

The More the Merrier – Revisiting CLIL-Based Vocabulary Growth in Secondary Education

Erwin Maria GIERLINGER*
Thomas Arno WAGNER**

Abstract

One crucial aspect of CLIL-based foreign language learning in instructional settings is vocabulary growth. As a consequence, research should be interested in how CLIL fosters vocabulary learning. Noticing an apparent shortage of data-driven quantitative research on vocabulary growth in this field of CLIL is, therefore, problematic. The present paper reports findings from a mixed-methods study of vocabulary growth in an Austrian lower secondary school CLIL setting, with English as the language of instruction and learning. The aim of the study was to analyse how the use of CLIL in the English classroom could benefit learners in their acquisition of vocabulary in the target language. First, a repeated-measure-design with experimental and control groups assessed receptive vocabulary growth by means of a standardized vocabulary size test. Second, students' questionnaire data as well as vocabulary profiling of the CLIL teachers' linguistic input explored possible covariates for the vocabulary test scores. We found that CLIL-related effects were only co-determined by input frequency, while extra-mural factors did not play any role in this study. As a consequence, overly optimistic expectations regarding the linguistic impact of CLIL in a mixed-ability setting guided by a predominantly implicit language teaching approach need to be re-evaluated critically.

Keywords: CLIL; receptive vocabulary growth; vocabulary size test; frequency effects; extra-mural factors.

* University College of Education Upper-Austria, Austria.

** University College of Education Upper-Austria, Austria.

Corresponding author: erwin.gierlinger@ph-ooe.at

Received: 2015-07-01 / Sent for peer review: 2015-07-05 / Accepted by peers: 2015-09-20 / Approved: 2015-11-01

To reference this article in APA style / Para citar este artículo en APA / Para citar este artigo

Gierlinger E. & Wagner T. (2016). The more the merrier – revisiting CLIL-based vocabulary growth in secondary education. *LACLIL*, 9(1), 37-63. doi:10.5294/laclil.2016.9.1.3

Entre más mejor: una revisión del incremento del vocabulario en un contexto AICLE de bachillerato

Resumen

Un aspecto relevante del aprendizaje de la lengua extranjera basado en AICLE en contextos instruccionales es el incremento del vocabulario. En consecuencia, la investigación debería interesarse en la promoción del aprendizaje de vocabulario en AICLE. Es alarmante observar una aparente falta de investigación del incremento del vocabulario basada en datos cuantitativos. El presente artículo da cuenta de los hallazgos de un estudio mixto del incremento del vocabulario en un colegio austriaco de bachillerato de educación básica en un contexto AICLE, en el cual se emplea el inglés como lengua de instrucción y de aprendizaje. El objetivo de este estudio fue analizar cómo el uso de AICLE en el salón de clase pudo beneficiar a los aprendices en cuanto a la adquisición de vocabulario en lengua extranjera. En primer lugar, a través de repetidas mediciones con el grupo experimental y de control, se aplicó una prueba estandarizada del tamaño de vocabulario. Luego, se empleó un cuestionario de datos de los estudiantes, y mediante un perfil de vocabulario del aporte lingüístico de los profesores de AICLE, se exploraron las posibles co-variables de las pruebas de vocabulario. Se encontró que los efectos relacionados con AICLE estuvieron co-determinados por la frecuencia, mientras que los factores extramurales no tuvieron un rol preponderante en este estudio. En consecuencia, las expectativas demasiado optimistas relacionadas con el impacto de AICLE en un contexto de habilidades mixtas guiadas predominantemente por un enfoque de enseñanza implícita de la lengua deben reevaluarse seriamente.

Palabras clave: AICLE; incremento del vocabulario receptivo; prueba del tamaño del vocabulario; factores extramurales.

Quanto mais melhor: uma revisão do aumento do vocabulário num contexto de Aprendizagem Integrada de Conteúdos e de Língua de ensino médio

Resumo

Um aspecto relevante da aprendizagem de língua estrangeira baseada na Aprendizagem Integrada de Conteúdos e de Língua (AICL) em contextos institucionais é o aumento do vocabulário. Por isso, a pesquisa deveria estar mais voltada a sua promoção. É alarmante observar uma aparente falta de pesquisa sobre esse tema em dados quantitativos. Este artigo dá conta dos achados de um estudo misto do aumento do vocabulário num colégio austríaco de ensino médio num contexto de AICL, no qual se emprega o inglês como língua de instrução e de aprendizagem. O objetivo deste artigo é analisar como o uso da AICL na sala de aula pôde beneficiar os aprendizes quanto à aquisição de vocabulário em língua estrangeira. Em primeiro lugar, por meio de repetidas medições com o grupo experimental e de controle, aplicou-se um teste para averiguar o tamanho de vocabulário. Em seguida, empregou-se um questionário de dados dos estudantes e, mediante um perfil de vocabulário da contribuição linguística dos professores de AICL, exploraram-se as possíveis covariáveis das provas de vocabulário. Constatou-se que os efeitos relacionados com a AICL num contexto de habilidades mistas guiadas predominantemente por um enfoque de ensino explícito da língua devem ser reavaliadas seriamente.

Palavras-chave: AICL, aumento do vocabulário receptivo, teste de tamanho do vocabulário, fatores extramurais.

INTRODUCTION

CLIL is “an educational approach where curricular content is taught through the medium of a foreign language, typically to students participating in some form of mainstream education at the primary, secondary, or tertiary level” (Dalton-Puffer, 2011, p. 183). One of its major premises holds that providing rich amounts of foreign language input in a mostly immersive context will lead to higher proficiency in the target language (Dalton-Puffer, 2011; Perez-Canado, 2011). However, some critical voices have been raised lately, see Bruton (2011) and Paran (2013).

This criticism became the driving force for the present study. It investigated CLIL-based vocabulary growth in lower secondary Austrian students. This sample arguably differed from more typical European CLIL contexts concerning student selection and target language contact time. The students in this CLIL research project, for example, worked within a non-selective, sub-optimal (Grandinetti, Langellotti & Ting, 2013), rather low-achieving learning background and a modular CLIL approach. The hallmark of modular CLIL is a sequence of various CLIL projects spread out throughout the school year, each interspersed with mother tongue teaching sequences (Krechel, 2005). Therefore, notwithstanding a certain lack of necessary CLIL criteria according to Tedick & Wesely (2015), this setting of non-selectivity of student population and modular input constitutes an “authentic” European CLIL context (author 1, 2007; Denman, Tanner & de Graaff, 2013; Krechel, 2005).

CLIL studies repeatedly report visible growth in areas such as receptive and productive vocabulary (Dalton-Puffer, 2011), and CLIL proponents have pointed out that such growth can be expected to happen even after a comparably short time of exposure within an immersive or incidental-language-learning-approach (Pérez-Cañado, 2011). This approach has had a marked influence on CLIL and foreign language pedagogy (Ellis & Shintani, 2013). Llinares & Whittaker (2009, p. 189), for example, maintain that “in most courses run by content teachers only, a foreign language is only used as a vehicle for learning content with the assumption that this will lead the student to learn the language naturally and incidentally”.

However, there seem to be three challenges to such a view. First, recent theoretical considerations in second language acquisition (SLA) research shed serious doubts on a simplistic relationship between input and language learning. Ortega (2015, p. 259), for example, considers input as only one of several ingredients of SLA to be “necessary but not sufficient, and perhaps not even the most crucial one”. Second, positive evidence for vocabulary growth in CLIL appears to be particular to studies coming from environments where access to CLIL depends on school-based selection procedures, such as language proficiency, parental background and school achievement (Bruton, 2011; Dalton-Puffer, 2011; Küppers & Trautmann, 2013; Rumlich, 2013). Third, there is a growing body of evidence on the limited and possibly non-optimal effect of incidental language growth in instructed settings in general (Laufer & Nation, 2012; Leow, 2015; Lyster, 2013; Nation, 2011).

All in all, vocabulary teaching and learning in CLIL is still strongly affected by beliefs in the effectiveness of such a language bath metaphor (Hüttner, Dalton-Puffer, & Smit, 2013)¹, even though Vollmer (2010, p. 50) states that there is “a paucity of representative and empirically valid studies concerning the strengths of CLIL students and thus a lack of evidence concerning the central assumptions about the benefits and the superiority of CLIL programmes”. This sentiment is also supported by Bonnet & Dalton-Puffer (2013) and not improved by the even higher scarcity of research into CLIL for low-achieving populations (Denman, Tanner & de Graaff, 2013; Grandinetti, Langellotti, & Ting, 2013; Schwab, 2013).

Finally, we need to point out that we are fully aware of the long tradition of bilingual and immersion programmes in many different parts of the world (Genesee, Lindholm-Leary, Saunders & Christian, 2006; Tedick & Wesely, 2015), but focusing on mostly European CLIL studies is, in our opinion, due to important differences between CLIL and immersion programmes and the particular research aim and context of our study (Dal-

1 Interestingly, some major CLIL methodology books (Coyle, Hood & Marsh, 2010; Dale & Tanner, 2012; Deller & Price, 2007; Mehisto, Frigols & Marsh, 2008) offer sets of activities and didactic advice going beyond this naturalistic and immersive approach.

ton-Puffer, Llinares, Lorenzo, & Nikula, 2014; Lasagabaster & Sierra, 2010). In Austria, for example, CLIL teaching is guided by a highly flexible legal context, which allows for schools to set up locally appropriate and tailor-made programs. Basically, these can range from short, project-based modules, to one-year courses in which English is used as a means of instruction for one or more subjects. The modular program, as described above, has turned into a very popular CLIL approach in lower secondary and primary education in Austria Replace with (Gierlinger, 2007).

To get a solid basis for our research design we reviewed 15 studies of quantitatively measured vocabulary growth in European CLIL classes. These studies, at first sight, unanimously show advantages in vocabulary growth for CLIL. However, a closer look revealed various caveats, and the purported advantages need therefore to be interpreted with caution.

For a start, since CLIL classes in these studies normally received extra language support, possible input frequency effects through additional exposure need to be taken into account (Dalton-Puffer, 2011; Jimenez Catalan & Ruiz de Zarobe, 2009). Second, reporting on absolute vocabulary gain can be misleading. When Pietilä & Merikivi (2014), for instance, described an advantage of CLIL classes with regard to absolute vocabulary growth, they failed to point out that the non-CLIL learners were eventually rapidly catching up, showing a far better relative gain. Third, similar to the above, vocabulary gain may need to be investigated more carefully in terms of its relative growth in comparison to control groups. As Mewald, Prenner, and Sprenger (2004, p. 12) report, significant differences between CLIL and control groups, as assessed in year six, were in fact levelling out by year eight. Therefore, their initially proposed *Scherenhypothese*² had to be discarded. A similar phenomenon was reported in Admiraal, Westhoff, and de Bot (2006), where vocabulary scores ceased to increase after four years of CLIL instruction. And fourth, effects attributed to CLIL exposure might actually be mediated by external factors. Sylven (2007), for example, pointed out that the CLIL-induced advantages they found, were in fact co-determined by extra-mural factors. And finally, there is the ongo-

2 This hypothesis predicts a widening gap between low and high achievers over a longer period of time through CLIL exposure.

ing methodological issue of finding both an appropriate tool and an appropriate design in order to measure vocabulary growth longitudinally (Schmitt, 2010; Dóczy & Kormos, 2016).

In light of the current state of affairs, there appears to prevail a certain ambiguity of empirical evidence on the superiority of CLIL-induced vocabulary learning, as, for example, voiced by Vollmer five years ago. This paper tries to address this gap by providing empirical data from a project with Austrian lower secondary school students. As already mentioned above, CLIL in these schools was carried out in a modular project format, in non-selective classes, and through mostly implicit language instruction (Ellis, et al., 2009, Gierlinger, 2015). We formulated two research hypotheses for this study.

After about six months of target language exposure through modular CLIL, CLIL learners will outperform non-CLIL learners with respect to their receptive vocabulary knowledge as measured in relative gains. In other words, CLIL learners will show a significantly higher receptive vocabulary growth in their post-testings.

A possible superiority of CLIL-induced receptive vocabulary growth will be, apart from the CLIL intervention, co-determined by extra-mural factors.

METHOD

Context and design of the study

A data-driven mixed-methods study was carried out between 2010 and 2011, combining both naturalistic qualitative and manipulated quantitative classroom data. Following a quasi-experimental non-randomised pre/post-test control-group-design, vocabulary test scores were taken twice from all students in order to quantify their vocabulary size before and after the instructional intervention. The instructional intervention was the exposure to CLIL teaching. In our setting, CLIL teaching was exclusively done through modular projects. Our CLIL teachers carried out around 5-7 CLIL projects extending for up to 4 weeks each throughout the school year. The overall contact time resulted in either 60 or 80 additional hours of CLIL teaching. The different figures are the result of one class being exposed to CLIL in two subjects and different project lengths.

We are acutely aware that such a research design could be challenged because of conflating variables, namely methodology and language input. However, CLIL (through English) in a European classroom almost necessarily entails additional exposure to the foreign language (Dalton-Puffer, 2011). Therefore, the teaching method (CLIL) and this additional exposure inevitably constitute conflating variables. Despite this dilemma, European CLIL research seems to have accepted this as an intrinsic design problem. In authentic educational settings, a clear-cut separation of these two variables can arguably and regrettably not be modelled as an experimental condition for English.

The following table summarises the instructional and learning environment of all five classes.

Table 1. Instructional and learning environment of the CLIL and non-CLIL classes

	Classes				
	CLIL 1	CLIL 2	Control 1	Control 2	Control 3
Students	21	21	16	24	16
English classes (hs / week)	3	3	3	3	3
Students' age	13-14	13-14	13-14	13-14	13-14
Course book (CB)	Subject CB	Subject CB	Subject CB	Subject CB	Subject CB
CLIL contact time (hours)	ca 60	ca 80	0	0	0
CLIL subjects	Geography	Chemistry History			
CLIL-language instruction	Predominantly implicit; modular projects; bilingual materials				

While the control group had received no extra language input, the students from the CLIL group had gone, depending on the number of CLIL modules throughout the school year, through either 60 or 80 hours of extra CLIL class time within the treatment period. This happened as part of the schools' language enrichment policy. As far as CLIL methodology was concerned, most of the teaching was held in English, and there was hardly any pre-planned and systematic language-focused work. Teachers' language interventions were predominantly reserved for quick content knowledge clarifications which also resulted in some code switching. The

following quote by a CLIL teacher seems to be representative of the language teaching policies: “Of course, students have to learn technical terms but that is not any focused vocabulary work, it is just the German translation so that one knows it when one needs it” (Gierlinger, 2015). Students were encouraged to speak English, but code switching was not strictly forbidden. CLIL class 1 used English subject course books, CLIL class 2 worked with English materials provided by the teachers. These materials were marginally but not systematically enhanced, such as providing translations or short definitions. The teacher in CLIL class 1 was a language and subject specialist, whereas the teachers in CLIL class 2 were only subject specialists.

Participants

In our study, 87 students from four different Austrian lower secondary schools, 45 boys and 42 girls ($M_{age} = 13.79$ years, $sd = 0.59$, range = 12-14.5 years) took the standardised vocabulary test twice ($t_1 =$ November 2010, $t_2 =$ May 2011). Such an interval may appear short for vocabulary growth in quasi-immersive settings, but other studies, such as Grandinetti, Langellotti, and Ting (2013), worked with even shorter intervals and less input. What is more, our control-group-design was geared towards tracing even minute growth effects across this comparably brief time-span.

Our sample consisted of an experimental group (two CLIL classes, $n = 39$) and the control group (three regular classes, $n = 48$). One class attended two CLIL subjects (chemistry and history). The CLIL students were not preselected but formed part of a whole-class and mixed-ability strategy within the school’s overall language policy.

The students’ mother tongues included, apart from German (83%), Albanian, Arabic, Mandarin Chinese, Romanian, Serbo-Croat-Bosnian (SCB), Tagalog, and Turkish. The CLIL students participated in this study as part of their school-wide CLIL enrichment project; the three classes that served as the control group were recruited in order to match the experimental group for type of school, students’ L1, age group, exposure to regular English classes, their English textbook, as well as the general communicative language teaching approach.

Materials and procedure

Three different tools were used for data elicitation. First, the effect of the instructional intervention (CLIL teaching) was assessed by measuring vocabulary size before and after exposure. For vocabulary measurement, the standardised and computer based vocabulary size test *X-Lex The Swansea Levels Test* (Meara & Milton, 2003) was chosen. *X-Lex* measures vocabulary size by prompting students to rate 120 English words from several vocabulary frequency bands as either known or not known (including nonce-words as distractors). From these ratings, a test score is calculated which reflects a student's vocabulary breadth. The rationale for this choice was manifold. First, *X-Lex* has already been recommended for vocabulary research in CLIL (Canga Alonso, 2013). Second, measuring receptive vocabulary proficiency, as tested by *X-Lex*, is apparently strongly related to word learning by incidental exposure, which is typical of CLIL environments (Jimenez Catalan & Ruiz de Zarobe, 2009, p. 84). Third, *X-Lex* has been standardised and validated, without resorting to one particular norming group, though, for English as a second language speakers in a number of studies (Huibregtse, Admiraal & Meara, 2002; Mochida & Harrington, 2006). Fourth, it measures vocabulary size against pre-defined corpora of different word frequencies, thus tracing growth constrained by word frequency. Fifth, students often react positively to computer-based applications, much more so than to paper-and-pencil designs. Sixth, from a pragmatic view, the schools only allowed for short periods of testing time, which in turn ruled out a more comprehensive assessment tool. Finally, the software automatically produced an output file that was easily saved and fed into spreadsheets and statistical software; this, in turn, ruled out well-known problems with computerisation procedures and the treatment of missing values.

In order to examine the lexical variety of teacher input (more than 11 hours of videoed classroom observation), we ran frequency analyses of the teachers' spoken input using *VocabProfile*. This software is part of the *New General Service List* and the *New Academic Word List* (Browne, Culligan, & Phillips, 2013) and was adapted for online application by Tom Cobb from the Université du Québec in Montreal, Canada (Cobb, 2014). *VocabProfile* performs lexical text analyses by dividing a given corpus into several categories by frequency. These include *k*₁, the most frequent

1.000 words of English, k_2 , the second most frequent thousand words of English, up to k_{25} (based on the *British National Corpus* - BNC20 as the reference corpus), as well as academic words of English and a residual category. It thereby assesses the proportions of low and high frequency vocabulary, indicating lexical variety. In addition to that, the software returns standard lexical statistics, such as type-token ratios of the corpus. The general reliability of this tool was assessed, among others, in studies by Meara and Fitzpatrick (2000) as well as Cobb and Horst (2001). In our study, the teachers' videoed input was transcribed and then fed as a text-file into the software.

The third methodological tool was a background questionnaire. Complementing the assessment of vocabulary breadth through *X-Lex*, all participants from the experimental group filled out a background questionnaire at the time of the first measurement in a paper-and-pencil fashion. The questionnaire explored extra-mural English-related activities along with bio-data such as gender, age, family background, school grades in the CLIL subjects, and self-assessed proficiency in the four skills listening, reading, writing, and speaking.

RESULTS

The analyses of the vocabulary test first focussed on the overall as well as k_1 scores (1,000 most frequent words of English) at t_1 and t_2 .³ First, it was checked that assumptions of normality within the data were met and that there was no over-homogeneity within the variances. Normality was first checked through the inspection of Q-Q-plots. Moreover, both Shapiro-Wilk tests and Anderson-Darling tests confirmed that there was no significant departure from normality (p -values of both groups and both measurements $> .05$). Variances across both groups were checked through Bartlett's tests of sphericity (all p -values $> .05$).

A first inspection of the two test scores and a visual display shows that both groups exhibit remarkable similarities (see Table 2, Figure 1).

3 The raw data from the vocabulary test and the context questionnaire contained a couple of missing values due to technical or organizational problems. Thus, the number of cases that were going into the analyses varies slightly.

Table 2. T-Test Results for the Groups' overall X-Lex Vocabulary Test Scores at t1 and t2

		CLIL group			control group			95% CI			
		M	SD	n	M	SD	n	M	d	t	df
t-test	t1	2758.97	746.56	39	2516.67	650.80	48	-0.09 - 0.79	.35	1.59	76
	t2	2842.86	773.19	35	2749.32	725.11	44	-0.33 - 0.58	.13	0.55	71

Note. M = mean, SD = standard deviation, d = Cohen's d.

Figure 1. Boxplots and interaction plot for the X-Lex scores at t1 and t2 by group

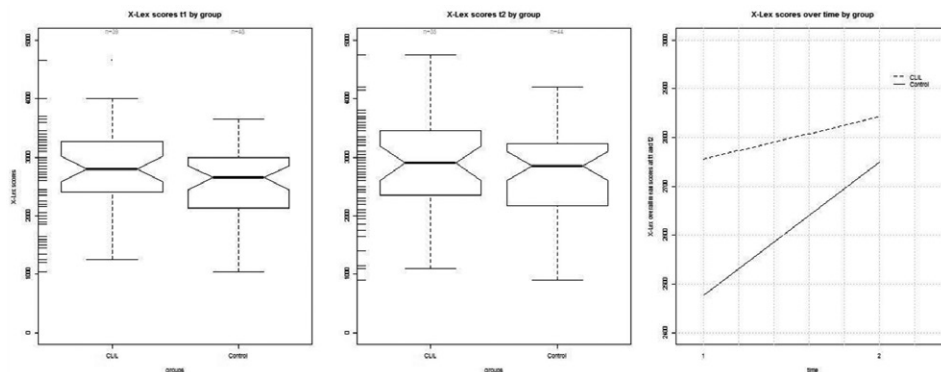
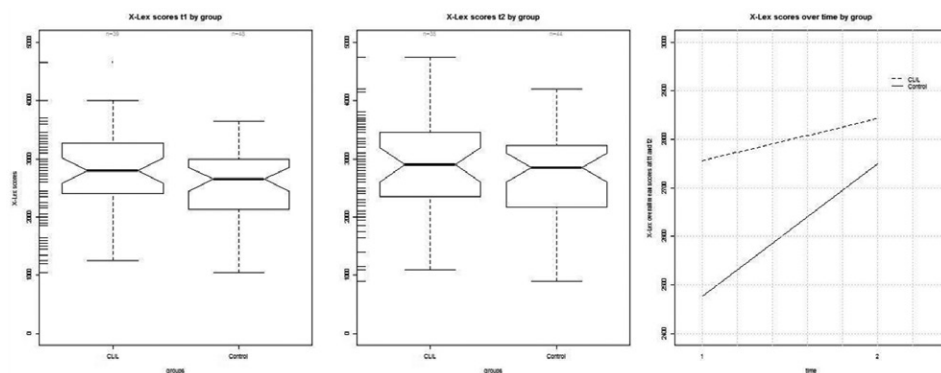


Figure 2. Boxplots and interaction plot for the X-Lex scores at t1 and t2 by group



What we can see in Table 2 and the two boxplots to the left of Figure 1 is that the groups' results do not differ much. CLIL students show the highest scores at both t_1 and t_2 , as can be seen in the length of the upper whiskers in the plot. Both groups show a considerable range and variability in test scores (all *SDs* above 650 points). The proximity of the medians in the boxplots and the overlap of the box notches (quasi confidence intervals) illustrate that group medians do not seem to differ significantly, neither at t_1 nor at t_2 . Independent-samples *t*-tests confirmed that the CLIL group had a slightly but not significantly⁴ better start at t_1 , with a small standardised effect-size of $d = 0.35$, and behaved similarly to the control group at t_2 , with a negligible effect size of $d = 0.13$.

The interaction plot on the right-hand side of Figure 1 illustrates a mild increase in test scores over time for the CLIL group as well as a pronounced increase for the control. A one-way repeated measure ANOVA over group interacting with time revealed that there was no significant main effect for the CLIL treatment ($F(1, 77) = 1.72, p = .19$), nor for the interaction over time ($F(1, 77) = 1.22, p = .27$). In order to check for a regression effect, the vocabulary scores at t_2 were modelled using an OLS regression with the test scores at t_1 and the grouping factor as predictors. However, even in such a model with the test scores at t_1 held constant ($F(11.9), df = 2;76, p < .001, R^2 = 22\%$), the treatment effect was still insignificant ($\beta = 49.38, t(151.67) = 0.33, p = .75$) and had a negligible magnitude of $\eta_p^4 = 0.0014$. Thus, coming back to hypothesis 1, it is not the CLIL exposure over time that predicts the vocabulary gain between t_1 and t_2 .

So far, these results are somewhat at odds with predictions made by many CLIL proponents. While we can see vocabulary growth in both groups, and while the CLIL group outperforms the control group in terms of absolute test scores, the relative gain of the control group exceeds the CLIL pupils by far. In order to investigate such idiosyncratic behaviour, two possible explanations will be explored in the following. The first one relates to frequency effects as well as the interaction of the vocabulary input students received and justifiable expectations about vocabulary growth

4 All *p*-values in this paper refer to a significance level of .05.

based on this specific input.⁵ The second one focuses on extra-mural influences (Sylven, 2007; Sylven, 2013), which might prove to be co-determining factors in our design.

Frequency effects of vocabulary input

Frequency is considered a key determinant of language acquisition, and higher frequency forms in the input are predicted to enable earlier automatization (Milton, 2009; Ortega, 2015). In order to find out if and to which extent vocabulary input by the CLIL teachers could have stimulated vocabulary growth in the CLIL group, teachers' input from three different subjects across 11 hours of class time was submitted to frequency analyses using *VocabProfile*. The video data were first transcribed, computerised, and then stripped off all proper names, since those would have skewed the true nature of the corpus' size (Milton, 2009). The three clean corpora were then fed into the software, which segmented them into frequency bands of the first (k_1), second (k_2), and third thousand (k_3) most frequent words, academic words (AWL), and a residual category called *off-list*. The profiler produced the frequency counts as illustrated in Table 3.

The three different subjects (chemistry, geography, history) show a surprisingly similar picture. First, there is the relatively high percentage (88.11 – 91.45%) of utterances belonging to the 1,000 most frequent English words (k_1). This and the low Guirard index of 7.54 – 8.65 indicate the repetitive nature of high-frequency words in the CLIL teachers' classroom language (Milton, 2009). While such a prevalence of basic lexis might well be pedagogically justified with respect to the entrenchment of English *BICS* vocabulary (Ellis, 2013), the low proportion of k_2 to k_3 words suggests that more advanced and broader word learning, especially with respect to the more closer frequency bands, may not have been fostered through this kind of linguistic input in our CLIL groups.

However, two reservations need to be raised immediately. First, following White (2013), it is still unclear what the optimal ratio of unknown

5 Note that a test which is considered to be appropriate for assessing receptive vocabulary in EFL as a subject may not be appropriate to assess EFL as a vehicular language (Jimenez Catalan & Ruiz de Zarobe, 2009).

Table 3. Corpus Analysis of Teachers Vocabulary Input from three CLIL Classes

	Chemistry 4h class time, 153 minutes coverage	Geography 2h class time, 92 minutes coverage	History 4h class time, 183 minutes coverage
k1 words	88.11%	91.45%	90.60%
k2 words	3.95%	4.69%	3.91%
k3 words	2.62%	1.35%	1.09%
AWL words	1.57%	0.49%	0.55%
off-list words	3.76%	2.02%	3.85%
tokens	7981	3473	7502
types	763	510	653
type-token ratio	.10	.15	.09
Guiraud index	8.54	8.65	7.54
tokens per type	10.46	6.81	11.49

Note. k1 = the first 1,000 most frequent words of English, k2 = 1,001 – 2,000 most frequent words of English, k3 = 2,001 – 3000 most frequent words of English, AWL = academic word list.

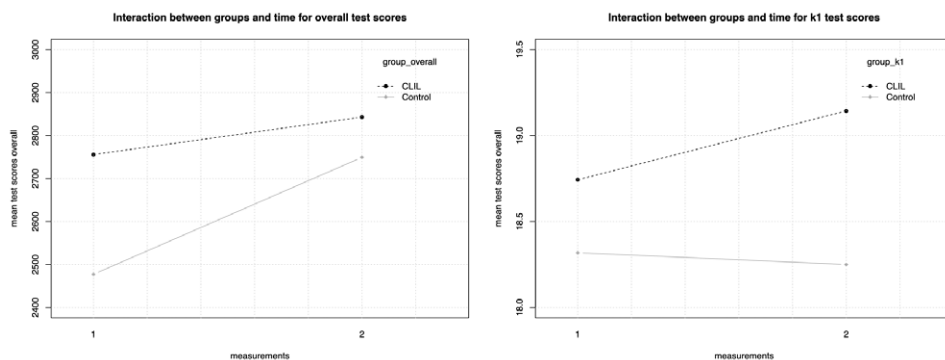
vocabulary items to the word total in a text would be; in other words, how many unknown words does a teacher's input need to exhibit in order to provide stimulating but comprehensible input? This reasoning does not refer to the well-known debate around the minimum vocabulary knowledge required to understand authentic texts in reading comprehension tasks (Hu & Nation, 2002; Nation, 2006). While there seems to be a general consensus that this minimal proportion of known words ranges around 95%, there is much less agreement as to the minimal proportion of known words in teachers' input in order to be both comprehensible and stimulating.

Second, a closer look at the *off-lists* in our research data raises doubts as to how and whether the corpus underlying this analysis reflects the desired subject learning in CLIL. For example, various subject-relevant words were relegated to this obscure category. In other words, the underlying algorithm *VocabProfile* employs may not align with the instructional goals of CLIL, as a considerable number of the subject-specific words are very important for a full understanding of the CLIL subject. In this respect, Hyland

and Tse (2007) call for more discipline-specific studies of vocabulary use in academic setting. Nevertheless, according to teachers' input analysis and the power of frequency effects (Ellis, 2013), vocabulary growth should have happened at least within the 1,000 most frequent English words, since those words featured prominently in the input corpora.

If we now take those insights from our corpus analysis and re-examine our quantitative vocabulary scores within the *k1* band, we get a rather different picture. Consider Figure 3 now, which contrasts the overall test scores with the *k1* scores (1,000 most frequent words).

Figure 3. Interactions plots for mean test scores at both measurements for overall (left) and *k1* (right) results



While the left panel of Figure 3 illustrates a pronounced increase in vocabulary growth for the control group, the right-hand-side panel shows that, within the 1,000 most frequent words of English, only the CLIL group benefits. When *k1* test scores at *t1* were centred ($M = 0, SD = 1$) and controlled for, a regression model over *t2* test scores ($F(3, 75) = 23.79, p < .001$, adjusted $R^2 = 47\%$) showed that this CLIL effect over time is significant ($\beta = 0.44, t(0.19) = 2.30, p = .024$). In other words, CLIL exposure is effective among the 1,000 most frequent words when measured against a standardised and constant *t1*-value.

Coming back to hypothesis 2, our results suggest that, within the *k1* vocabulary band, vocabulary development was co-determined by CLIL exposure. Let us now have a look at the influence of background variables as determiners for receptive vocabulary growth.

The influence of background variables

As Sylven (2007; 2013) & Pietilä and Merikivi (2014) pointed out, extra-mural factors might co-determine CLIL-induced vocabulary growth. Consequently, we also explored possible background variables. These included pupils' sex and age, the families' education level, if they had been on a stay abroad or not, their grade in the CLIL subject (geography), their English proficiency (amalgamation of school grades and self-assessment in the four skills), their English activities outside school as well as the grouping factor CLIL vs. control. After these dimensions of our questionnaire were pooled using principal component analyses, the vocabulary scores at t_2 were examined in regression models with eight predictors. After all predictors were checked for variance inflating factors and collinearity, a step-wise linear regression revealed that in the final model ($F(2,69) = 13.67, p < .001, R^2$ (adjusted) = 0.26) only pupils' English proficiency ($\beta = -344.64, t(74.64) = -4.62, p < .001, \eta_p^2 = 0.24$) had a significant and substantial partial effect. The β -coefficient in this model was negative, because a lower value for this predictor (fed into the model as the scores from the principal component analysis), corresponded to a high proficiency level. A complementing classification analysis confirmed that only pupils' English proficiency level predicted the second test scores. Coming back to hypothesis 2, extra-mural factors outside the CLIL setting did not play a significant role in our data.

DISCUSSION

Vocabulary growth is one of CLIL's major language learning driving forces (Bonnet & Dalton-Puffer, 2013). Thus, the aim of this study was to revisit language growth in CLIL classrooms to find out whether new production data match concepts such as frequency effects in teachers' input (Ellis, 2013) and extra-mural factors (Sylven, 2007; 2013).

Let us now discuss the major two findings from our study. First, CLIL students fail to outperform the controls in terms of overall receptive vocabulary growth. However, the frequency analyses of teachers' input revealed that CLIL exposure actually centred mainly on the 1,000 most frequent

words of English (k_1). And this is probably the reason why it was only within this band that we found significant vocabulary growth for CLIL students.

We can think of two possible explanations for this unorthodox result. The first one relates to the power of frequency effects. Since CLIL students were vastly more exposed to vocabulary from the k_1 band, deeper learning and entrenchment was to be expected (Ellis, 2013). The reasons for such a high occurrence of k_1 words may lie in the particular pedagogical context of CLIL, in which subject-specific content comprehension and clarification are considered to be of utmost importance by the teachers (Gierlinger, 2007; 2015; Hüttner et al., 2013; Llurda & Lasagabaster, 2010; Nikula, 2010). And one of the strategies to reach this aim is explaining and elaborating on subject-specific concepts through basic, high-frequency vocabulary. The high type-token ratio plus the high coverage of the k_1 band in our CLIL teachers' input suggests a deliberate effort towards content comprehensibility and clarification. Other research by Nation & Webb (2011) points out that by keeping the vocabulary load lower and increasing its repetitions, the amount of vocabulary learnt will increase⁶.

Second, the CLIL specific vocabulary growth may reside more significantly in the area of subject specific vocabulary, which was not covered by the testing tool. However, this raises the question why the use of subject specific vocabulary apparently only had a negligible priming effect (Hoey, Mahlberg, Stubbs, & Teubert, 2007) on academic and general vocabulary? In other words, one would have expected a much more pronounced effect between academic and subject specific vocabulary within the subject classroom discourses. Arguably, this linguistic puzzle may result from a more general lack of academic language use at this age level and within the context of a broad spectrum of learner achievements. Comparative research between the use of academic language in CLIL classes and in mainstream classes could shed more light on this issue.

6 Arguably, such teaching techniques would not be applied as systematically and frequently in a foreign language learning classroom setting. The foreign language curriculum typically expects an incremental vocabulary progression along pre-defined proficiency levels (CEFR, Council of Europe, 2001; Milton, 2009). Or, as Decoo (2010) contended, foreign language learning is a progressive endeavor where the learner should advance from one point to another, constantly improving.

Third, as suggested by Zydatiss (2012, pp. 27-28), visible receptive vocabulary growth within CLIL may only be expected after a certain critical mass of treatment exposure. Thus, a period of 5-6 months of project-based exposure of CLIL might simply fail to reach such a critical mass and thereby prove less effective. This critical mass phenomenon may be further aggravated by an implicit teaching approach, which may have a tendency to delay noticing and hence language learning (Svalberg, 2007; Williams, 2013). A possible language threshold for CLIL is also tentatively pointed out by Agustin Llach (2014) in her research on primary CLIL. Summing up, frequency and noticing effects may play a vital role in CLIL vocabulary growth.

The other main finding of our study pertains to the role of extra-mural factors. While Sylven (2007) found that extra-mural factors did play a significant role in her study, in our data only learners' proficiency level in English predicted the final vocabulary results. Notice, however, that Sylven (2013), in a theoretical article, related her research outcomes to the extraordinary linguistic situation of Sweden, where English, in her words, is "omnipresent" (p. 310).

These findings raise at least three more issues. First, the question remains whether richer teacher input may not result in a broader vocabulary gain at least for more advanced learners. Second, would the results have turned out to be the same in a less immersive, more instructed and vocabulary-focused teaching and learning context? The massive amount of recent SLA literature pertaining to the important role of language awareness, the noticing hypothesis, and explicit knowledge for language learning (Bot, Lowie, & Verspoor, 2006; Ellis, 2015; Ellis & Shintani, 2013; Leow, 2015; Williams, 2013) suggest that these issues need to be addressed by future CLIL research. Third, on a more general level, the overall heterogeneity of CLIL contexts and implementations makes it dangerous to jump to foregone conclusions, or, as Bonnet and Dalton-Puffer (2013, p. 279) put it, "CLIL is itself subject to existing teaching cultures rather than an omnipotent agent of systemic change".

To sum up, our results remain puzzling but maybe also pioneering for the moment. Although we believe, we can trace parts of these issues to methodological design problems that come with *X-Lex's* – and other vocabulary tools' – difficulty to deal with subject-essential CLIL vocabulary.

Notwithstanding these issues, as implications of our research we propose some tentative recommendations for further research and CLIL practice. Given the importance of vocabulary growth in CLIL, there is ample room for further research into the development of CLIL-induced receptive vocabulary development over time. We believe that researching CLIL's potential over longer periods, together with a careful description of the methodological instantiations, will reveal a more realistic picture of the effect of CLIL on the learning of subject and language content. In addition to this, we need more studies on learner and teacher vocabulary with respect to frequency and typology (general, academic, and technical) but also its relationships to CLIL methodologies, ranging from (totally) immersive to (more) form focussed approaches. Our data suggest that the mainstream *language bath* CLIL metaphor needs to be complemented by more deliberate and form-focused instructional approaches (Grandinetti et al., 2013; Lyster, 2013; Nation, 2011).

Finally, our research turned out to be intrinsically complex, because it studied the development of a complex phenomenon (vocabulary growth) in complex ecologies (classroom learning) among a multi-variant dynamic population (school learners, teachers). Controlling quasi-experimental conditions in such a setting appears challenging. Such factors can make it extremely difficult to adopt traditionally formulated, linearly framed research methods. By applying a longitudinal and mixed method approach, we tried to go beyond a popular but possibly too simplistic comparison of CLIL and non-CLIL outcomes only. Although the majority of these comparative studies paints a positive picture with respect to language growth in CLIL, the results of our study prove to be much less straightforward and point towards a complex set of factors influencing language growth in CLIL. All in all, the benefits of our study lie in its explorative and critically reflective nature. Despite these constraints, we hope that our results prove to be sufficiently interesting to merit further investigations.

ACKNOWLEDGEMENTS

This research was supported by a grant from the University College of Education Upper-Austria.

REFERENCES

- Admiraal, W., Westhoff, G. & de Bot, K. (2006). Evaluation of bilingual secondary education in the Netherlands: Students' language proficiency in English. *Educational Research and Evaluation*, 12, 75-93.
- Agustin Llach, M. D. P. (2014). Exploring the lexical profile of young CLIL learners: Towards an improvement in lexical use. *Journal of Immersion and Content-Based Language Education*, 2(1), 53-73.
- Aiken, L. & West, S. (1991). *Multiple regression: Testing and interpreting interactions*. Thousand Oaks, CA: Sage.
- Bonnet, A. & Dalton-Puffer, C. (2013). Great expectations? Competence and standard related questions concerning CLIL moving into mainstream. In S. Breidbach & B. Viebrock (Eds.), *Content and language integrated learning (CLIL) in Europe* (pp. 269-284). Frankfurt: Peter Lang.
- Bot, K. D., Lowie, W. & Verspoor, M. (2006). *Second language acquisition*. New York: Routledge.
- Browne, C., Culligan, B. & Phillips, J. (2014, October 1). Retrieved from New General Service List: <http://www.newgeneralservicelist.org/>
- Bruton, A. (2011). Are the differences between CLIL and non-CLIL groups in Andalusia due to CLIL? A reply to Lorenzo, Casal and Moore (2010). *Applied Linguistics*, 32(2), 236-241.
- Burger, S. & Weinberg, A. (2014). Three factors in vocabulary acquisition in a university French immersion adjunct context. *Journal of Immersion and Content-Based Language Education*, 2(1), pp. 23-52.
- Canga Alonso, A. (2013). The receptive vocabulary of Spanish 6th-grade primary-school students in CLIL instruction: A preliminary study. *LACLIL*, 6(2), 22-41.
- Cenoz, J., Genesee, F. & Gorter, D. (2014). Critical analysis of CLIL: Taking stock and looking forward. *Applied Linguistics*, 35(3), 243-262.
- Cobb, T. (2016, January 19). *The Compleat Web VP*. Retrieved from Compleat Lexical Tutor: <http://www.lextutor.ca/vp/comp/>

- Cobb, T. & Horst, M. (2001). Growing academic vocabulary with a collaborative online database. In B. Morrison, D. Gardner, K. Koebke, & M. Spratt (Eds.), *LT Perspectives on IT & Multimedia* (pp. 189-226). Hong Kong: Polytechnic University Press.
- Costa, F. & D'Angelo, L. (2011). CLIL: A suit for all seasons. *LACLIL*, 4(1), 1-13.
- Council of Europe. (2001). *Common european framework of reference for languages: Learning, teaching, assessment*. Cambridge: CUP / Council of Europe.
- Coyle, D., Hood, P. & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge: Cambridge University Press.
- Cummins, J. (2008). BICS and CALP: Empirical and theoretical status of the distinction. In B. Street & N. Hornberger (Eds.), *Encyclopedia of language and education* (pp. 71-83).
- Dale, L. & Tanner, R. (2012). *CLIL Activities with CD-ROM: A resource for subject and language teachers*. Cambridge: Cambridge University Press.
- Dalton-Puffer, C. (2011). Content-and-language integrated learning: From practice to principles. *Annual Review of Applied Linguistics*, 31, 182-204.
- Dalton-Puffer, C., Llinares, A., Lorenzo, F. & Nikula, T. (2014). You can stand under my umbrella": Immersion, CLIL and bilingual education. A response to Cenoz, Genesee & Gorter (2013). *Applied Linguistics*, 35(2), 213-218.
- Decoo, W. (2010). *Systemization in foreign language teaching: Monitoring content progression*. New York: Routledge.
- Deller, S. & Price, C. (2007). *Teaching other subjects through English*. Oxford: Oxford University Press.
- Denman, J., Tanner, R. & de Graaff, R. (2013). CLIL in junior vocational secondary education: challenges and opportunities for teaching and learning. *International Journal of Bilingual Education and Bilingualism*, 16(3), 285-300.
- Doczi, B. & Kormos, J. (2016). *Longitudinal developments in vocabulary knowledge and lexical organization*. New York, NY: Oxford University Press.
- Ellis, N. C. (2015). Implicit and explicit language learning: Their dynamic interface and complexity. In P. Rebuschat (Ed.), *Implicit and explicit*

- learning of languages* (pp. 1-24). Amsterdam: John Benjamins Publishing Company.
- Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J. & Reinders, H. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching*. Bristol: Multilingual Matters.
- Ellis, R. & Shintani, N. (2013). *Exploring language pedagogy through second language acquisition Research*. New York: Routledge.
- Genesee, F., Lindholm-Leary, K., Saunders, B. & Christian, D. (2006). *Educating English language learners: A synthesis of research evidence*. Cambridge: Cambridge University Press.
- Gierlinger, E. M. (2007). Modular CLIL in lower secondary education: some insights from a research project in Austria. In C. Dalton-Puffer & U. Smit (Eds.), *Empirical perspectives on classroom discourse* (pp. 79-118). Frankfurt: Peter Lang.
- Gierlinger, E. M. (2015). 'You can speak German, sir': on the complexity of teachers' L1 use in CLIL. *Language and Education*, 29(4), 347-368. doi:10.1080/09500782.2015.1023733
- Grandinetti, M., Langellotti, M. & Ting, T. Y. (2013). How CLIL can provide a pragmatic means to renovate science education – even in a sub-optimally bilingual context. *International Journal of Bilingual Education and Bilingualism*, 16(3), 267-284.
- Guirard, P. (1960). *Problèmes et Méthodes de la Statistique Linguistique*. Paris: Presses universitaires de France.
- Hoey, M., Mahlberg, M., Stubbs, M., & Teubert, W. (2007). *Text, Discourse and Corpora*. London: Continuum.
- Huibregtse, I., Admiraal, W. & Meara, P. (2002). Scores on a yes/no vocabulary test: Correction for guessing and response style. *Language Testing*, (19), pp. 227-245.
- Hüttner, J., Dalton-Puffer, C. & Smit, U. (2013). The power of beliefs: Lay theories and their influence on the implementation of CLIL programmes. *International Journal of Bilingual Education and Bilingualism*, 16(3), 267-284.

- Hyland, K. & Tse, P. (2007). Is there an “academic vocabulary”? *TESOL Quarterly*, 41(2), pp. 235-253.
- Jimenez Catalan, R., Maria, & Ruiz de Zarobe, Y. (2009). The receptive vocabulary of EFL learners in two instructional contexts: CLIL versus non-CLIL instruction. In R. Jimenez Catalan, Maria & Y. Ruiz de Zarobe (Eds.), *Content and language integrated learning: Evidence from research in Europe* (pp. 81-92). Bristol: Multilingual Matters.
- Krechel, H.-L. (2005). Situation des mehrsprachigen Unterrichts und der Lehrerbildung in Deutschland. In H.-L. Krechel (Ed.), *Mehrsprachiger Fachunterricht in Ländern Europas* (pp. 9-33). Tübingen: Gunter Narr.
- Küppers, A. & Trautmann, M. (2013). It is not CLIL that is a success - CLIL students are! Some critical remarks on the current CLIL boom. In S. Breidbach & B. Viebrock (Eds.), *Content and language integrated learning (CLIL) in Europe* (pp. 285-296). Frankfurt: Peter Lang.
- Lasagabaster, D. & Sierra, J., Manuel. (2010). Immersion and CLIL in English: More differences than similarities. *ELT Journal*, 64(4), 367-374.
- Leow, R. P. (2015). *Explicit learning in the L2 classroom: A student-centered approach*. New York, NY: Routledge.
- Llinares, A. & Whittaker, R. (2009). Teaching and learning history in secondary CLIL classrooms: from speaking to writing. In E. Dafouz & M. C. Guerrini (Eds.), *CLIL across educational levels*. Oxford: Richmond Publishing.
- Llinares, A., Morton, T. & Whittaker, R. (2012). *The Roles of Language in CLIL*. Cambridge University Press.
- Llurda, E. & Lasagabaster, D. (2010). Factors affecting teachers’ beliefs about interculturalism. *International Journal of Applied Linguistics*, 20(3), 327-353.
- Loewen, S. (2011). Focus on form. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 576-592). New York: Routledge.
- Lorenzo, F., Casal, S. & Moore, P. (2009). The effects of content and language integrated learning in European education: Key findings from the Andalusian bilingual sections evaluation project. *Applied Linguistics*, 31(3), 418-442.

- Lorenzo, F., Moore, P. & Casal, S. (2011). The causes, effects, and breadth of content and language integrated learning - a reply to Bruton. *Applied Linguistics*, 32(4), 450-455.
- Lyster, R. (2013). Content-based Language teaching. In P. Robinson (Ed.), *The Routledge Encyclopedia of Second Language Acquisition* (pp. 125-127). New York: Routledge.
- Lyster, R., Quiroga, J., & Ballinger, S. (2013). The effects of biliteracy instruction on morphological awareness. *Journal of Immersion and Content-Based Language Education*, 1(2), 169-197.
- Meara, P. & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, 4(2), pp. 142-154.
- Meara, P. & Fitzpatrick, T. (2000). Lex30: An improved method of assessing productive vocabulary in an L2. *System*, 28(1), pp. 19-30.
- Meara, P. & Milton, J. (2003). *X-Lex, the Swansea Levels Test*. Newbury: Express.
- Mehisto, P., Marsh, D. & Frigols, M. J. (2008). *Uncovering CLIL: Content and language integrated learning in bilingual and multilingual education*. Oxford: Macmillan.
- Mewald, C., Prenner, M., & Sprenger, H. (2004). *Englisch als Arbeitssprache (EAA) auf der Sekundarstufe 1*. BMUK: GZ 20.233/4-VI/A/3/01.
- Milton, J. (2009). *Measuring second language vocabulary acquisition*. Bristol: Multilingual Matters.
- Mochida, A. & Harrington, M. (2006). The Yes-No test as a measure of receptive vocabulary knowledge. *Language Testing*, 23, pp. 73-98.
- Nation, I. S. P. (2011). Second Language Speaking. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning. vol II* (pp. 443-454). New York and London: Routledge.
- Nikula, T. (2010). Effects of CLIL on a teacher's classroom language use. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 105-124). Amsterdam: John Benjamins.
- Ortega, L. (2014). Second language learning explained? SLA across 10 contemporary theories. In B. VanPatten & J. Williams (Eds.),

- Theories in second language acquisition: An introduction* (pp. 245-272). New York: Routledge.
- Paran, A. (2013). Content and language integrated learning: Panacea or policy borrowing myth. *Applied Linguistics Review*, 4(2), 317-342.
- Pérez-Cañado, M. L. (2012). CLIL research in Europe: past, present, and future. *International Journal of Bilingual Education and Bilingualism*, 15(3), 315-341.
- Pérez-Cañado, M. L., Marsh, D. & Padilla, J. R. (2015). *CLIL in action: Voices from the classroom*. Cambridge: Cambridge Scholars Publishing.
- Pietilä, P. & Merikivi, R. (2014). Vocabulary in CLIL and in mainstream education. *Journal of Language Teaching and Research*, 5(3), 487-497.
- Rumlich, D. (2013). Students' general English proficiency prior to CLIL: Empirical evidence for substantial differences between prospective CLIL and non-CLIL students in Germany. In S. Breidbach & B. Viebrock (Eds.), *Content and language integrated learning (CLIL) in Europe: Research perspectives on policy and practice* (pp. 181-202). Frankfurt: Peter Lang.
- Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York: Palgrave MacMillan.
- Schwab., G. (2013). Bili für alle? Ergebnisse und Perspektiven eines Forschungsprojektes zur Einführung bilingualer Module in einer Hauptschule. In S. Breidbach & B. Viebrock (Eds.), *Content and language integrated learning (CLIL) in Europe: Research perspectives on policy and practice*. Frankfurt: Peter Lang.
- Svalberg, A. M.-L. (2007). Language awareness and language learning. *Language Teaching*, 40(04), 287-308.
- Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), *Handbook of research in second language teaching and learning* (pp. 471-483). Mahwah, NJ: Lawrence Erlbaum Associates.
- Sylvén, L., K. (2007). Swedish CLIL students' extracurricular contact with English and its relation to classroom activities. In D. Marsh & D. Wolff (Eds.), *CLIL in diverse contexts - converging goals* (pp. 237-252). Frankfurt: Peter Lang.

- 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.
- Tedick, D. J. & Wesely, P. M. (2015). A review of research on content-based foreign/second language education in US K-12 contexts. *Language, Culture and Curriculum*, 28(1), 25-40.
- Vollmer, H. (2010). Bilingualer Unterricht als Inhalts- und Sprachenlernen. In G. Bach & S. Niemeier (Eds.), *Bilingualer Unterricht: Grundlagen, Methoden, Praxis, Perspektiven* (pp. 47-70). Frankfurt: Lang.
- White, J. (2013). Flooding. In P. Robinson (Ed.), *The Routledge Encyclopaedia of Second Language Acquisition* (pp. 238-239).
- Williams, J., N. (2013). Attention, awareness, and noticing in language processing and learning. In J. M. Bergsleithner & S. N. Frota (Eds.), *Noticing and second language acquisition: Studies in honor of Richard Schmidt* (pp. 39-57). National Foreign Language Resource Center.
- Zydatiś, W. (2012). Linguistic thresholds in the CLIL classroom? The threshold hypothesis revisited. *International CLIL Research Journal*, 1(4), 17-28.