sat down to eat even though they were not hu_____.

(3) This work is not up to your usu_____ standard.

According to Laufer and Nation (1999), the first few letters of the target words are always provided in order to prevent the test-takers from filling in some other word which may be semantically suitable in the given context but which comes from a wrong frequency level.

3. Calculating Vocabulary Size

As for the Vocabulary Levels Test, each frequency level section consists of 30 items. Therefore, the maximum score for each level is 30 and for the whole test 150. The answers were scored as correct or incorrect, and each correct match was given one point. As regards the Productive Vocabulary Levels Test, the total number of items is 90, with 18 items at each frequency level. This holds that the maximum score for the test is 90 and for each level 18. As in the VLT, the scoring was in terms of correct (one point) or incorrect/blank (zero points). An answer was considered correct when the item was semantically appropriate. If used in a wrong grammatical form, for instance the wrong tense, it was nevertheless accepted. A word spelled wrongly was not marked as incorrect either if the word was nonetheless recognisable and if the error did not distort the word (e.g. **dwel* instead of *dwell*). Most of the incorrect answers were non-words or existing words incorrect in the particular context, like ro*om* instead of ro*ots* in example (4) from the 2000 word level.

(4) Plants receive water from the soil through their ro_____.

As discussed above, the scores are thought to indicate the proportion of words a learner may know at a particular frequency level. This entails that they do not actually reveal much of the total number of words the learner knows, and,

consequently, the size of the lexicon has to be calculated separately. We decided to apply a formula introduced by Laufer (1998) since her way of treating the missing levels and the additional university word level is generally accepted. Laufer had excluded the last frequency level from her test, so the calculations she made represent a vocabulary of merely 5000 word families. Furthermore, she used the older version of the VLT which holds that the maximum score for each level was 18, as opposed to 30. According to Laufer (1998), the first and the second thousand levels can be assumed to have a corresponding score, whereas the fourth level score can be taken as an average of the third and fifth levels. The sum of the scores at all the levels is then multiplied by 5000 and divided by 108 (eighteen items per level for six levels).

Up to the fifth level, the calculation in the study at hand was done following Laufer. As for the upper levels, we finally chose to extend her model to cover also the sixth, seventh, eighth and ninth levels. That is to say, the missing levels were taken as an average of the fifth and tenth levels. This solution entails a great deal of rough estimation and gives a disproportionate weighting to the results of the third and fifth sections of the tests but, on the other hand, it exploits all the information that is available and builds directly upon the results elicited in the tests. After having calculated a score for each frequency level, we multiplied the sum of the scores at all the levels by 10000 (as the tests with all the levels included represent a lexicon of 10000 word families) and divided it by 330 in the case of the VLT (thirty items per level for eleven levels – 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, academic words) and by 198 in the case of the PVLT (eighteen items per level for eleven levels).

VIII. Results

The participants' passive vocabulary knowledge was measured with the Vocabulary Levels Test and their active vocabulary knowledge was evaluated with the Productive Vocabulary Levels Test. The scores of the tests represent estimated receptive and productive vocabulary sizes of the participants, calculated in word families. The calculations were done according to the principles described in the previous section.

Research question one asked how large sixth-graders' passive and active vocabularies in mainstream and in CLIL education are. Table III displays the results of the receptive test for sixth-graders by frequency levels.

TABLE III.
SIXTH-GRADERS' RECEPTIVE VOCABULARY SIZE

	Vocabulary size		Difference betw. groups		
	GEN6	CLIL6	t	df	p
2000	410	702	10.08	147	<0.001
3000	251	547	9.77	137	<0.001
5000	141	417	10.97	120	<0.001
AWL	94	364	9.16	107	<0.001
10000	23	153	7.24	91	<0.001
Total	1853	4505	11.39	116	<0.001

As indicated in Table III, the mean passive vocabulary size of the GEN6 learners was approximately 1850, while the corresponding figure for the CLIL group was 4505. The difference between the groups was examined by using the t-test. As can be seen in Table III, significant differences were found across all frequency levels, together with the total score.

As regards active vocabulary knowledge, Table IV shows that the GEN6 pupils could, on average, produce nearly 800 word families, whereas the CLIL6 pupils' total estimated productive vocabulary size was somewhat less than 2300 words. As was the case with receptive vocabulary, also productive vocabulary showed significant differences between the two groups across all frequency levels.

TABLE IV.
SIXTH-GRADERS' PRODUCTIVE VOCABULARY SIZE

	Vocabulary size		Difference		
	GEN6	CLIL6	t	df	p
2000	198	488	11.48	136	<0.001
3000	154	381	9.9	129	<0.001
5000	39	128	5.84	91	<0.001
AWL	24	154	7.36	88	<0.001
10000	1	41	4.19	75	<0.001
Total	788	2271	9.22	102	<0.001

In conclusion, the results of the Levels Tests for the sixth-graders indicate that the CLIL learners scored, on average, significantly better than the pupils studying in the general classroom, i.e. they appear to have larger receptive and productive vocabularies.

Research question two addressed the passive and active vocabulary sizes of ninth-grade students in mainstream and in CLIL education. The scores in Table V suggest that the CLIL9 learners have generally succeeded better in the VLT than the mainstream ninth-graders. In other words, their passive vocabulary size seems to be, on average, larger. Significant differences were found across all of the tested frequency levels along with the total score.

TABLE V.
NINTH-GRADERS' RECEPTIVE VOCABULARY SIZE

	Vocabulary size		Difference		
	GEN9	CLIL9	t	df	p
2000	741	841	5.54	151	<0.001
3000	608	724	4.46	179	<0.001
5000	451	575	4.34	179	<0.001
AWL	472	625	4.87	179	<0.001
10000	239	325	3.18	179	0.002
Total	5161	6379	4.82	179	<0.001

Regarding active vocabulary knowledge, Table VI displays the scores of the PVLT for the ninth-graders. Significant differences were found for all the frequency levels and for the total score alike, with the CLIL students exhibiting larger productive vocabulary sizes, on average.

TABLE VI.
NINTH-GRADERS' PRODUCTIVE VOCABULARY SIZE

	Vocabulary size		Difference		
	GEN9	CLIL9	t	df	p
2000	499	646	5.77	179	<0.001
3000	427	559	4.66	179	<0.001
5000	152	257	4.52	179	<0.001
AWL	203	320	6.23	179	<0.001
10000	64	130	3.52	154	0.001
Total	2565	3742	5.26	179	<0.001

Research question three asked about the effect of word frequency on the relationship between passive and active vocabularies. The results presented here will show, first of all, how the two knowledge types correlated with each other. Table VII displays the correlation coefficients and the significance levels for each group individually and for all subjects together as measured by Pearson's correlation.

TABLE VII.
CORRELATION BETWEEN RECEPTIVE AND PRODUCTIVE VOCABULARY SIZE BY LEARNING ENVIRONMENT AND GRADE

		Vocabulary size		Correlation	
	N	Receptive	Productive	r	p
GEN6	74	1853	788	0.74	<0.001
CLIL6	75	4505	2271	0.87	<0.001
GEN9	93	5161	2565	0.84	<0.001
CLIL9	88	6379	3742	0.91	<0.001
All groups	330	4595	2462	0.91	<0.001

As indicated in Table VII, very strong positive correlations (r>0.8) were found between the receptive score and the productive score, indicating that generally pupils who scored high in the receptive test were likely to score high in the productive test as well. That is to say, it appears that wide passive knowledge usually means wide active knowledge. This holds across all groups, save GEN6 in which case the effect size can be considered strong (r>0.5). These results are not surprising, as they are in line with most earlier studies. To further investigate the relationship between the two vocabularies, the dimension of word frequency was introduced.

In what follows, the lexical profiles of the subjects are illustrated by frequency levels. Fig. 1 depicts the passive and active vocabularies of all of the participants at the tested word frequency levels and Table VIII presents the situation for each group separately. The maximum score for each level is 1000 word families.

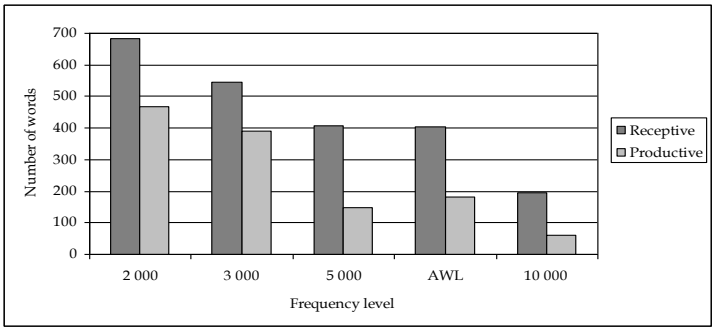

Figure 1. Receptive and productive vocabulary size at different word frequency levels

TABLE VIII.
RECEPTIVE AND PRODUCTIVE VOCABULARY SIZE BY LEARNING ENVIRONMENT AND GRADE AT DIFFERENT WORD FREQUENCY LEVELS

	GEN6	CLIL6	GEN9	CLIL9	All groups
2 000	410 / 198	702 / 488	741 / 499	841 / 646	648 / 468
3 000	251 / 154	547 / 381	608 / 427	724 / 559	545 / 390
5 000	141 / 39	417 / 128	451 / 152	451 / 257	407 / 149
Academic	94 / 24	364 / 154	472 / 203	625 / 320	404 / 183
10 000	23 / 1	153 / 41	239 / 64	325 / 130	194 / 62

Fig. 1 and Table VIII show a similar tendency: both passive and active scores decrease systematically with decreasing word frequency, suggesting that word frequency is in direct proportion to vocabulary knowledge. (That some of the scores at the academic word level are higher than at the 5000 word level does not contradict this finding. As pointed out earlier, the academic level consists of words from 4000 to 6000 frequency levels, which means that the academic words cannot necessarily be regarded as more infrequent than the words at the 5000 level.) Concretely, the subjects seem to have known more frequent than infrequent vocabulary across the entire sample, the smallest lexicons comprising almost exclusively common words.

Whether the differences in the numbers of words known at distinct frequency levels are statistically significant was examined by a one-way ANOVA. The academic word level is not a frequency level of its own and, therefore, it was not treated separately but as a part of the 5000 word level (see above). The ANOVA showed that word frequency had a significant effect on both receptive and productive vocabulary knowledge overall ($F=269$, $p<0.001$ and $F=395$, $p<0.001$, respectively), and the Scheffe post hoc test confirmed that the differences were highly significant across all of the frequency levels ($p<0.001$). That is, the learners appear to have known notably more common than rare words.

To view the relationship between word frequency and vocabulary learning further, an active-passive ratio for each frequency level was calculated. Table IX presents the ratios for each group of learners as well as for all groups together.

TABLE IX.
PRODUCTIVE-RECEPTIVE RATIO BY LEARNING ENVIRONMENT AND GRADE AT DIFFERENT WORD FREQUENCY LEVELS

	GEN6	CLIL6	GEN9	CLIL9	All groups
2000	47.3	68.8	66.4	76.2	68.4
3000	62.4	74.4	70.0	76.4	71.6
5000	30.9	28.4	32.6	41.3	36.6
Academic	23.1	43.3	46.2	50.7	45.3
10000	8.8	15.6	26.4	41.1	32.0

Table IX exposes an interesting phenomenon: the gap between active and passive vocabularies differs at different frequency levels, entailing that the receptive and productive scores do not diminish at the same rate from one level to another. The gap is narrower at the more frequent levels and wider at the less frequent levels, suggesting that rarer words were less likely to be part of the participants' active vocabulary knowledge. Rather unexpectedly, the active-passive ratio did not drop right after the second frequency band but only after the third, it actually being at its highest at that level. The difference between the 2000 and 3000 word levels is, however, small.

The significance of the differences between the ratios was explored by conducting a one-way ANOVA, the result of which implies that word frequency influences the active-passive ratio significantly ($F=170$, $p<0.001$). According to the post hoc Scheffe test, the differences were pronounced between all of the frequency bands ($p<0.001$), save the 2000 and 3000 word levels ($p>0.05$). This implies that the first three thousand words are more readily available for production than the more infrequent words.

In conclusion, the results suggest that word frequency and vocabulary learning are related. The learners knew significantly more vocabulary both receptively and productively at the higher word frequency levels than at the lower frequency levels. However, the relationship between active and passive vocabularies was not uniform across the levels in that the proportion of actively known words lessened drastically after the 3000 word level. This indicates that infrequent words were less likely to be active than the highly frequent first 3000 words.

IX. DISCUSSION

The primary objective of the present study was to estimate and compare the vocabulary sizes attained in mainstream and in CLIL education by the end of the lower and upper levels of the Finnish comprehensive school. Both receptive and productive lexicons were measured and examined in relation to frequency bands. The vocabulary tests used were the Vocabulary Levels Test (Nation, 1983, 1990) and the Productive Vocabulary Levels Test (Laufer & Nation, 1999).

The first and the second research questions attempted to estimate the vocabulary sizes of sixth- and ninth-graders studying in the traditional classroom and in CLIL. Table X shows a summation of the approximated vocabularies, as calculated in word families.

TABLE X.
SUMMATION OF RECEPTIVE AND PRODUCTIVE VOCABULARY SIZE

	Sixth grade		Ninth grade	
	GEN (N=74)	CLIL (N=75)	GEN (N=93)	CLIL (N=88)
Receptive size	1800	4500	5200	6400
Productive size	800	2300	2600	3700

As could be expected, the vocabulary sizes elicited in the present study lag behind native speaker word knowledge. At the age of thirteen an L1 learner would probably know well over 10000 word families and at the age of sixteen the number would typically be around 15000 to 16000 (cf. Nation & Waring, 1997). On the other hand, the results imply that the Finnish comprehensive school can provide a good setting for foreign language vocabulary acquisition in that the pupils were shown to learn receptively an average of around 700 word families and productively about 400 word families per year during their comprehensive education, estimations which are in line with previous research on English vocabulary learning in a foreign language context (cf. Table I).

Regarding the difference between the learning streams, our results confirm the prevailing perception that the CLIL environment is more fruitful for foreign language development than the monolingual stream. Namely, the scores of the vocabulary tests, summarized in Table X above, indicate that the average CLIL pupil could both comprehend and produce notably more words than the average GEN pupil in the same grade. The observed differences were statistically highly significant as measured by the t-test.

It can safely be assumed that the results are largely due to the different learning environments, CLIL classes providing the learners more opportunities to be exposed to English and to use it than regular classes. It is important to note that the CLIL participants had been admitted to the CLIL programmes (before starting the first grade) on the basis of their achievement in an entrance procedure which had not tested English language skills but sought to measure school readiness and general language awareness. Furthermore, it is noteworthy that in Finland, children start the CLIL programme typically without any knowledge of the foreign language, from zero, as it were, and they do not take additional language classes outside school as seems to be customary in some other countries, for example in Spain (Bruton, 2011).

As discussed above, vocabulary acquisition appears to be both an implicit and an explicit process, which implies that optimal learning may be best attained by combining explicit and implicit learning conditions (cf. e.g. R. Ellis, 1994; N. Ellis, 1993). As suggested by N. Ellis (1994), the process begins by unconscious form acquisition after which the meanings are learnt consciously. Furthermore, the faster the lower-level acquisition of forms becomes efficient, the sooner learners can focus on the higher-level acquisition of meanings and thereby extend their lexical knowledge. Besides involving formal instruction, CLIL provides frequent exposure to versatile and meaningful input and a multitude of opportunities to use the target language in situations that have a communicative function. This means that CLIL learners have more opportunities for both implicit form acquisition and explicit establishment of form-meaning links than GEN learners, accumulating larger lexicons. It is worth noting that the CLIL participants also reported reading in English outside school more often than their GEN peers, their interest in reading probably being a by-product of CLIL education. This issue is developed and discussed more thoroughly in another article by the same authors (Pietilä & Merikivi, 2014).

The third research question addressed the connection between active and passive vocabulary knowledge in relation to frequency bands of the lexicon. Our results reassert the prevalent view that the most frequent items are typically learnt before the rarer items (e.g. Read, 1988; Nation, 1990; Milton, 2006) in that the typical frequency pattern was observable across the data (cf. Richards & Malvern, 2007; Milton, 2007). This could be expected, as meeting and using the most common words of a language cannot, in practice, be avoided.

Most of the participants knew infrequent items without having full-scale competence of the first three frequency bands, strengthening the idea that learners do not acquire vocabulary strictly frequency level by frequency level (cf. Schmitt & Meara, 1997) but tend to pick up at least some lower-frequency items alongside learning the most common words. This seems only natural when it comes to our data, as comprehensive school pupils are commonly exposed to learning materials that often present words in semantic fields rather than purely in their order of frequency. That all the groups showed strong receptive and productive knowledge of common words at the end of the comprehensive school is another sign of successful language training, frequent lexicon forming the base for extensive reading and being indispensable in all communication.

Pupils in a CLIL context probably have substantially more time and opportunities to process the target language than pupils in a GEN context, and CLIL environments may require the students to produce the second language more frequently than GEN environments, factors likely to further lexical activation. Moreover, we believe that learners in CLIL may invest more effort into acquiring vocabulary in general than learners in GEN, as they have to cope in the CLIL language in their studies. If this is correct, the greater mental effort required for conscious learning may result in more forceful activation of the receptive lexicon in the CLIL pupils. This explanation is consistent with Swain and Lapkin (1995) and Carter (2001) who state that acquisition is more profound when students stretch their linguistic resource, and with N. Ellis (1994) who stresses the importance of explicit processes in thorough lexical learning, including word activation. On the other hand, Laufer (1998) found that by concentrating on vocabulary acquisition, her monolingual stream participants' active vocabulary size increased by 850 words on average during a single school year. In our opinion, the result implies that GEN is not automatically less efficient than CLIL and that also the CLIL stream could well afford to aim for larger active vocabularies. This is consistent with Järvinen (1999; see also Swain, 1993) who presumes

that in CLIL word acquisition should be emphasised and the students' own production encouraged even more to ensure optimal development of linguistic competence.

X. Conclusion

Many questions concerning vocabulary acquisition and different learning contexts remain unanswered. For example, according to our results, the CLIL environment does seem to be more conducive to word acquisition, especially the development of active vocabulary, than the traditional language learning classroom, but it may not make full use of its potential as regards lexical learning. It would therefore be interesting to examine which pedagogical solutions, if any, would alter the situation most favourably. Furthermore, whether mainstream pupils would attain the same results as, or even better results than, CLIL pupils if they also began receiving English instruction from the first grade onwards, can be speculated. Regarding the relationship between active and passive knowledge, the results of the present study strengthened the perception that a learner's receptive knowledge is greater than his/her productive knowledge and that the two vocabularies correlate strongly. However, we suggest that it would be too simplistic to express the connection merely in terms of mean ratio and correlation for all foreign language learners because the linkage was found to be neither uniform nor stable. On one hand, the proportion of active items became larger along development and, on the other, the active-passive ratio altered within a learner's lexicon at distinct frequency levels, being the highest at the most frequent levels. In the Finnish context, it would be of particular interest to study the development of a foreign language not as outstandingly present in our society as English is. In other words, it would be enlightening to examine whether decidedly fewer contacts with the target language would result, firstly, in a lower ratio between active and passive lexicons and, secondly, in a widening gap at higher levels of proficiency, as opposed to the tapering gap in the case of English.

Acknowledgement

The authors are very grateful to the anonymous reviewers for their observant and insightful comments and suggestions.

References

[1] Beglar, D. & A. Hunt. (1999). Revising and validating the 2000 Word Level and University Word Level Vocabulary Tests. *Language Testing* 16.2, 131-162.
[2] Bruton, A. (2011). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System* 39, 523-532.
[3] Carter, R. (2001). Vocabulary. In R. Carter & D. Nunan (eds.), *The Cambridge Guide to Teaching English to Speakers of Other Languages*. Cambridge: Cambridge University Press, 31-57.
[4] Coxhead, A. (2000). A new academic word list. *TESOL Quarterly* 34.2, 213-238.
[5] Dalton-Puffer, C., T. Nikula & U. Smit (eds.) (2010). Language use and language learning in CLIL classrooms. Philadelphia/Amsterdam: John Benjamins.
[6] Ellis, N. (1993). Rules and instances in foreign language learning: Interactions of explicit and implicit knowledge. *The European Journal of Cognitive Psychology* 5.3, 289-318.
[7] Ellis, N. (1994). Vocabulary acquisition: The implicit Ins and Outs of explicit cognitive mediation. In N. Ellis (ed.), *Implicit and Explicit Learning of Languages*. London: Academic Press, 211-282.
[8] Ellis, R. (1994). The Study of Second Language Acquisition. Oxford: Oxford University Press.
[9] Järvinen, H.-M. (1999). Acquisition of English in Content and Language Integrated Learning at elementary level in the Finnish comprehensive school. Turku: University of Turku.
[10] Järvinen, H.-M. (2005). Language learning in content-based instruction. In A. Housen & M. Pierrard (eds.), *Investigations in instructed second language acquisition*. Berlin: Mouton de Gruyter, 433-456.
[11] Järvinen, H.-M. (2012). What does a CLIL teacher need to know about language - the linguistic competence of a CLIL teacher. In A. Koskensalo, J. Smeds, R. de Cillia & A. Huguet (eds.), *Language: competence-change-contact; Sprache: Kompetenz-Kontakt-Wandel*. Munster: Lit Verlag, 47-54.
[12] Karonen, S. (2003). Semantic organization of L2 learners' mental lexicon: comparing pupils in mainstream education and CLIL classes. Unpublished MA Thesis, University of Turku.
[13] Krashen, S. (1983). Principles and practice in second language acquisition. Oxford: Pergamon Press.
[14] Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics* 19, 255-271.
[15] Laufer, B. & P. Nation. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics* 16, 307-322.
[16] Laufer, B. & P. Nation. (1999). A vocabulary-size test of controlled productive ability. *Language Testing* 16.1, 33-51.
[17] Laufer, B. & T. S. Paribakht (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning* 48.3, 365-391.
[18] Lim Falk, M. (2008). Svenska i engelskspråkig skolmiljö. Ämnesrelaterat språkbruk i två gymnasieklasser (Swedish in an English-language school environment. Subject-based language use in two upper secondary classes). Acta Universitatis Stockholmiensis. Stockholm Studies in Scandinavian Philology.
[19] Meara, P. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjær & J. Williams (eds.), *Performance and competence in second language acquisition*. Cambridge: Cambridge University Press, 35-53.

[20] Milton, J. (2006). Language Lite? Learning French vocabulary in school. *French Language Studies 1*6, 187-205.
[21] Milton, J. (2007). Lexical profiles, learning styles and the construct validity of lexical size tests. In H. Daller, J. Milton & J. Treffers-Daller (eds.), *Modelling and Assessing Vocabulary Knowledge*. Cambridge: Cambridge University Press, 47-58.
[22] Nation, P. (1983). Testing and teaching vocabulary. *Guidelines* 5, 12-25.
[23] Nation, P. (1990). Teaching and learning vocabulary. Boston: Heinle & Heinle Publishers.
[24] Nation, P. (2001). Learning vocabulary in another language. Cambridge: Cambridge University Press.
[25] Nation, P. & R. Waring (1997). Vocabulary size, text coverage and word lists. In N. Schmitt & M. McCarthy (eds.), *Vocabulary: description, acquisition and pedagogy*. Cambridge: Cambridge University Press, 6-19.
[26] Nemati, A. (2010). Active and passive vocabulary knowledge: the effect of years of instruction. *Asian EFL Journal* 12.1, 30-46.
[27] Nikula, T. (2005). English as an object and tool of study in classrooms: interactional effects and pragmatic implications. *Linguistics and Education* 16.1, 27-58.
[28] Nikula, T. & D. Marsh. (1999). Case study: Finland. In D. Marsh & G. Langé (eds.), *Implementing content and language integrated learning: a research-driven TIE-CLIL foundation course reader*. Jyväskylä: University of Jyväskylä, 17-72.
[29] Pietilä, P. & R. Merikivi. (2014). The impact of free-time reading on foreign language vocabulary development. *Journal of Language Teaching and Research* 5.1, 28-36.
[30] Read, J. (1988). Measuring the vocabulary knowledge of second language learners. *RELC Journal* 19, 12-25.
[31] Richards, B. & D. Malvern. (2007). Validity and threats to the validity of vocabulary measurement. In H. Daller, J. Milton & J. Treffers-Daller (eds.), *Modelling and Assessing Vocabulary Knowledge.* Cambridge: Cambridge University Press, 79-92.
[32] Schmidt, R. (2001). Attention. In P. Robinson (ed.), *Cognition and second language instruction*. Cambridge: Cambridge University Press, 3-32.
[33] Schmitt, N. (2000). Vocabulary in language teaching. Cambridge: Cambridge University Press.
[34] Schmitt, N. & P. Meara. **(**1997). Researching vocabulary through a word knowledge framework: word associations and verbal suffixes. *Studies in Second Language Acquisition* 19.1, 17-35.
[35] Schmitt, N., D. Schmitt & C. Clapham. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing* 18.1, 55-88.
[36] Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal* 36.2, 139-152.
[37] Swain, M. (1996). Integrating language and content in immersion classrooms: research perspectives. *The Canadian Modern Language Review* 52.4, 529-548.
[38] Swain, M. & R. Johnson. (1997). Immersion education: a category within bilingual education. In R. Johnson & M. Swain (eds.), *Immersion Education: International Perspectives*. Cambridge: Cambridge University Press, 1-16.
[39] Swain, M. & S. Lapkin. (1995). Problems in output and the cognitive processes they generate: a step towards second language learning. *Applied Linguistics* 16.3, 371-391.
[40] Sylvén, L.K. (2004). Teaching in English or English teaching? On the effects of content and language integrated learning on Swedish learners' incidental vocabulary acquisition. Ph.D. dissertation, University of Gothenburg.
[41] Sylvén, L.K. (2013). CLIL in Sweden – why does it not work? A metaperspective on CLIL across contexts in Europe. *International Journal of Bilingual Education and Bilingualism* 16.3, 301-320.
[42] Valtanen, J. (2001). The English language proficiency of 9th grade comprehensive school students in bilingual content and language integrated learning. Unpublished MA Thesis. University of Turku.
[43] Verspoor, M. & W. Lowie. (2003). Making sense of polysemous words. *Language Learning* 53.3, 547-586.
[44] Waring, R. (1997). A study of receptive and productive vocabulary learning from word cards. *Studies in Foreign Languages and Literature* 21, 94-114.
[45] Wode, H. (1999). Incidental vocabulary acquisition in the foreign language classroom. *Studies in Second Language Acquisition* 21.2, 243-258.

Riika Merikivi comes originally from Huittinen, but lives now in Turku, Finland. Her MA degree, with English philology as her major subject, is from the University of Turku, Finland (2012). She has formerly worked as a CLIL teacher and is now teaching English in a vocational institute in Turku. She is interested in researching foreign language learning, particularly vocabulary acquisition in a CLIL environment. Her research interests also include the relationship between receptive and productive vocabularies in different target languages and, more recently, child language development.

Päivi Pietilä was born in Pori, Finland. She received her MA degree, majoring in English, from the University of Turku in 1978; a *maîtrise ès lettres* degree from l'Université de Franche-Comté, France, in 1979; a Licentiate in Philosophy from the University of Joensuu, Finland, in 1983; and completed her PhD at the University of Turku, Finland, in 1990.

Her permanent position is that of Professor of English at the University of Turku, Finland, but in 2013, she divided her time as a visiting scholar at the University of Cambridge, UK, and the City University of New York, USA. Her research interests include second/foreign language learning, second language attrition, L2 speaking skills, L2 academic writing, vocabulary development and the lexis-grammar interface.