

*Vocabulary development in a CLIL context:
A comparison between French and English L2*

Kristof Baten

Ghent University, Belgium

<https://orcid.org/0000-0003-2125-8011>

kristof.baten@ugent.be

Silke Van Hiel

Ghent University, Belgium

<https://orcid.org/0000-0002-6616-3959>

silkevanhiel@gmail.com

Ludovic De Cuypere

Ghent University, Belgium

Free University of Brussels, Belgium

<https://orcid.org/0000-0002-0050-1097>

ludovic.decuypere@ugent.be

Abstract

Content and language integrated learning (CLIL) has expanded in Europe, favored by the large body of research, often showing positive effects of CLIL on L2 development. However, critical voices have recently questioned whether these positive findings apply to any language, given that most research focuses on English. Taking into account this concern, the present study investigated the (productive and receptive) vocabulary development in L2 English and L2 French of the same group of learners *within* a CLIL context. The aim was not to evaluate the benefits of CLIL over non-CLIL, but, instead, to examine whether vocabulary gains in CLIL learning are language-dependent. More specifically, this study included 75 Flemish eight-grade pupils who had CLIL lessons in both English and French. The results show that although the pupils have a larger English vocabulary, the level of improvement (from pretest to

posttest) is not different across the languages. The findings indicate that within CLIL vocabulary knowledge also develops in languages other than English.

Keywords: CLIL; vocabulary; L2 English; L2 French; productive and receptive levels tests

1. Introduction

Content and language integrated learning (CLIL), that is, the teaching of subjects, such as history or economy, in a foreign language, has gained increasing popularity in the European educational landscape over the past 20 years (EACEA, 2012) and in many other geographical contexts, such as Asia (e.g., Lin, 2016) or South America (e.g., Banegas, 2011). Undoubtedly, this growth in popularity is partly driven by the substantial body of research on the effects of CLIL, including large-scale studies on learning outcomes (e.g., Admiraal, Westhoff, & de Bot, 2006 for the Netherlands; Lasagabaster, 2008 for Spain; Zydaliß, 2007 for Germany) as well as specific studies dealing with individual aspects of language, such as vocabulary, pronunciation and morphosyntax, or the four language skills. The image that emerges from these studies is that CLIL learners generally attain higher proficiency levels than non-CLIL learners, especially in listening skills (Aguilar & Muñoz, 2014) and vocabulary (Jexenflicker & Dalton-Puffer, 2010). Furthermore, CLIL learners are found to display greater fluency and creativity in speaking (Mewald, 2007) and to reach higher levels on CEFR-based diagnostic tests in reading, listening, writing and speaking (Lorenzo, Casal, & Moore, 2010).¹

Despite the positive picture that is commonly associated with CLIL, a number of critical voices have recently raised concerns about the role of English within CLIL. For example, Cenoz, Genesee, and Gorter (2014, p. 257) point out that “much, if not most, research on CLIL has been conducted by ESL/EFL scholars.” Consequently, most of the (positive) research findings are based on the acquisition of English as a second or foreign language. In this respect, Dalton-Puffer, Nikula, and Smit (2010, p. 286) refer, somewhat provocatively, to content and *English* integrated learning instead of content and *language* integrated learning. Similarly, Pérez, Lorenzo, and Pavón (2016, p. 485) speak of an “empirical vacuum” of how CLIL functions in languages other than English, as it is not implausible that the positive findings for CLIL in English are, at least partly, connected to English itself. Cenoz et al. (2014) call for more *inner*-CLIL research, a

¹ In areas such as morphosyntax, writing performance and pronunciation CLIL seems to have little effect (Aguilar & Muñoz, 2014; Dalton-Puffer, 2008; Dalton-Puffer et al., 2010; Ruiz de Zarobe, 2011).

need which was also underlined by Breidbach and Viebock (2012) in their review of recent research on CLIL in Germany. Seeing that there is indeed no one unified CLIL approach (Coyle, 2007; Lasagabaster, 2008; Wolff, 2002), it is worthwhile to get into the black box and examine CLIL also in relation to its variation. CLIL approaches vary according to curriculum variables such as intensity, duration, age of onset, starting linguistic level, the subjects and languages involved (Coyle, 2007), but also according to pedagogic variables such as types of input, types of output practice, and use of strategies (De Graaff, Koopman, Anikina, & Westhoff, 2007). The present study will single out the variable of language, which within CLIL is most commonly a combination of English with another national, regional, border or minority language (Pérez-Cañado, 2012).

Therefore, in an attempt to answer this call for more inner-CLIL research, the present study compares the (receptive and productive) vocabulary development in both L2 English *and* L2 French *within* the same group of CLIL learners. We focus on vocabulary because positive effects of CLIL have especially been observed in this area (Dalton-Puffer, 2008). Moreover, vocabulary knowledge is generally taken to be one of the most salient components of linguistic ability (Hulstijn, 2010). It should be emphasized, however, that this study does not include a non-CLIL group and thus refrains from evaluating the effectiveness of CLIL compared to non-CLIL. Instead, the present study examines to what extent vocabulary development within a specific CLIL context varies according to the target language. With this comparison of English CLIL versus French CLIL, the present investigation seeks to fill a gap in the existing CLIL research, which has largely focused on the *lingua franca* English (Pérez et al., 2016). The goal is to examine how CLIL works for languages that do not have the status of a global language. To this aim, the present study zooms in on Belgium's Dutch-speaking region (i.e., Flanders), where, since recently, CLIL is organized in English and French (and German as a matter of fact). Before the findings are presented, some brief background is given on CLIL in general as well as on the specific context of CLIL in Flanders.

2. Literature review

2.1. Content and language integrated learning

CLIL has become a popular and widespread practice (and term) in Europe (Coyle, Hood, & Marsh, 2010). It refers to an alternative didactic method in which school subjects are taught in a second or foreign language (L2). Crucial to the method is the dual focus on language and content, which implies that the language is not the main (or only) goal, but rather serves as the means of communication in authentic situations. The method is believed to be more effective than

traditional language education on linguistic, subject content, cognitive and affective-attitudinal grounds. As such, it would replicate the positive effects found in the wealth of previous research on content/language integration in Canadian immersion, US bilingual education and European international schools. However, Pérez-Cañado (2012) rightly pointed out that, while positive effects have been attested in these specific North American and European contexts, the possibility of similar effects arising from CLIL still remains largely assumed instead of empirically backed-up. Indeed, even though CLIL represents an emerging field, strong empirical evidence in this area is still scarce (Gierlinger, 2017). Also, the context-specificity of each type of bilingual education shows that extrapolation of findings from one situation to another should be treated with caution. In addition, it should also be taken into account that CLIL approaches vary considerably, which again limits the generalizability of research findings (Coyle, 2007; Lasagabaster, 2008; Wolff, 2002).

Nevertheless, a recurring finding in CLIL research seems to be the positive effect on vocabulary knowledge. For example, in a large-scale study with 180 pupils in Berlin, Zydatiņ (2007) observed that the CLIL learners outperformed the other learners on lexical competence on an English proficiency test (and also on grammatical and communicative competences, for that matter). In writing, Ackerl (2007) and Jexenflicker and Dalton-Puffer (2010) found that Austrian CLIL learners have a larger vocabulary size, use more complex and less frequent words, and show more word variation. Similarly, Catalanian CLIL learners obtained significantly higher levels on lexical complexity in their writing performance than their non-CLIL counterparts (Navés, 2011). In a study with Finnish learners, using the receptive and productive Vocabulary Levels Test (see below), Merikivi and Pietilä (2014) analogously found larger vocabulary sizes in the CLIL-group compared to the non-CLIL group. Receptive scores were higher than productive scores, and both were correlated, meaning that CLIL learners with high receptive vocabulary size also scored high for productive vocabulary size. Furthermore, studies have shown that CLIL pupils rely less on their mother language and increase their level of lexical inventions, which indicates a higher proficiency level (Jiménez Catalàn, Ruiz de Zarobe, & Cenoz, 2006; Ruiz de Zarobe, 2010).

The observed lexical advantage of CLIL learners is attributed to the interaction of explicit and implicit learning conditions (Merikivi & Pietilä, 2014). Because of the more frequent exposure to versatile and meaningful input, students unconsciously learn the form of the words. In SLA this type of learning is termed “incidental language learning” (Hulstijn, 2003). However, “contextual language learning” (Elgort, Brysbaert, Stevens, & Van Assche, 2018) may be a better term, because unlike what is suggested by *incidental*, the learning is not accidental but, rather, the result of particular activities in meaningful contexts.

Indeed, in CLIL, the students get more opportunities to use the target language in meaningful communicative situations, which leads to conscious learning of the meaning of these words. This is closely related to the “involvement load hypothesis” (Laufer & Hulstijn, 2001), which suggests that the higher the degree of involvement on the part of the learner, the better it is for acquisition. The hypothesis consists of three components (i.e., need, search and evaluation), which refer to the knowledge (e.g., of a word) that is required to complete a task, the attempt to acquire this knowledge (e.g., the meaning of an unknown word), and the evaluation of one’s own performance (e.g., appropriate use of a word). More so than in traditional L2 classes, CLIL incorporates a greater involvement load, which may positively affect vocabulary learning.

Furthermore, the non-threatening atmosphere in the CLIL classroom (Nikula, 2010), for example in terms of error correction, most likely adds to the uptake of new words. In this regard, MacIntyre and Gregersen (2012) argue that positive associations and emotions act as facilitators in the language learning process. Because language learning is not the main and only goal in CLIL, the fear of using the target language and making mistakes eases off. Indeed, Nikula (2010) has shown that there is more student-teacher interaction in the CLIL class compared to regular L2 classes. This increased interaction may explain the lexical advantage of CLIL as it fits well with general theories of learning, suggesting that frequency of encounters is one of the most powerful predictors of learning (Ellis, 2002). However, it should be noted that an increase of student-teacher interaction is not necessarily a given in CLIL classrooms. Lo and Macaro (2015), for example, observed little interaction in classrooms in Honk Kong that had just started to experience the CLIL approach.

Some researchers claim that other factors, such as reading or extracurricular contact, also affect vocabulary development. Sylvén (2004), for instance, found that CLIL students had significantly more contact with English outside of school than their non-CLIL counterparts. Interestingly, the CLIL students in this study already scored higher on the first receptive vocabulary test, which was administered before CLIL instruction had started. This led the author to state that it is not possible to conclude that CLIL per se led to larger vocabulary gains. In addition, it should also be noted that not every study reported higher levels of vocabulary knowledge. For example, in a large-scale longitudinal study on the effects of CLIL education in the Netherlands, Admiraal et al. (2006) found no differences in receptive vocabulary knowledge between CLIL and non-CLIL learners.

What is striking about the studies overviewed above is that the lexical advantage has been observed for English, which may not be a coincidence, given that it is the uncontested global language, in the case of which the out-of-class exposure is significant. This raises questions on how CLIL functions in languages

other than English. In this regard, Pérez et al. (2016) examined the linguistic and sociolinguistic competences as well as the socio-educational outcomes of a French CLIL program in Andalusia. The findings of this study suggest that CLIL also works for French, especially with regard to linguistic competence, but at the same time the study warns of possible detrimental effects when not taking into account important social issues. Furthermore, also in the UK and Switzerland a few studies have included other languages than English. Wiesemes (2009), for example, found that UK pupils in a French CLIL program reported increased levels of motivation and enjoyment. In Switzerland, Gassner and Maillat (2006) found positive effects of CLIL on productive skills in French. On the other hand, Serra (2007) did not observe differences between CLIL and non-CLIL on language skills in Italian and Romansch. Closer to the specific CLIL context of the present study, De Smet, Mettwie, Galand, Hilgsmann, and Van Mensel (2018) examined the levels of anxiety and enjoyment of French-speaking Belgians in Dutch CLIL and non-CLIL contexts. Interestingly, the study also included English CLIL and non-CLIL contexts, which makes it possible to examine how CLIL interacts with different target languages. In this regard, the results reveal that, in addition to significant differences between CLIL and non-CLIL in general, the levels of anxiety and enjoyment diverge even more in English than in Dutch. This finding suggests an important role of the target language within CLIL. Therefore, De Smet et al. (2018) call for further empirical investigations in this area.

2.2. CLIL in Flanders

The CLIL context in the present study is situated in a secondary school in Flanders, the Dutch-speaking part of Belgium. Although Belgium is officially trilingual on state-level (Dutch, French, German), the educational system of the communities (i.e., the Flemish, Francophone and German-speaking communities) is organized unilingually. This means that the language of the community is the medium of instruction for all subjects (i.e., Dutch in Flanders), except for L2 classes, where it is common practice to use the foreign language as the language of instruction as soon as possible. This educational separation on community-level is the result of a long process of linguistic legislation since the birth of Belgium (see Bollen & Baten, 2010; Buyl & Housen, 2014). With the independence of Belgium in 1830, French was declared the only official language and it became the dominant language in government and administration. French was also the dominant language used by the Catholic clergy, the nobility, the industrials and the bourgeoisie. However, the 19th century witnessed the emergence of the so-called Flemish Movement, which aimed for the dutchification of education, the judiciary, the army and official administration. In 1963 the tug-of-war between the

Dutch-speaking and French-speaking parts of Belgium culminated in the instalment of the Dutch-French language border (Willemyns, 2002). The struggle for linguistic rights – often considered a struggle for social rights – has had many consequences, including the educational separation on community-level, and, as a result, different language policies in Flanders and Wallonia emerged (Bollen & Baten, 2010; Buyl & Housen, 2014; Van de Craen, Surmont, Mondt, & Ceuleers, 2011). This particularly applies to the communities' organization of bilingual education: Whereas the Franco-phone community already began with CLIL in 1998 (Chopey-Paquet, 2007), Flanders only started the program 16 years later, that is, in the school year 2014/2015.

Currently, five years after its implementation, CLIL programs are offered in more than 100 Flemish schools. In order to obtain the permission to implement CLIL, these schools had to submit a bulky application, stating, among other things, the aims of the CLIL program, the characteristics of the CLIL curriculum (e.g., the language(s) and subject(s) involved, teaching material, etc.), the staff policy, the quality assurance policy, and so on. In this application the schools also have to comply with a number of restrictions imposed by the Flemish government (De Vlaamse Regering, 2004). First, CLIL can only be organized on the level of secondary education. Second, the number of courses taught in a foreign language, other than traditional L2 courses, is limited to 20% of the total curriculum. Third, it is obligatory to offer a parallel Dutch-speaking program, enabling pupils who opt out to take the same courses in Dutch. And fourth, the only languages in which CLIL is allowed are English, French and German. Although actually all languages (i.e., national languages, migrant languages, minority languages, and border languages) are eligible as medium of instruction in CLIL – at least, this was the spirit of the European language planners (Pérez et al., 2016) – the reality in Flanders shows otherwise, in that the policy does not allow for teaching in minority languages (e.g., Spanish or Italian) or migrant languages (e.g., Turkish). In fact, financial government support of minority-language teaching projects (Dutch: *onderwijs in eigen taal en cultuur*, OETC) was discontinued in 2011 (Bollen & Baten, 2010). For example, the Foyer project in Brussels, which was launched in 1981 and which provided part of the education in Spanish, Italian and Turkish, no longer receives public funding. However, on a voluntary basis, the project still runs on a smaller scale.

Closer inspection of the presently available CLIL programs in Flanders reveals that history and geography are the most popular courses, and English is the most popular language (offered in 64 schools). CLIL in French ranks second (offered in 45 schools), while CLIL in German is only offered in four schools.² Clearly, also Flanders favors English as medium of instruction in CLIL. The fact

² <http://onderwijs.vlaanderen.be/nl/clil-content-and-language-integrated-learning>

that English is the most popular CLIL-language and that history and geography are the most popular CLIL-subjects is most likely related to the competence requirements set by the Flemish government. CLIL teachers are obliged to attest competences in both the CLIL subject (by a bachelor or master degree) and the CLIL language (again by a bachelor or master degree, or by an official language proficiency certification). The required language proficiency level is the C1 level of the CEFR in all skills. The school and the teachers experience this requirement as too high, and it is a perception that the C1 level is more easily achieved for English than for French. Therefore, as a side effect, the schools and teachers seem more inclined to choose English.³ The frequent choice for history and geography is the result of the Flemish bachelor program for secondary teacher training, which involves a combination of the two subjects. It is mostly the teachers that have a foreign language in this combination of two who will (be asked to) become CLIL-teachers. Apparently, among the available teachers, the combination with history/geography was a popular one. In a recent survey of the Flemish schools inspectorate regarding the present-day CLIL practice in Flanders, the schools and the teachers indicate that these competence requirements as well as the abovementioned restrictions hinder the rollout of CLIL.⁴ So, while CLIL is now successfully launched in Flanders, the rigid rules and regulations may need some adjustments.

Given that CLIL has only recently been introduced in Flanders, research into the linguistic as well as extra-linguistic effects of CLIL remains rather scarce. The limited published findings so far relate to the research activities that took place in CLIL pilot projects: Before 2014-2015, a small number of schools were granted permission by the Minister of Education to embark on experimental CLIL projects. For example, in the so-called STIMOB-project (Stimulerend Meertalig Onderwijs in Brussel), Van de Craen, Ceuleers, Lochtman, Allain, and Mondt (2006) found that CLIL learners in primary education in Brussels had equal and sometimes better knowledge of both the L1 (Dutch) and the L2 (French) compared to the non-CLIL children. In another experimental CLIL project, Strobbe, Sercu, Strobbe, and Welcomme (2013) investigated the outcomes of CLIL in nine Flemish secondary schools. Quantitatively, no significant positive effect on the target language (either French or English) was found. Although the CLIL learners were capable of communicating fluently, their fluency was restricted to the content provided in the CLIL classroom and their language was often ungrammatical and unidiomatic. Qualitatively, however, the researchers observed higher self-confidence among CLIL learners when they had to express

³ According to an inquiry of the Flemish schools inspectorate: www.onderwijsinspectie.be/sites/default/files/atoms/files/CLIL-RAPPORT%20zonder%20bijlagen%20-%2020170112.pdf

⁴ See fn. 4

themselves in the target language as well as increased motivation and enthusiasm with regard to the course-specific content. In addition, and more importantly for the present study, learners reported having noticed considerable improvement in relation to course-specific vocabulary development.

In addition to these pilot project studies, a recent study examined the effects of CLIL (in French) on mathematical content learning in a Flemish secondary school (Surmont, Struys, Van Den Noort, & Van de Craen, 2016). The results showed that the 35 CLIL-learners in seventh grade outperformed the 72 non-CLIL counterparts on a mathematical test. The researchers ascribe the difference to the CLIL approach. However, their conclusion that the data provide “clear proof” (p. 328) might be overstated. As pointed out by the authors, the groups were self-selected, which means that the pupils were able to choose whether or not to participate in the CLIL program (this was also the case in the pilot project studies above). This means that other lurking factors, such as motivational levels, parental support, and the like, could have had an impact. For instance, it is possible that CLIL is chosen by higher achieving and well-supported pupils in the first place. A number of studies in the Spanish context indeed showed that the parents of the children in the CLIL stream often have a university degree, indicating higher socio-economic status (SES; e.g., Alonso, Grisaleña, & Campo, 2008; Bruton, 2011; Pérez et al., 2016). Pérez et al. (2016) even observed that the effect of SES is reinforced when the CLIL program incorporates languages other than English, thus making non-English CLIL a program for the select few.

Nevertheless, in light of the tendency of the existing literature to focus on CLIL in English, it is interesting to point out that the above studies on CLIL in Flanders have examined CLIL with French as the medium of instruction. The moderately positive findings of the (pilot project) studies in CLIL French indicate that previous findings for L2 English in CLIL may be transferable to L2 French, especially with regard to the acquisition of vocabulary, in which domain most positive effects of CLIL have been observed so far, at least for L2 English (Dalton-Puffer, 2008). It is the present study's aim to explore this.

2.3. Research questions

The present study was conducted in Flanders, a region where both English and French are considered an asset. However, attitudes and exposure to these languages are different. Attitudes towards French, for instance, are rather negative, compared to the positive attitudes towards English (Lochtman, Lutjeharms, & Kermarrec, 2005). Mettewie (2015) suggests that this negative attitude may hinder the acquisition of L2 French. Indeed, Dewaele (2005) found a relationship between negative attitudes and poor achievement among Flemish students.

In addition to these different attitudes, English and French differ in the extent to which learners are exposed to them outside the classroom: Whereas English is pervasive in the daily lives of children in Flanders (De Wilde, Brysbaert, & Eyckmans, 2019), exposure to French, despite it being an official language, is limited. The difference in exposure undoubtedly has consequences for L2 acquisition. In a study focusing on the success in learning French and English in regular foreign language classes in Flanders, Housen et al. (2001) found that pupils had better receptive and productive skills in English than in French. With regard to vocabulary, students were not only able to recognize English words easier than French words, but they also commanded richer and more varied English vocabulary compared to French. According to Housen, Janssens, and Pierrard (2001), this difference is not only due to greater typological affiliation between English and Dutch, but is also the result of more extracurricular contact with English than with French. On the other hand, the pupils demonstrated better knowledge of formal language use in French than in English. This finding is not surprising because the extracurricular contact with English generally takes place in informal settings, whereas exposure to French is limited to the formal language use in class. Informal encounters with French, for example through television, occur, but to a considerably lesser extent than could be expected in a bilingual Dutch-French country.

Given the different findings with regard to vocabulary knowledge in French and English of Flemish pupils in regular foreign language classes and given the newly emerged educational context of CLIL in Flanders, the present study sought to assess the vocabulary knowledge in French and English of Flemish pupils in a CLIL context. It is important to note explicitly that the study does not evaluate the benefits of CLIL over non-CLIL, but aims to establish the initial level of vocabulary and to address the possibly differential vocabulary development in a French and English CLIL context. Two research questions guided the study:

1. What is the level of receptive and productive vocabulary knowledge in French and English of beginning Flemish CLIL learners? (RQ1)
2. How do English and French CLIL students differ in receptive and productive gains? (RQ2)

With respect to RQ1, we expected a clear difference between English and French, with higher scores for English. Such a finding would be in line with Housen et al. (2001) and reflect the differences that exist in terms of attitudes and exposure. With respect to RQ2, we formulated two alternative hypotheses, the first assuming a larger gain for French than for English, and the second predicting the opposite. The reasoning behind the first hypothesis was that learners are expected to have more

room for improvement in their French vocabulary and less in their English vocabulary, because they already had considerable English vocabulary knowledge before they started the CLIL program. The rationale for the second hypothesis was that more knowledge leads to more gains, that is, “the rich get richer.”

3. Methodology

3.1. Participants

We collected data from 104 pupils in a large secondary school in the province of Antwerp (Flanders). However, participants with either English or French as home language ($N = 21$) and participants that only partially completed the tests were excluded from the study ($N = 8$). Therefore, the data reported here comes from 75 pupils (28 females, 47 males). All participants were L1 speakers of Dutch and aged 12 to 14 at the time of testing ($M = 12.9$, $SD = 0.3$). Data from a language background questionnaire further revealed that, in terms of previous experience with L2 learning, all the pupils had received formal instruction in French for three years prior to the time of study (in Grades 5-7). In contrast, none of the pupils had received formal language instruction in English before. However, due to heavy media exposure, Flemish children acquire English from an early age onwards and before any formal education gets under way (Goethals, 1997; Simon, Lima Jr, & De Cuyper, 2016). In this regard, the participants reported different engagement with French and English media. On a 5-point Likert-scale from *never* to *daily*, the pupils' average media-engagement with English (3.19) was significantly higher than their media-engagement with French (1.95; Wilcoxon signed-rank test: $z = -6.54$, $p < .001$, $r = -0.53$). The school started the CLIL program in 2015 and runs the program in both L2 English and L2 French in Grades 7 and 8. The program involves such subjects as economics, history, computer science, music education, and religion.

Table 1 Distribution of hours per week of CLIL and regular language classes

	CLIL	Language class
English	1	2
French	2	4 or 5

For this study, we examined pupils who were in Grade 8, taking both history in French and music in English. More specifically, the pupils received two hours of history in French and one hour of music in English per week. They additionally took four or five hours of French and two hours of English per week. The difference for the number of hours of French was related to program of study: Pupils taking the modern languages program received one extra hour of

French ($N = 15$). Table 1 summarizes the distribution of language and CLIL education in hours per week. It can be noticed that the exposure to French was double the exposure to English. In other schools in Flanders other choices are made, which indicates that CLIL comes in different shapes and colors, not only in Flanders but also across Europe (Coyle, 2007; Lasagabaster, 2008; Wolff, 2002). Although this variation in CLIL implementations obviously means that findings in CLIL research are context-specific, they are revealing for CLIL in general.

3.2. Research instruments

The participants took for each language two Vocabulary Levels Tests (VLT): a reception test and a production test. In the receptive VLT, participants matched decontextualized words with definitions, while in the productive VLT, they completed words that appeared in short sentences.

For English, we used the second Receptive Vocabulary Levels Test, developed by Schmitt, Schmitt, and Clapham (2001) and the Productive Vocabulary Levels Test, constructed by Laufer and Nation (1999). Both tests contain five levels of word frequency: the 2000-word level, the 3000-word level, the 5000-word level, the 10 000-word level and the university/academic word level. The first two levels (2000-level and 3000-level) contain high-frequency words. Research has established that knowledge of these word families is required for basic daily conversation (Adolphs & Schmitt, 2003), movie viewing (Webb & Rodgers, 2009) and reading texts (Schmitt & Schmitt, 2014). As a mid-frequency level, the 5000-word level represents the boundary towards low-frequency items. This level is seen as the threshold for dealing with authentic texts/discourse fluently (Schmitt & Schmitt, 2014). The 10000-word level contains low-frequency items. Having reached this level, L2 users are able to read practically any texts without major difficulty (Nation, 2006). Finally, the university/academic word level is not based on frequency but contains words that occur widely in academic discourse and textbooks. Contrary to the common belief that university/academic words represent infrequent and specialized words inaccessible from general language, Masrai and Milton (2018) demonstrated that the majority of these words actually fall within the 3000 most frequent words. Because we did not expect beginning secondary school pupils to know low-frequency words, we only included the first three levels in the receptive vocabulary test and the first two levels in the productive vocabulary test. Indeed, as Kremmel and Schmitt (2018) point out, administering the levels representing low-frequency words to beginner learners would be time poorly spent.

In the receptive test, each level consists of thirty definitions and sixty words organized in groups of three definitions and six words. Participants are asked to match

the meanings of the right-hand column with a word from the left-hand column. The words in both columns are representative of the words at that frequency level. The following is an excerpt from the test:

Receptive vocabulary levels test

- | | |
|-------------|--------------------------------|
| 1. Business | |
| 2. Clock | |
| 3. Horse | ___ part of a house |
| 4. Pencil | ___ animal with four legs |
| 5. Shoe | ___ something used for writing |
| 6. Wall | |

The productive test contains 18 words per level. The participants are presented with sentences including a missing word. They are asked to fill in the blanks with appropriate target words. The first few letters of the target words are provided, as well as an indication of the total number of letters, in order to prevent the participants from filling in other words which, although semantically suitable, would come from non-targeted frequency levels. The following is an example test item:

Productive vocabulary levels test

There are a doz___ eggs in the basket.

Equivalent tests were used to measure the vocabulary size for French: the receptive vocabulary test as developed by Batista (2014) and the productive vocabulary test as developed by Peters, Velghe, and Van Rompaey (2015). Analogous to our decision regarding the English tests, we only included the first three levels of the receptive test and the first two of the productive test. Likewise, the receptive test aimed to elicit 30 word-definition matches per level, while the productive test aimed to elicit 18 word completions per level. Different from the English test, we sometimes explained or translated difficult words in the French sentences. The following are two excerpts from the tests:

Receptive vocabulary levels test

- | | |
|----------------|----------------------------------|
| 1. Ambassadeur | |
| 2. Enfance | |
| 3. Portrait | ___ conflit |
| 4. Rayon | ___ première partie de la vie |
| 5. Trouble | ___ représentant du gouvernement |
| 6. Vœu | |

Productive vocabulary levels test

Après une guerre de plus de quatre ans, les deux pays voisins (=buurlanden) ont signé la p____.

The vocabulary levels tests estimate the total number of words a learner knows. The estimates can be used to compare groups of learners, measuring vocabulary growth within a language. Because the tests for English and French were developed and designed according to the same principles, we assumed that the estimates could be useful for comparing vocabulary knowledge and growth across the particular language.

3.3. Procedure and scoring

The participants completed the four tests twice over a 3-month-interval. The pre-test took place at the beginning of the CLIL program (October 2015), and the post-test was performed after three months (January 2016). After a briefing session with one of the researchers, the tests were administered by the teachers themselves in one of their CLIL classes: First the productive levels test and then the receptive levels test. The order of testing for French and English was random and dependent on practical matters (e.g., different week calendars, other assignments, etc.). The participants had approximately 30 minutes to complete the productive test and 20 minutes to complete the reception test. At the beginning of each session, the participants were given instructions regarding the content of the tests and the manner in which they should be solved. Moreover, they received explanation about the goal of these tests, that is, that their vocabulary knowledge was examined solely for research purposes and that the results would not have an impact on course grades. Furthermore, in order to avoid guessing, the participants were asked to only provide answers of which they were certain.

Taking the same testing instruments twice may entail the possibility of a practice effect from the first to the second time. Considering that parallel test versions of the VLT exist, in principle, these parallel versions could have been administered to the group of CLIL learners in the present study, which would have ruled out any memory effect confounding the results. However, Kremmel and Schmitt (2018, p. 4) state that parallel tests are “not found to be equivalent enough to be used to measure the learning gains of any individual learner, and for this purpose, the same version should be used twice.” They point out, though, that a substantial amount of time should elapse between the administrations; ideally, more than a month.

With regard to the scoring of the tests, we marked test items as either correct (1 point) or incorrect (0 points). Mistakes in grammatical form or in spelling

were not penalized in the Productive Vocabulary Levels Tests. Answers were marked as correct as long as the meaning of the word could be derived from a phonologically recognizable form. The maximum score that could be obtained for each level in the receptive vocabulary test was 30, whereas the maximum score for each level in the productive vocabulary test was 18, resulting in the maximum test score of 90 for the receptive vocabulary test and 36 for the productive vocabulary test. For our analysis, we used the pupils' individual test scores. This accuracy approach is different from the reached levels approach. For example, according to Schmitt, Schmitt, and Clapham (2001), learners have to obtain a score of 26/30 on a particular level in order to conclude that the level is mastered. However, the accuracy approach enabled us to make extrapolations on vocabulary size (see below). Table 2 lists the VLTs that were used in the present study together with their maximum scores.

Table 2 Overview of vocabulary levels tests per type of knowledge and language

L2	Type of knowledge	Adapted from	Levels	Maximum score
English	Receptive	Schmitt, Schmitt, & Clapham (2001)	2000, 3000, 5000	/90
	Productive	Laufer & Nation (1999)	2000, 3000	/36
French	Receptive	Batista (2014)	2000, 3000, 5000	/90
	Productive	Peters et al. (2015)	2000, 3000	/36

3.4. Data analysis

Our response variable of interest was the gain score achieved by each participant on the French and English tests (with gain score = posttest score - pretest score). We analyzed the data for the production and reception tests separately as the difference between both types of knowledge was in itself not a matter of interest. It should be noted that all participants followed the CLIL program in both English and French CLIL simultaneously. This setting allowed us to compare the two CLIL languages in a paired study design.

For both the productive and the receptive results, we first compared the results for the pretests and posttests separately by means of a Mann-Whitney-Wilcoxon test (i.e., pretest: English vs. French and posttest: English vs. French). A non-parametric test was chosen because of a clear violation of the equal variances assumption. Then we examined whether there was a significant improvement in the gain scores for both languages separately, using a paired-samples *t*-test (two-tailed; i.e., English: pretest vs. posttest and French: pretest vs. posttest). Finally, we evaluated the differences between the mean gain scores of both languages, again by means of a paired-samples *t*-test (two-tailed; i.e., gain: English vs. French). In total, 10 significance tests were performed. We controlled

for the family-wise error rate by means of Bonferroni correction, which consists of dividing the alpha level by the number of tests, and accordingly adjusted our significance level to .005 (= .05/10).⁵

4. Results

4.1. Productive levels tests

Table 3 presents an overview of the main statistics for the results of the production tests per language. The paired individual scores for the pre- and posttest per language are visualized in Figure 1. Cronbach's alpha estimates were .91 (95% CI = 0.86 to 0.95) and .58 (95% CI = 0.40 to 0.76) for the English and French test, respectively.⁶

Table 3 Statistics for the production tests: minimum, 1st quartile, median, mean, 3rd quartile, maximum

English	Min	Q1	Med	<i>M</i>	Q3	Max
Pre	1	7	12	12.0	15.0	27
Post	6	13	16	16.8	21.5	30
Gain	-4	3	5	4.9	7.0	11
French	Min	Q1	Med	<i>M</i>	Q3	Max
Pre	0	4.0	5	5.3	7.0	12
Post	0	6.5	9	8.7	10.5	16
Gain	-8	2.0	3	3.5	6.0	10

⁵ We did not perform ANCOVA, which is often the preferred model in a pretest-posttest study design, for three reasons. First, gain score analysis is better suited to answer RQ2 than ANCOVA. Gain score analysis answers the question how groups differ, on average, in gains, whereas ANCOVA evaluates whether post test means differ between independent groups, adjusted for pretest scores (Fitzmaurice, Laird, & Ware, 2004). Second, the groups in our data, that is, English vs. French, are not independent (as in ANCOVA). This allowed us to compare the paired gains in both languages, which in turn simplified the statistical analysis to a paired-samples *t*-test. Third, there was a substantial difference in range between the results for French and English. As there were no observations for French for the higher scores for English, we would have to extrapolate the results for French, which would arguably hamper the reliability of the regression analysis.

⁶ The low Cronbach's alpha for the French Production Tasks is partially related to two participants who performed much worse on the posttest than on the pretest (both scored 3 and 12 on the pretest but 0 and 4 on the posttest, respectively). Eliminating the two participants from the dataset would have improved the value to 0.67 (95% CI = 0.53 to 0.81). We do not know why the gain scores for these two participants are negative. We decided to retain the two participants in the dataset, so as not to artificially inflate Cronbach's alpha.

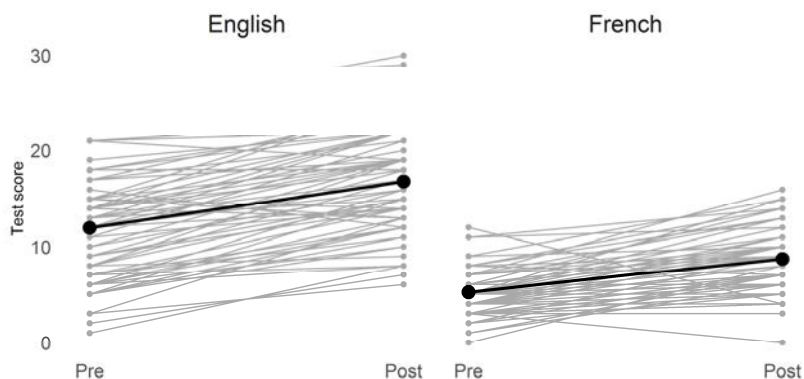


Figure 1 Individual scores for the production tests per language (Grey lines indicate changes in scores. The black line connects the mean results for each test)

The first general observation is the large difference in the range of scores for both languages, with much higher scores on the English test than on the French one, both on the pretest and the posttest. Tellingly, more than half of the participants ($n = 41$; 54%) achieved a score on the English pretest that was equal to or higher than the maximum score of 12 on the French pretest, which indicates that the initial level of vocabulary was generally higher in English than in French. Comparing the pre- and posttest scores for both languages, we find that participants tended to attain a higher score for English, both on the pretest and the posttest ($p < .001$, for both pretest: English vs. French and posttest: English vs. French). An extrapolation of the mean correct responses on these VLTs (36 items) to knowledge of the 3,000 most frequent words yielded productive vocabularies at the time of the pretest of 1000 ($12/36 \times 3000$) words for English and only 442 ($5.3/36 \times 3000$) words for French, while at the time of the posttest the extrapolation revealed productive vocabularies of 1400 and 725 words, respectively. The second finding was that productive vocabulary knowledge improved in the case of both languages. For English, there was an average gain of 4.9 ($SD = 3.4$) for English ($t = 12$, $df = 74$, $p < .0001$, 95% $CI = 4$ to 5.6). For French, there was an average gain of 3.5 ($SD = 2.9$, $t = 10.4$, $df = 74$, $p < .0001$, 95% $CI = 2.8$ to 4.2).⁷

⁷ The normality assumption was violated in the case of French (Shapiro-Wilk test: $p = .009$). We therefore performed an additional non-parametric sign test, which also appeared highly significant ($p < .0001$; only three out of 75 pupils achieved a lower score on the posttest than on their pretest).

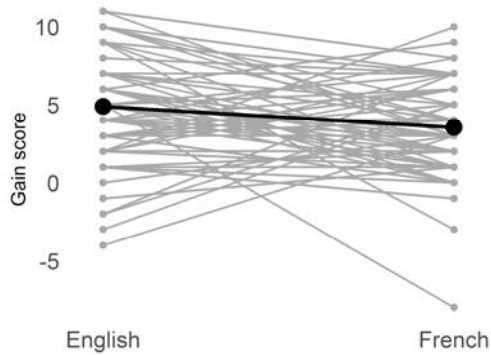


Figure 2 Individual gain scores per individual participant for the English and French production tests (gain score = posttest score - pretest score; The black line connects the mean gains for each language)

The comparison of the gain scores between the two languages (i.e., English posttest – pretest vs. French posttest – pretest) is visualized in Figure 2. The difference in paired means equals 1.3 (95% *CI* = 0.6 to 2.0), which is statistically significant based on a paired-samples *t*-test ($t = 3.7$, $df = 149$, $p < .001$, $r = 0.29$), but nevertheless rather low given that the maximum score equaled 36. The estimated effect size was also rather low, as is indicated by the *r* value (Rosnow & Rosenthal, 2005, p. 328) and the maximum difference of 2.0 points (out of 36) on a 95% *CI*. In terms of vocabulary size, these numbers represent an average gain of 408 English words (with a 95% *CI* between 333 and 467 words) and an average gain of 292 French words (with a 95% *CI* between 233 and 350 words).

4.2. Receptive levels tests

Table 4 presents an overview of the main statistics for the results of the reception tests per language. The paired individual scores for the pre- and posttest for English and French are depicted in Figure 3. Cronbach's alpha was .83 (95% *CI* = 0.76 to 0.91) for both the English and French test, a value that can be considered satisfactory.

Overall, we can see again that pupils tended to achieve higher scores in this case for English than for French, both on the pretest and the posttest ($p < 0.001$ for both the pretest: English vs. French and the posttest: English vs. French). It should be emphasized again that 24 participants (32%) achieved an English pretest score that was higher than or equal to the maximum pretest score for French (53). An extrapolation of the mean correct responses on the receptive VLT (90 items) to knowledge of the 5,000 most frequent words yielded receptive vocabularies at the time of the pretest of 2,500 ($45/90 \cdot 5000$) words

for English and 1,694 ($30.5/90 \times 5000$) words for French while at the time of the posttest the extrapolation revealed receptive vocabularies of 2,850 and 2,100 words, respectively.

Table 4 Summary for the reception tests: minimum, 1st quartile, median, mean, 3rd quartile, maximum

English	Min	Q1	Med	<i>M</i>	Q2	Max
Pre	13	34	44	45.0	55.5	84
Post	19	40	51	51.3	65.5	85
Gain	-34	0	5	6.2	13.0	34

French	Min	Q1	Med	<i>M</i>	Q3	Max
Pre	9	23.5	29	30.5	37.0	53
Post	14	29.5	37	37.8	45.0	63
Gain	-10	2.0	7	7.3	13.5	22

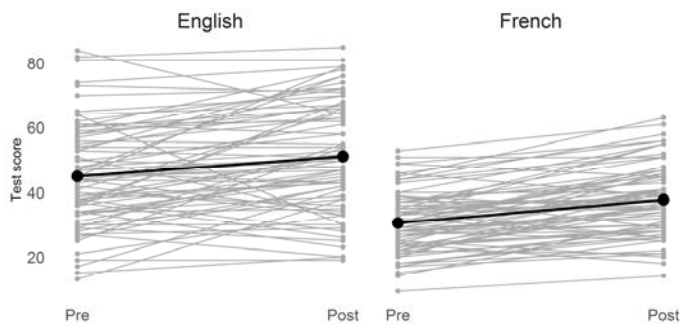


Figure 3 Individual scores for the reception tests per language (Grey lines indicate individual changes in score. The black line connects the mean results for each test)

There was also significant improvement on the receptive tests for both languages. On average, there was an approximate gain of 6.2 ($SD = 12.2$) for English ($t = 4.4$, $df = 74$, $p < .0001$, $95\% CI = 3.4$ to 9) and of 7.2 ($SD = 7.7$) for French ($t = 8.2$, $df = 74$, $p < .0001$, $95\% CI = 5.5$ to 9), which corresponds to a receptive vocabulary size gain of 344 English words (with a $95\% CI$ between 189 and 500 words) and 400 French words (with a $95\% CI$ between 306 and 500 words). If we compare the gain scores on both languages, illustrated in Figure 4, we can see that the mean gain was in this case slightly higher for French than for English. However, the difference in paired means was negligible, amounting to only 1 point out of the total of 90. The difference is also not significant ($t = 1$, $df = 149$, $p = .28$).

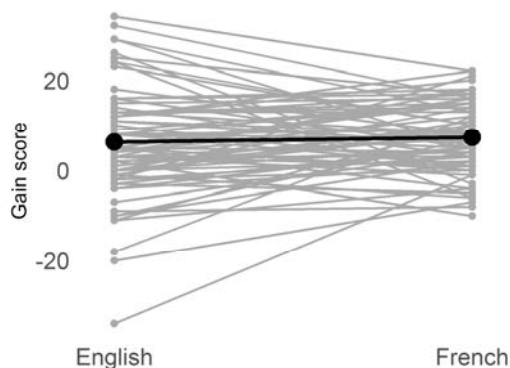


Figure 4 Individual gain scores per individual for the English and French reception tests (The black line connects the mean gains for each language)

5. Discussion

In response to the dearth of comparative empirical work on the role of target languages in CLIL, the present study examined the level of vocabulary knowledge in English and French of first-time CLIL learners in Flemish education (RQ1) and whether the vocabulary improvement in this CLIL context is target language-dependent (RQ2).

With respect to RQ1, the present study found that the scores on the receptive and productive vocabulary levels tests were significantly higher for English than for French, both on the pretests and the posttests. This result is in line with the findings reported by Housen et al. (2001), who examined the foreign language acquisition of French and English in regular foreign language classes in Flanders. The results for the pretests may seem remarkable, given that the pupils received no previous formal instruction in English in contrast to three years of formal instruction in French. Housen et al. (2001) referred to the typological affiliation between English and Dutch as part of the explanation for the difference between English and French vocabulary knowledge. In fact, the positive impact of typological resemblance particularly applies to the most frequent words, although negative influences through the so-called “false friends” are also to be reckoned with. Another, and arguably more plausible explanation for why Flemish pupils have such a high starting proficiency level of English in comparison to French, is the high amount of exposure to English (see De Wilde et al., 2019; Simon, Lima Jr, & De Cuypere, 2016; Sundqvist & Sylvén, 2016). Whereas contact with French is mostly limited to the classroom (even with French being the second official language of Belgium), English is omnipresent in the daily lives of Flemish pupils. Indeed, also the participants of the present study reported significantly higher media-engagement with English compared to French. In addition, the fact that the attitudes of Flemish students towards French and

English are different (Lochtman et al., 2005) may have influenced the results in the sense that negative attitudes towards French may decelerate the acquisition process (Dewaele, 2005; Mettewie, 2015).

Naturally, not every pupil is exposed to English to the same extent. Substantial differences in this regard may explain why the scores for English ranged widely. For example, an extrapolation of the lowest and the highest scores on the English pretests revealed a vocabulary size, ranging from 83 to 2,250 words for productive knowledge and from 722 to 4,667 words for receptive knowledge. By comparison, the scores for French were closer to each other, suggesting that the exposure to this language was largely similar across the participants (i.e., limited to the classroom). Interestingly, our results for English are reminiscent of Sylvén's (2004) study, which investigated the vocabulary development in English of Swedish CLIL and non-CLIL learners. She observed that the learners with the most exposure to English outside the classroom scored best on vocabulary tests. Remarkably, this observation applied to both the CLIL and the non-CLIL group. In other words, extramural exposure can be more important for vocabulary acquisition than CLIL itself. The wide range of the scores on the English tests that was obtained in the present study may suggest the same thing. Unfortunately, because we only procured minimal data on extramural language engagement (and there was no non-CLIL group), the link between the amount of exposure and general vocabulary acquisition could not be further pursued.

Moving on to RQ2, the present study found that, despite the clear differences between English and French, the productive and receptive vocabulary knowledge of the participants improved significantly in both languages. In the case of productive knowledge, improvement constitutes an average of 408 English and 292 French words, while for receptive knowledge it amounted to 344 English and 400 French words. This improvement was not as self-evident as it may seem, considering that it happened in the relatively short timeframe of three months, meaning that the learning rate was between 3.2 and 4.5 new words per day. By comparison, estimates of vocabulary knowledge by adult native speakers indicate that they learn about seven new words per day (Brybaert, Stevens, Mandera, & Keuleers, 2016). In other words, the learning rate in the present study was considerable. In addition, it should be emphasized that the improvement relates to general vocabulary knowledge in English and in French, and not to course-specific vocabulary (in this study history and music). This finding is different from previous research on CLIL, which provided evidence for vocabulary gains particularly in technical and semi-technical terms (Dalton-Puffer, 2008, 2009; Ruiz de Zarobe, 2011). It should be emphasized that also the Flemish learners in Strobbe et al. (2013) only reported course-specific vocabulary gains.

Interestingly, in addition to the finding that general vocabulary knowledge improves in both English and French, the present study shows that the level of improvement is comparable across the languages. This goes against the hypotheses formulated above. Instead of seeing a larger effect for English than for French, or the other way around, the present study's findings suggest that vocabulary development in French keeps pace with that in English, even though the respective gains occur at different levels. Considering this, the fact that the vocabulary development seems to run parallel in English and in French puts the effect of extramural exposure into perspective. While it is true that the higher media-engagement with English leads to higher initial levels for English (both in production and reception), as well as results in a wider range in the scores for English compared to French (reflecting in all probability various levels of engagement with English among participants), it does not automatically bring about significantly greater gains for English compared to French. Actually, given that the progress is similar in both languages, it shows that intramural language exposure plays an equally important role in lexical development. Indeed, it should be recalled that the in-school exposure to French, both in CLIL and foreign language classes, is twice as high as in the case of English. In other words, despite the rather difficult initial state for CLIL in French in Flanders (i.e., negative attitudes towards French and minimal exposure outside of school), the increased in-school exposure yields hopeful results for L2 French, at least in terms of vocabulary knowledge.⁸

The implication of this finding seems to be that CLIL can safely include other languages than English, thus complying with the original multilingual aspirations of this approach. However, in order to avoid drawing false conclusions from the present study, it is important to note that the purpose of the investigation was not to compare the effectiveness of CLIL in the case of two different languages (and, as a matter of fact, nor was the purpose to assess its general effectiveness over non-CLIL). To examine this question, other variables such as out-of-school exposure and classroom input should be kept constant, which is obviously impossible in this kind of classroom-oriented research. Instead, the present study was intended to determine the vocabulary knowledge in English and French and to evaluate whether the degree of the development depends on the target language involved. With this aim in mind, the pairwise comparison in the present study suggests that learners in a CLIL class are capable of developing the vocabulary knowledge of a language to which they are less exposed

⁸ The results can be regarded as promising in light of the clock-like regularity with which Flemish media report that knowledge of French among Flemish pupils is on the decline. These reports are usually not based on empirical academic studies but on fragmentary results of entry or placement tests, or finals.

outside of the school and of which the initial knowledge is rather limited. However, an important provision seems to be that there is sufficient parallel exposure in L2 classes. Obviously, this finding only applies to French in the Dutch-speaking part of Belgium. It remains to be seen if the same results would be obtained for the other national/border language in Belgium (German) or for languages of a different nature (migrant or non-neighboring languages). Also, we do not think that it makes sense to extrapolate our findings to other educational settings in different countries or regions involving different languages (e.g., CLIL in English and Dutch in the French-speaking part of Belgium).⁹ In other words, the findings of the present investigation cannot be taken as a basis for a recommendation to implement CLIL for languages other than English all over.

In light of the above, even though the present results seem to provide general support for CLIL, also for languages other than English, caution in terms of implementation is advised for a number of reasons. First of all, one limitation of the present study is that only one CLIL-school was included. This shortcoming restricts the generalizability of the findings to other CLIL schools in Dutch-speaking Belgium, because as Coyle et al. (2008, p. 101) pointed out, “there is a lack of cohesion around CLIL pedagogies. There is neither one CLIL approach nor one theory of CLIL.” In other words, more research is needed, not only with respect to different foreign languages involved, but also with respect to *inner-CLIL* differences. It is crucial in future CLIL research to examine the diversity within CLIL (e.g., in terms of amount and type of input, amount of interaction and output, different types of CLIL teaching; see de Graaff et al., 2007), and how this diversity impacts learners’ language attainment. In this regard, it should be recalled that the present study involved two different content subjects, that is, history and music. Naturally, teaching differs according to the subject and such differences (e.g., history is likely to be more text-based than music) may have an impact on “picking up” new words. Further investigation is required to investigate the effect of the subject matter on CLIL learners’ L2 development.

The second limitation of the present study is that we did not examine the influence of other factors, such as motivation, socioeconomic status, parental support, and the like. In this respect, Bruton (2011) argued that (self-selected) CLIL learners are usually different from the outset, and, as a result, it remains unclear whether the positive results obtained in many CLIL studies are attributable to CLIL or to other factors. It should be clarified that his critique was directed at control group studies, comparing the results of CLIL versus non-CLIL

⁹ At present a large-scale project is being conducted in French-speaking Belgium, correlating linguistic performance with cognitive, educational and socio-affective variables in Dutch and English CLIL (Bulon, Hendriks, Meunier, & Van Goethem, 2017).

learners. Such control-group comparisons are a common research design to investigate the effectiveness of pedagogical programs. However, Cenoz et al. (2014) criticize the use of control-group comparisons (see also Berthele & Vanhove, 2020). The main reason is that the participants in these studies are rarely randomly assigned to treatment and control groups, and that, in effect, a range of factors remain uncontrolled for. The absence of random group assignment clearly constitutes a major drawback and precludes drawing conclusions for the beneficial effect of CLIL. Therefore, claims in support of CLIL should be taken with circumspection (e.g., Cenoz, Genesee, & Gorter, 2014, p. 257).¹⁰ However, the present study was not intended to evaluate the effectiveness of CLIL over non-CLIL, and, for this reason, it did not include a non-CLIL group. Nevertheless, even if the current investigation had included a non-CLIL group showing less improvement or non-significant changes, it would be premature to associate the differences in language gains (or content gains, for that matter) with CLIL. Actually, that is not even the point. The aspiration of building multilingual and multicultural societies may be reason enough to foster CLIL in Flanders (for French and other languages).

6. Conclusion

The present study examined the vocabulary development of Flemish pupils in a CLIL-context. In terms of language development, research on CLIL has so far mainly focused on the *lingua franca* English, leaving much unknown about the development of other languages taught in this manner (Cenoz et al., 2014; Pérez et al., 2016). This study, therefore, measured the initial level of receptive and productive vocabulary knowledge in both English and French (i.e., at the start of the CLIL program), as well as the gains after three months. The results show that although vocabulary knowledge was generally better for English than for French, the level of improvement in productive and receptive vocabulary knowledge was the same across these languages. In line with the widespread explanation in CLIL-related discourses, the significant gains in productive and receptive vocabulary knowledge can be related to increased exposure, which functions as a trigger for acquisition processes (Dalton-Puffer, 2008). Most importantly, however, the present study has shown that despite the numerous differences that exist regarding particular target languages (in this study English and French) in terms of status, extramural and in-school exposure, attitudes, and so on, progress can be made in both languages, even though the respective gains in English and French vocabulary knowledge occur at different levels.

¹⁰ However, for a recent good practice including a *randomized* controlled field experiment on the effects of CLIL employing a repeated-measures design (i.e., pre-test, posttest, follow-up), see Piesche, Jonkmann, Fiege, and Keßler (2016).

References

- Ackerl, C. (2007). Lexico-grammar in the essays of CLIL and non-CLIL students: Error analysis of written production. *Vienna English Working Papers*, 16, 6-11.
- 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.
- Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, 24, 425-438.
- Aguilar, M., & Muñoz, C. (2014). The effect of proficiency on CLIL benefits in Engineering Students in Spain. *International Journal of Applied Linguistics*, 24, 1-18.
- Alonso, E., Grisaleña, J., & Campo, A. (2008). Plurilingual education in secondary schools: Analysis of results. *International CLIL Journal*, 1, 36-49.
- Banegas, D. L. (2011). Content and language integrated learning in Argentina 2008 - 2011. *Latin American Journal of Content & Language Integrated Learning*, 4, 33-50.
- Batista, R. (2014). *A Receptive Vocabulary Knowledge Test for French Learners with Academic Reading Goals*. Montreal, Quebec: Concordia University.
- Berthele, R., & Vanhove, J. (2020). What would disprove interdependence? Lessons learned from a study on biliteracy in Portuguese heritage language speakers in Switzerland. *International Journal of Bilingual Education and Bilingualism*, 23, 550-566
- Bollen, K., & Baten, K. (2010). Bilingual education in Flanders: Policy and press debate (1999-2006). *Modern Language Journal*, 94, 412-433.
- Breidbach, S., & Viebrock, B. (2012). CLIL in Germany: Results from recent research in a contested field of education. *International CLIL Research Journal*, 1(4). <http://www.icrj.eu/14/article1.html>
- 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). *International Journal of Applied Linguistics*, 32, 236-241.
- Brybaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, 7, 1116.
- Bulon, A., Hendriks, I., Meunier, F., & Van Goethem, K. (2017). Using global complexity measures to assess second language proficiency: Comparing CLIL and non-CLIL learners of English and Dutch in French-speaking Belgium. *Papers of the Linguistic Society of Belgium*, 11, 1-25.

- Buyl, A., & Housen, A. (2014). Factors, processes and outcomes of early immersion education in the Francophone community in Belgium. *International Journal of Bilingual Education and Bilingualism*, 17, 178-196.
- Cenoz, J., Genesee, F., & Gorter, D. (2014). Critical analysis of CLIL: Taking stock and looking forward. *International Journal of Applied Linguistics*, 35, 243-262.
- Chohey-Paquet, M. (2007). CLIL in Belgium (French-Speaking). In A. Maljers, D. Marsh, & D. Wolff (Eds), *Windows on CLIL. Content and Language Integrated Learning in the European spotlight* (pp. 25-32). Den Haag: European Platform for Dutch Education.
- Coyle, D. (2007). Content and language integrated learning: Towards a connected research agenda for CLIL pedagogies. *International Journal of Bilingual Education and Bilingualism*, 10, 543-562.
- Coyle, D., Hood, P., & Marsh, D. (2010). *Content and Language Integrated Learning*. Cambridge University Press.
- Coyle, D., Van Dusen-Scholl, N., & Hornberger, N. (2008). CLIL: A pedagogical approach from the European perspective. *Encyclopedia of language and education* (pp. 97-111). New York: Springer.
- Dalton-Puffer, C. (2008). Outcomes and processes in Content and Language Integrated Learning (CLIL): Current research from Europe. In W. Delanoy & L. Volkman (Eds.), *Future perspectives for English language teaching* (pp. 139-157). Heidelberg: Carl Winter.
- Dalton-Puffer, C., Nikula, T. & Smit, U. (2010). *Language use and language learning in CLIL Classrooms*. Amsterdam: John Benjamins.
- de Graaff, R., Koopman, G. J., Anikina, J., & Westhoff, G. (2007). An observation tool for effective L2 pedagogy in Content and Language Integrated Learning (CLIL). *International Journal of Bilingual Education and Bilingualism*, 10, 603-624.
- De Smet, A., Mettewie, L., Galand, B., Hiligsmann, P., & Van Mensel, L. (2018). Classroom anxiety and enjoyment in CLIL and non-CLIL: Does the target language matter? *Studies in Second Language Learning and Teaching*, 8, 47-72
- De Vlaamse Regering. (2014). *Besluit van de Vlaamse regering tot bepaling van de kwaliteitsstandaard voor Content and Language Integrated Learning (CLIL) in het gewoon secundair onderwijs en de leertijd en aanwijzingen van de bevoegde dienst voor de goedkeuring van de CLIL-plannen*. <http://data-onderwijs.vlaanderen.be/edulex/document.aspx?docid=14721>
- Dewaele, J.-M. (2005). Sociodemographic, psychological and politicocultural correlates in Flemish students' attitudes towards French and English. *Journal of Multilingual and Multicultural Development*, 26(2), 118-137.
- De Wilde, V., Brysbaert, M., & Eyckmans, J. (2019). Learning English through out-of-school exposure. Which levels of language proficiency are attained and which types of input are important? *Bilingualism: Language and Cognition*, 23(1), 171-185.

- EACEA. (2012). *Key data on teaching languages at school in Europe*. http://eacea.ec.europa.eu/education/eurydice/documents/key_data_series/143en.pdf
- Elgort, I., Brysbaert, M., Stevens, M., & Van Assche, E. (2018). Contextual word learning during reading in a second language: An eye-movement study. *Studies in Second Language Acquisition*, 40, 341-366.
- Ellis, N. (2002). Frequency effects in language processing: A review with Implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition*, 24, 143-188.
- Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley.
- Gassner, D., & Maillat, D. (2006). Spoken competence in CLIL: A pragmatic take on recent Swiss data. *Vienna English Working Papers*, 15, 15-22.
- Gierlinger, E. M. (2017). Teaching CLIL? Yes, but with a pinch of SALT. *Journal of Immersion and Content-Based Language Education*, 5, 187-213.
- Goethals, M. (1997). English in Flanders (Belgium). *World Englishes*, 16, 105-114.
- Housen, A., Janssens, S., & Pierrars, M. (2001). *Frans en Engels als vreemde talen in Vlaamse scholen*. Brussel: VUBPRESS.
- Hulstijn, J. H. (2003). Incidental and intentional learning. In C. Doughty & M. H. Long (Eds.), *The handbook of second language acquisition* (pp. 349-381). Oxford, UK: Blackwell.
- Hulstijn, J. H. (2010). Measuring second language proficiency. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 185-200). Amsterdam: Benjamins.
- Jexenflicker, S., & Dalton-Puffer, C. (2010). Comparing the writings of CLIL and non-CLIL students in higher colleges of technology. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds.), *Language use and language learning in CLIL classrooms* (pp. 169-189). Amsterdam: John Benjamins.
- Jiménez Catalán, R. M., Ruiz de Zarobe, Y., & Cenoz, J. (2006). Vocabulary profiles of English foreign language learners in English as a subject and as a vehicular language. *Vienna English Working Papers (VIEWS)*, 15, 23-27.
- Kremmel, B., & Schmitt, N. (2018). Vocabulary levels test. In J. I. Lontos, M. DeliCarpini, & J. C. Riopel (Eds.), *The TESOL Encyclopedia of English language teaching* (pp. 1-7). New York: Wiley-Blackwell.
- Laufer, B., & Nation, P. (1999). A vocabulary size test of controlled productive ability. *Language Testing*, 16, 33-51.
- Laufer, B., & Hulstijn, J. H. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22, 1-26.
- Lasagabaster, D. (2008). Foreign language competence in Content and Language Integrated Courses. *The Open Applied Linguistics Journal*, 1, 31-42.

- Lin, Angel, M. Y. (2016). *Language across the curriculum: CLIL in English as an additional language (EAL) contexts. Theory and practice*. Singapore: Springer.
- Lo, Y., & Macaro, E. (2015). Getting used to content and language integrated learning: What can classroom interaction reveal? *The Language Learning Journal*, 43, 1-17. <http://doi.org/10.1080/09571736.2015.1053281>.
- Lochtman, K., Lutjeharms, M., & Kermarrec, G. (2005). Langues étrangères à Bruxelles: Recherche sur les attitudes d'étudiants Bruxellois des écoles d'ingénieur commercial ULB et VUB. In E. Witte, L. Van Mensel, M. Pierrard, L. Mettwie, A. Housen, & R. De Groof (Eds.), *Language, attitudes and education in multilingual cities* (pp. 211-233). Brussels: Koninklijke Vlaamse Academie van België voor Wetenschappen en Kunst.
- Lorenzo, F., Casal S., & Moore, P. (2010). The effects of content and language integrated learning in European education: key findings from the Andalusian sections evaluation project. *Applied Linguistics*, 31, 418-442.
- MacIntyre, P., & Gregersen, T. (2012). Emotions that facilitate language learning: The positive-broadening power of the imagination. *Studies in Second Language Learning and Teaching*, 2, 193-213.
- Masrai, A., & Milton, J. (2018). Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement. *Journal of English for Academic Purposes*, 31, 44-57.
- Merikivi, R., & Pietilä, P. (2014). Vocabulary in CLIL and in mainstream education. *Journal of Language Teaching and Research*, 5, 487-497.
- Mettwie, L. (2015). Apprendre la langue de "l'Autre" en Belgique: La dimension affective comme frein à l'apprentissage. *Le Langage et l'Homme*, 1, 23-42.
- Mewald, C. (2007). A comparison of oral foreign language performance of learners in CLIL and mainstream classes at lower secondary level in Lower Austria. In C. Dalton-Puffer & U. Smit (Eds.), *Empirical perspectives on CLIL classroom discourse* (pp. 139-177). Frankfurt: Lang.
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review*, 63, 59-82.
- Navés, T. (2011). How promising are the results of integrating content and language for EFL writing and overall EFL proficiency? In Y. Ruiz de Zarobe, J. M. Sierra, & F. Gallardo del Puerto (Eds.), *Content and foreign language integrated learning. Contributions to multilingualism in European contexts* (pp. 155-186). Frankfurt am Main: Peter Lang.
- 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.
- Pérez, A., Lorenzo, F., & Pavón, V. (2016). European bilingual models beyond lingua franca: Key findings from CLIL French programs. *Language Policy*, 15, 485-504.

- Pérez-Cañado M. L. (2012). CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism*, 15, 315-341.
- Peters, E., Velghe, T., Van Rompaey, T. (2015, July). *A post-entry English and French vocabulary size for Flemish learners*. Paper presented at EALTA. Copenhagen, Denmark.
- Piesche, N., Jonkmann, K., Fiege, Ch., & J.-U. Keßler (2016). CLIL for all? A randomized controlled field experiment with sixth-grade students on the effects of content and language integrated science learning. *Learning and Instruction*, 44, 108-116.
- Rosnow, R., & Rosenthal, R. (2005). *Beginning behavioral research: A conceptual primer*. Englewood Cliffs, NJ: Pearson/Prentice Hall.
- Ruiz de Zarobe, Y. (2010). Written production and CLIL. An empirical study. In C. Dalton-Puffer, T. Nikula, & U. Smit (Eds), *Language use and language learning in CLIL classrooms* (pp. 191-209). Amsterdam: John Benjamins.
- Ruiz de Zarobe, Y. (2011). Which language competencies benefit from CLIL. In Y. Ruiz de Zarobe, J. Sierra, & F. Gallardo del Puerto (Eds), *Content and foreign language integrated learning. Contributions* (pp. 129-153). Bern: Peter Lang.
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18, 55-88.
- Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, 47, 484-503
- Serra, C. (2007). Assessing CLIL at primary school: A longitudinal study. *International Journal of Bilingual Education and Bilingualism*, 10, 582-602.
- Simon, E., Lima Jr, R., & De Cuypere, L. (2016). Acquiring non-native speech through early media exposure: Belgian children's productions of English vowels. *Poznan Studies in Contemporary Linguistics*, 52, 719-743.
- Strobbe, L., Sercu, L., Strobbe, J., & Welcomme, A. (2013). *Je vak in een vreemde taal? Wegwijzers voor de CLIL-onderwijspraktijk*. Leuven: Acco.
- Sundqvist, P., & Sylvén, L. K. (2016). *Extramural English in teaching and learning*. London: Palgrave Macmillan.
- Surmont, J., Struys, E., Van Den Noort, M., & Van de Craen, P. (2016). The effects of CLIL on mathematical content learning: A longitudinal study. *Studies in Second Language Learning and Teaching*, 6, 319-337.
- Sylvén, L. (2004). *Teaching in English or English teaching? On the effects of Content and Language Integrated Learning on Swedish learners' incidental vocabulary acquisition* (Unpublished doctoral dissertation). Göteborg University, Sweden.
- Van de Craen, P., Ceuleers, E., Lochtman, K., Allain, L., & Mondt, K. (2006). An interdisciplinary research approach to CLIL learning in primary schools in Brussels. In C. Dalton-Puffer & U. Smit (Eds.), *Empirical perspectives on*

- CLIL classroom discourse – CLIL: Empirische Untersuchungen zum Unterrichtsdiskurs* (pp. 253-274) Frankfurt: Lang.
- Van de Craen, P., Surmont, J., Mondt, K., & Ceuleers, E. (2011). Twelve years of CLIL practice in multilingual Belgium. In G. Egger & C. Lechner (Eds.), *Primary CLIL around Europe: Learning in two languages in primary education* (pp. 81-97). Marburg: Tectum Verlag.
- Webb, S., & Rodgers, M. P. H. (2009). The lexical coverage of movies. *Applied Linguistics*, 30, 407-442.
- Wiesemes, R. (2009). Developing theories of practices in CLIL: CLIL as post-method pedagogies? In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (Eds.), *Content and language integrated learning. Evidence from research in Europe* (pp. 41-59). Bristol: Multilingual Matters.
- Willemyns, R. (2002). The Dutch-French language border in Belgium. *Journal of Multilingual and Multicultural Development*, 23, 36-49.
- Wolff, D. (2002). On the importance of CLIL in the context of the debate on plurilingual education in the European Union. In D. Marsh (Ed.), *CLIL/EMILE. The European dimension. Actions, trends, and foresight potential* (pp. 47-48). Jyväskylä: University of Jyväskylä.
- Zydatiś, W. (2007). Bilingualer Fachunterricht in Deutschland: eine Bilanz. *Fremdsprachen Lehren und Lernen*, 36, 8-25.