Gutiérrez Ortiz, Mar. 2020. Assessing the Development of Second Language Syntax in Content and Language Integrated Learning. *Revista de lenguas para fines específicos 26.2*, pp. 111-130 ·
http://dx.doi.org/XXXXXXXXXXXXX
https://ojsspdc.ulpgc.es/ojs/index.php/LFE/index

Paper received: 30 April 2019
Paper received in revised form and accepted for publication: 7 October 2019

Assessing the development of second language syntax in Content and Language Integrated Learning

Mar Gutiérrez Ortiz[1] · https://orcid.org/0000-0002-1063-0933

Universidad de Sevilla

Departamento de Literatura Inglesa y Norteamericana. Facultad de Filología. C/ Palos de la Frontera, S/N. 41004 Sevilla.

ABSTRACT

This study examined the effect of the Content and Language Integrated Learning (CLIL) approach on the English proficiency and syntax of a group of Andalusian learners. High-school students (n = 22) enrolled in CLIL and non-CLIL classes in the same school (I.E.S. Mariana Pineda) took an English proficiency test adapted from the University Entrance Examination. Students were assessed for syntactic development in English using an Elicited Imitation (EI) task in conjunction with an experimental design based on the syntactic properties of English. Pre-intervention academic achievement, socio-economic status (SES), and gender were included as covariates in the analyses. CLIL students scored significantly higher than non-CLIL students on the syntactic task ($p = 0.002$) and the proficiency test ($p < 0.001$). Results revealed a significant positive correlation between the scores obtained in the proficiency test and the EI task for the CLIL group ($p < 0.01$), but not for the non-CLIL group ($p = 0.39$). These findings advance a methodology to assess the English syntax of CLIL students.

Keywords: CLIL, second language assessment, syntax, L2 proficiency

RESUMEN

Este estudio examinó la efectividad del Aprendizaje Integrado de Contenidos y Lenguas Extranjeras (AICLE) respecto al dominio del inglés y la incorporación de su sintaxis en un grupo de estudiantes andaluces. Alumnos de instituto (n = 22) matriculados en clases AICLE y clases no-AICLE del mismo centro educativo (I.E.S. Mariana Pineda) realizaron un examen de inglés adaptado de la Prueba de Evaluación de Bachillerato para el Acceso a la Universidad. Se evaluó el grado de desarrollo de la sintaxis del inglés en los alumnos empleando una prueba de Imitación Elicitada (EI) en conjunción con un diseño experimental basado en las propiedades sintácticas del inglés. Se incluyeron como covariables en el análisis el nivel académico y socioeconómico previos a la intervención, así como el sexo del alumnado. Los alumnos AICLE obtuvieron mejores resultados que los no-AICLE en la prueba sintáctica ($p = 0.002$) y en el examen de inglés ($p < 0.001$). Los resultados evidenciaron una correlación significativa positiva entre la puntuación obtenida en el examen inglés y en la prueba EI para el grupo AICLE ($p < 0.01$), pero no para el no-AICLE ($p = 0.39$). Estos resultados proponen una metodología para evaluar la sintaxis inglesa de los estudiantes CLIL.

Palabras clave: AICLE, evaluación de la L2, sintaxis, dominio de la L2

1. Introduction

Ever since Content and Language Integrated Learning instruction (henceforth CLIL) was implemented across Spain, a large body of research has evaluated its impact on the students' acquisition of the target language. Numerous studies have found a consistent advantage held by CLIL students over their non-CLIL peers in the command of English as a second language (henceforth L2). However, the analyses of the students' elicited narrations and written productions have obtained inconclusive results about the development of English morphosyntactic structures by learners immersed in CLIL classrooms when compared to those only taking

[1] **Corresponding author** · Email: margutort@alum.us.es

English-as-a-foreign-language lessons (henceforth EFL) (see Ruiz de Zarobe, 2015 for a review on the effect of the CLIL approach). Because the methodological shortcomings of some of these previous papers compromised the validity of their findings (Pérez Cañado, 2016), a recent longitudinal study has been conducted comparing homogeneous samples of CLIL and non-CLIL strands (Pérez Cañado, 2018). CLIL students have been shown to surpass their non-CLIL peers in every linguistic skill, especially upon completion of compulsory education (Pérez Cañado, 2018). Nevertheless, there is still a need for research studies that statistically guarantee the homogeneity of treatment and control groups in the comparison of specific language areas, such as morphology and syntax.

The central objective of this study was to investigate the effect of participation in a CLIL program on the students' implicit knowledge of the L2. The study was conducted in a Spanish high school located in Andalusia, a traditionally monolingual region where CLIL programs have been implemented in the vast majority of public schools to foster the acquisition of English as an L2. In addition to an Elicited Imitation (henceforth EI) task designed to evaluate the development of language syntax in the acquisition of English as a foreign language, students took a standardized exam that measured their proficiency. In order to ensure the estimates of the effect were not due to self-selection bias, socio-economic status (henceforth SES) and academic achievement (measured as Grade Point Average) were included as covariates in the study. This paper will inform us about the development of grammar in CLIL, and by doing so, a new approach is simultaneously advanced for measuring English language acquisition. The results suggest that direct assessment of students' syntactic knowledge through the EI task may be an effective approach to evaluating high school CLIL programs in different educational contexts.

2. Literature review

2.1. The CLIL advantage

A general advantage in English proficiency has been repeatedly reported for learners enrolled in CLIL programs when compared to learners who only receive EFL lessons (Dalton-Puffer, 2008). In the Spanish context, CLIL students have been shown to surpass non-CLIL students in reading comprehension (Lorenzo, Casal & Moore, 2010; Navés, 2011; Pérez-Vidal & Roquet, 2015; Prieto-Arranz, Rallo Fabra, Calafat-Ripoll & Catrain-González, 2015), vocabulary (Jiménez-Catalán & Ruiz de Zarobe, 2009; Moreno Espinosa, 2009), fluency, and lexical and syntactic complexity in written productions (Gené-Gil, Juan-Garau, & Salazar-Noguera, 2015; Lasagabaster, 2008; Lorenzo et al., 2010; Navés & Victori, 2010; Pérez-Vidal & Roquet, 2015; Ruiz de Zarobe, 2010). CLIL has also been shown to improve oral production (Lasagabaster, 2008; Lorenzo et al., 2010; Ruiz de Zarobe, 2008) and metacognitive awareness for selecting appropriate learning strategies (Ruiz de Zarobe & Zenotz, 2012, 2015). Research has obtained contradictory results for listening comprehension: whereas Navés (2011) and Pérez-Vidal and Roquet (2015) did not find a difference between CLIL and non-CLIL learners in the Catalonian context, Lorenzo et al. (2010), Lasagabaster (2008) and Prieto-Arranz et al. (2015) found a positive effect of CLIL on listening comprehension.

A number of researchers have questioned the validity of these previous findings (Pérez Cañado, 2016) on the grounds that CLIL groups are not comparable to the control groups used in the experiments due to self-selection bias (Bruton, 2011a, 2011b, 2013, 2015). It is not unreasonable to suspect that SES, academic performance or motivation may predispose students to choose a CLIL program, and that these very same reasons may also have

an impact on their proficiency. Therefore, the advantage of CLIL students may not only be due to their participation in the CLIL program, but to pre-existing differences between groups. In fact, a study conducted in the Catalonian context found that CLIL students from grades 7 and 9 performed equally or better on writing and overall proficiency tasks than non-CLIL learners who were enrolled in higher grades (Navés & Victori, 2010). Over time, the differences were ameliorated, which shows that the initial results may have reflected a pre-existing imbalance. To fill in this research gap, Pérez Cañado (2018) has conducted a longitudinal study with 2,024 students in three monolingual communities: CLIL students have been shown to outperform their non-CLIL peers on all the linguistic skills assessed, even when controlling for motivation, verbal intelligence, extramural exposure to English, setting, and SES.

The few studies that have compared the grammatical development of CLIL and non-CLIL groups in Spain have made use of natural communication. By asking students to retell the picture story "Frog, where are you?", they have obtained disparate results about different aspects of the students' morphosyntax. In a longitudinal analysis of the oral narratives of Basque/Spanish bilinguals, Ruiz de Zarobe (2008) found that CLIL groups outperformed non-CLIL groups in grammatical accuracy. Villarreal and García Mayo (2009) demonstrated that CLIL students omitted affixal –s and –ed less than non-CLIL students, but they did not find significant differences between the omission rate of suppletive forms. Martínez Adrián and Gutiérrez Mangado (2009) found significant differences between the number of placeholders and embedded clauses used by CLIL and non-CLIL students, but not between the production of null subjects, null objects, and correct negative structures. Lázaro Ibarrola (2012) demonstrated that CLIL students produce higher rates of inflected verbs, higher correction rates in the use of pronouns, and a higher total number of pronouns and subordinate sentences.

By eliciting the narration of the picture story, all these studies compared the morphosyntactic structures that spontaneously emerge in the students' discourse. However, the use of the L2 in oral production is not only determined by the implicit knowledge of the language, but also by external factors such as the characteristics of the language task, the personality and socio-psychological features of the learner, and the features of the pedagogical intervention (Housen & Kuiken, 2009). In fact, a study comparing successive and simultaneous bilinguals found that the children's performance in the narration task was not always correlated with their actual knowledge of the grammar (Kim, Park & Lust, 2018). In turn, the EI task allows us to measure the acquisition of specific grammatical aspects by asking participants to reproduce sentences "that are designed to differ only with respect to a particular grammatical factor" (Lust et al., 1996: 57). In L2 acquisition research, this task has proven valid, reliable and fit for discriminating between learners of different proficiencies (Gaillard & Tremblay, 2016).

The present paper makes use of an EI task modelled after Flynn's (1986) findings about the constraints that the principal branching direction of a language imposes on the acquisition of syntax to measure the impact of CLIL programs on the development of English subordinate clauses. The results obtained are compared to the students' scores in a standardized English test adapted from the University Entrance Examination.

2.2. Principal Branching Direction

In recent decades, a vast number of theories have tried to account for the nature of language development and the role of input in the process of language acquisition (see Lust, 2011 for a review). Research has provided evidence that Universal Grammar (UG), the innate human language faculty, remains intact and constantly accessible during the process of first and second language acquisition (Epstein, Flynn & Martohardjono, 1996).

To account for the development of the specific languages despite the invariability of UG, Epstein et al. (1996) highlight the need to draw a distinction between the innate language faculty and the grammar of specific languages. This view is fleshed out in the Grammatical Mapping (GM) paradigm, which reconciles the role of Universal Grammar (UG) and input in the process of language acquisition (Lust, 2012). According to GM, children use UG to develop Specific Language Grammar (SLG), after being exposed to the data of a particular language (Lust, 2012). UG is constituted by principles, invariant properties common to all languages, and parameters, binary switches that are set to one of the values depending on the linguistic input (Chomsky, 1980). Principles account for language universals, whereas parameters account for language variation. L2 learners are guided by UG to develop the L2 from exposure, so they need to reset the parameters in the creation of the L2 grammar (Epstein et al., 1996). Since the knowledge of previous languages also plays a role in subsequent language acquisition (Flynn, Foley & Vinnitskaya, 2004), the SLG of the L2 is determined by UG, influenced by the SLG of the first language, and computed over the L2 data.

In the creation of the SLG of a first or second language, binary parameters are set to one of the values after exposure to the data and determine certain fundamental properties of the language grammar. The principal branching direction (PBD), also called the head initial/head final parameter, is a parameter that classifies languages according to their predominant branching direction.

> Principal Branching Direction refers to the branching direction which holds consistently in unmarked form over major recursive structures of a language, where "major recursive structures" are defined to include embeddings of sentence complements under either NP or S "heads." Specifically, relative clauses in complex NP and adverbial subordinate clauses are critical to the definition of this parameter. (Lust et al., 1995, p. 199)

Thus, right-branching languages (head-initial languages) tend to include embedding to the right of their heads, i.e. the unmarked position for subordinate clauses is to the right of the sentence (Lust et al., 1995). Although not all languages have a perfectly consistent branching direction, and all languages appear to allow alternations in the order of components (Flynn, 1983), some languages, like English, allow head-initial characterization, as exemplified in (1) and (2) (Flynn & Espinal, 1985):

(1) [The child [who is eating rice]] is crying
(2) [The child drank the milk [after he ate the rice.]]

Research has shown that in early child language, the development of grammatical anaphora is determined by the principal branching direction of the language. In English, the acquisition of forward anaphora (3) precedes the acquisition of backward anaphora (4), because it coheres with the head-initial configuration of the language (Lust, 1981):

(3) John read the play while he smoked a pipe.
(4) While he smoked a pipe, John read the play.

In the same way, Flynn (1983) has demonstrated that the PBD is also a structural constraint in the acquisition of English as an L2. Thus, Spanish speakers learning English as an L2 are sensitive to the head-direction of the L2 and prefer forward to backward anaphora in complex sentences. Flynn (1986) showed that Spanish L2 learners of English acquired complex sentences with postposed embedded clauses and forward anaphora, such as in (5), before complex sentences with preposed embedded clauses and backward anaphora in (6):

(5) The man answered the boss when he installed the television.
(6) When he entered the office, the professor questioned the man.

Beginners did not show significant differences between the two types of sentences because they had not acquired any of them. Intermediate students mastered significantly more sentences with postposed clauses and forward anaphora (5) than sentences with preposed clauses and backward anaphora (6). Advanced students had command over both types. In the present study, Flynn's findings are used to assess the grammatical knowledge of students in CLIL and non-CLIL groups. The amount of correct imitations that students produce for each type of sentence indicates their level of the L2.

2. Rationale and design

The present study exemplifies the use of a comprehensive methodology to measure the effect of the CLIL approach on the students' level of English as an L2. CLIL and non-CLIL students are compared for the scores obtained in a proficiency test and on an EI task that taps into their grammatical knowledge through an experimental design. The research design was cross-sectional and quasi-experimental as the assignment to each condition (intervention vs. no intervention) was by self-selection. Since quasi-experimental designs lack random assignment, a comparison group was identified that was as similar as possible to the experimental group in terms of pre-intervention characteristics. Thus, the CLIL group was compared with 12th-grade students that attend the non-CLIL program in the same high school. To ensure the comparability between groups, SES, and academic achievement before the intervention were included as covariates.

3.1. Elicited Imitation Task

The EI task is an experimental task used in conjunction with experimental designs to assess grammatical knowledge (Lust, Flynn & Foley, 1996). In the EI task, the researcher reads a sentence aloud and asks the participant to repeat it according to standardized methods for administration. This task is based on the assumption that in order to repeat the sentence, the participant must analyze the structure of the sentence they have heard (including vocabulary, syntax and meaning), and then reconstruct this structure in their production of the model. The stimulus sentences must be too long to be stored in the short-term memory without being analyzed (Flynn & Espinal, 1985). So in order to "imitate" a structure, the structure must be part of the speaker's grammatical competence (Lust et al., 1996). The researcher can analyze the participants' knowledge of specific aspects of the grammar by modeling sentences that vary only in critical factors with all others held constant.

The critical factors that vary in our experimental design are based on Flynn's findings (1986): our two groups had to repeat a series of complex sentences with adverbial subordinate clauses and pronoun anaphora. The direction of the pronoun anaphora was coherent with the position of the subordinate clause in relation to the main clause. Thus, in postposed clauses the antecedent preceded the pronoun (forward anaphora), and in the preposed clauses the pronoun preceded the antecedent (backward anaphora):

(7) Postposed clause with forward anaphora: The man answered the boss when he installed the television.
(8) Preposed clause with backward anaphora: When he installed the television, the man answered the boss.

Because English is a right-branching language, postposed clauses with forward anaphora (7) are unmarked, and so they are acquired before preposed clauses with backward anaphora (8). This developmental pattern was used to determine the students' level of syntax by computing not only the total number of sentences imitated, but also the type of sentences that each group produced correctly. In line with previous findings (Flynn, 1986), significant differences between (7) and (8) are only expected for the non-CLIL group (equivalent to Flynn's (1986)

intermediate group). If the CLIL group has already acquired the two types of sentences, they will produce a similar number of correct imitations for all of them.

Even though Flynn (1983) did not find an independent effect of anaphora or branching direction on the number of correct imitations, a third type of sentence with preposed subordinate clause and forward anaphora (9) has been included to control whether the direction of the anaphora (backward vs. forward) alone had an effect on the number of correct imitations within left-branching clauses and whether the position of the clause (preposed vs. postposed) had an effect on the production of sentences with forward anaphora:

(9) Preposed clause with forward anaphora: When the doctor received the results, he called the gentleman

The independent variables are the program (CLIL or non-CLIL) X gender with covariates, SES and Grade Point Average (GPA). The dependent variable is the total number of correct imitations. To measure the difference in the number of correct imitations for each type of sentence, we used a generalized mixed model with subject as a random effect and type of sentence, gender, SES, and GPA as fixed effects.

3.2. Proficiency Test

The students' English performance was assessed using the mandatory University Entrance examination. This test is designed to evaluate the minimum level that all 12th-grade students need to achieve at the end of secondary education. Because passing the test contributes to the students' access to higher education, the results of the present study are informative about the potential impact of CLIL on the students' future careers. In addition to learning about the effect of CLIL on the students' English proficiency, we intend to gain further insight into the effect of CLIL on young people's potential to enter a globalized job market.

The independent variables are the program (CLIL or non-CLIL) X gender with covariates, SES, and GPA. The dependent variable is the score obtained in the proficiency test. To check the effect of CLIL participation on the students' performance in each part of the test, we used the score obtained in the different parts of the tests as dependent variables.

3.2. Hypotheses

The first hypothesis of this study is that students in the CLIL group would outperform students in the non-CLIL group in the proficiency test and the EI task. The second hypothesis is that students in the CLIL group would score similarly for the three types of complex sentences in the EI imitation task, given that previous research has shown that their English language is highly developed. In turn, non-CLIL students, who have not attained full English proficiency, are expected to be weaker on complex sentences with preposed clauses and backward anaphora, showing thus some of the intermediate steps in learning that have been documented in the literature. The third hypothesis is that if both the standardized proficiency tests and the EI test measure language knowledge, then the results of the proficiency test and the EI task would be strongly correlated.

4. Method

4.1. Program description

We investigated the achievements of the students in I.E.S. Mariana Pineda, a public high school located in Montequinto, a town near Seville (the main city in Andalusia). The school offers a CLIL program from 1º E.S.O. (7th grade) through 2º de Bachillerato (12th grade). The content that courses teach in the foreign language vary every year:

- 1º E.S.O (7th grade): Music, Natural Sciences, and Social Sciences.
- 2º E.S.O (8th grade): Technology, Music, Natural Sciences and Social Sciences.
- 3º E.S.O (9th grade): Technology, Natural Sciences, and Social Sciences.
- 4º E.S.O (10th grade): Social Sciences, Ethics and Integrated Project.
- 1º Bachillerato (11th grade): Science of Contemporary World, Philosophy and Integrated Project.
- 2º Bachillerato (12th grade): Integrated Project.

All students attended EFL during six years of primary education. In secondary education, the experimental group attended CLIL content classes (CLIL) and EFL, whereas the control group only learned English during the EFL lessons (Table 1).

	1º E.S.O. EFL/CLIL		**2º E.S.O. EFL/CLIL**		**3º E.S.O. EFL/CLIL**		**4º E.S.O. EFL/CLIL**		**1º Bac. EFL/CLIL**		**2º Bac EFL/CLIL**		**TOTAL**
Non-CLIL	4		4		4		4		3		3		22
CLIL	4	8	4	11	4	8	4	6	3	6	3	1	62

Table 1. Number of CLIL and EFL hours that each group receives weekly during secondary school.

4.2. Participants

The participants were 22 Spanish students between the ages of 17 and 18. The CLIL group (n = 11) had been attending a CLIL program for 6 years, although one of the students had only been in the program for 2 years. The non-CLIL group (n = 11) received all content lessons in Spanish and EFL three hours a week. All the parents of underage students and 18-year-old students completed consent forms before testing began. No monetary compensation was given for students' participation in this study.

The differences between groups were controlled with a linguistic and demographic questionnaire. In Spain, the fact that both groups attend the same public high school largely determines their SES. Even so, students provided their parents' occupation and highest academic attainment. SES was obtained from the fathers' occupation and education using the International Socio-Economic Index of Occupational Status (Ganzeboom, De Graaf & Treiman, 1992). Since the students' GPA prior to entering the program could not be retrieved, the grades obtained in the first semester of high school were used as covariates. High school grades usually reflect academic performance better than elementary school grades, which take into account attitudinal factors that are irrelevant to the present study. In addition, at that point, students had only attended three CLIL classes for three months (mid-September to mid-December). Because we did not have pre-intervention scores for English proficiency,

the students completed a survey about their linguistic background, which did not reveal significant differences between the groups.

4.3. Materials

4.3.1. Language Questionnaire

A questionnaire adapted from the Virtual Linguistics Lab Child Multilingualism Questionnaire (Lust & Blume, 2016) was used for gathering linguistic and demographic background information of students. Students provided the age and context of L2 and L3 acquisition, as well as information about their use of the language outside the classroom. Information was gathered about the parents' education and occupation prior to enrollment in the program. Students reported their GPA from the first year of compulsory education.

4.3.2. Elicited Imitation Task

This experimental task was adapted from Flynn (1986) and consisted of a battery of 9 sentences (3 iterations for each type) to specifically examine the learners' knowledge of embedded clauses in English (see appendix I). Participants were asked to repeat the sentences one by one after the experimenter. All the sentences were matched for length (15 syllables). The stimuli were counterbalanced across trials using an online randomizer.

4.3.3. Proficiency Test

The proficiency test was adapted from the reserve English test in the University Entrance Examination of September 2014. The exam was shortened to one hour, so students had to read a text, answer four reading comprehension questions (instead of five), complete 11 Use of English exercises (vocabulary and grammar), and write a text of 100 words (instead of 150 words).

4.4. Procedure

Two teachers in Mariana Pineda high school informed students that they would have the opportunity of participating in a study that would evaluate the effect of the CLIL program on the acquisition of English as an L2. Students that expressed willingness to participate received consent forms. Once the consent forms had been signed, the researcher attended the school to answer specific questions about the procedure. During the same session, the researcher examined students on the vocabulary of the EI task, and asked them to memorize the words they did not know. After a short time interval, a specific computerized link for the proficiency test was assigned to the participants. In the first part of the experiment, participants completed the proficiency test during the Computer Science class. Students were not allowed to consult any extra materials and only questions about the format of the exam were answered. The exam was proctored through collaboration between teachers and researcher.

After an average of three days, students were called individually to do the second part of the experiment. First, students were administered an online questionnaire about demographic information and previous experience with the language. Then they listened to the list of complex sentences in random order and repeated them one

by one. Their answers were recorded with Audacity. The recordings and their transcriptions were coded by researchers trained in language analysis according to standardized procedures.

4.5. Coding

4.5.1. Language Questionnaire

The highest academic qualification attained by both parents was divided into five levels: 1- None/Primary, 2- Secondary/Intermediate professional training, 3- Pre-university/Advanced professional training, 4- Tertiary and 5- Postgraduate. In addition, the occupation of the father was coded using the ISEI scores for Occupation categories (Ganzeboom et al., 1992), which transforms the father's education and occupation into SES. The ISEI score was standardized for the descriptive tables and centered for the statistical analyses. The students reported their GPA from the first year of secondary education on a scale of 0 to 10, and this score was centered for the statistical analyses.

4.5.2. Elicited Imitation Task

In order for an EI response to be scored correct (1 point), the participant had to repeat the stimulus sentence given without any major syntactic or semantic change. Lexical substitutions that did not imply a substantial change in the meaning of the sentence were not considered a semantic error. For example, the substitution of *gentleman* for *man* did not constitute an error, but the substitution of *actor* for *lawyer* did. Incidental changes made by the participant that did not alter the syntactic structure or meaning in ways that are relevant to the focus of this research, such as singular to plural or present tense to past tense, were not scored as incorrect, and neither were mispronunciations. Changes that altered the syntactic structure or meaning of the original stimulus sentences (e.g., repetition of only one clause, alteration of the original anaphora/antecedent relation, change in clause order, or repetition of a lexical item not considered to be a synonym) were scored as incorrect (0 points). Because there were three iterations of each type of sentence, the maximum score students could obtain was 9. The researcher transcribed all the sentences. Reliability was determined by having a native Spanish speaker who was bilingual in English and a native English speaker code all the data. Agreement between coders was 100%.

4.5.2. Proficiency test

The guidelines established for the correction of the University Entrance examination were adapted for this proficiency test. In the reading comprehension section, students could obtain up to 3 points, 1.5 for two multiple-choice questions and 1.5 for two true or false questions. The 11 Use of English questions were worth 4 points. A specific set of guidelines was taken into account for the writing task. The score (up to 3 points) was based on grammatical correction (repeated mistakes were only taken into account once), lexical richness and accuracy, and textual and communicative aspects. The maximum total score in the proficiency test was 10. To ensure scoring reliability, the composition was scored by the researcher and a research assistant. Disagreements were resolved by discussion, and in the event that consensus could not be reached, the two grades were averaged.

5. Results

This chapter presents the results of the statistical analyses conducted for the study. The first section elucidates the students' linguistic background and demographic information. Then the results of the proficiency test, the EI task, and the correlation between them are reported.

Before all the analyses, assumption checks were performed. Normality of residuals was checked by graphical methods (normal q-q plot). A plot of the residuals against the predicted value revealed fairly linear relationships between continuous data. There was no discernable pattern to the Scale Location plots (square root of the standardized residuals vs. fitted values), so the homoscedastic error assumption was not violated. Neither the Cook's D plot nor the Residuals vs. Leverage plot pointed to the presence of influential outliers.

5.1. Demographic and linguistic information

The questionnaire revealed the students' homogeneous background. All of them are native speakers of Spanish who started attending EFL classes (3 hours a week) in the first year of primary education (age 6-7), and 16 students (CLIL = 9 and non-CLIL = 7) took French in secondary education. None of the students had spent more than one week in an English-speaking country.

GROUP	GENDER	Extracurricular Mean (SD)	SES (z-score) Mean (SD)	GPA Mean (SD)
Non-CLIL	Male (n = 8)	0.37 (0.99)	-0.25 (1.12)	6.85 (0.65)
	Female (n = 3)	0.00	0.64 (1.52)	6.54 (0.42)
	Both genders (n = 11)	0.27 (0.86)	0.008 (1.22)	6.77 (0.61)
CLIL	Male (n = 6)	3.5 (3.54)	-0.28 (0.66)	7.17 (0.84)
	Female (n = 5)	1.6 (3.2)	0.36 (8.38)	7.69 (0.82)
	Both genders (n = 11)	2.63 (3.52)	-0.008 (0.78)	7.41 (0.87)

Table 2. Summary of the background information of the students by group and gender.

Students in the CLIL group attended extracurricular English classes for more years than students in the non-CLIL group before entering secondary education, but the results of a t-test did not indicate significant differences between groups ($t(20) = 1.58, p = 0.13$). In addition, parents in the CLIL group had a slightly higher education level (Figures 1 and 2). However, according to the ISEI class scheme, students in the CLIL class did not have significantly higher SES than students in the non-CLIL class, $t(20) = 0.37, p = 1.00$. Although students in the CLIL group had slightly higher GPA than students in the non-CLIL group at the onset of the intervention, the difference was not significant ($t(20) = 1.9, p = 0.70$).

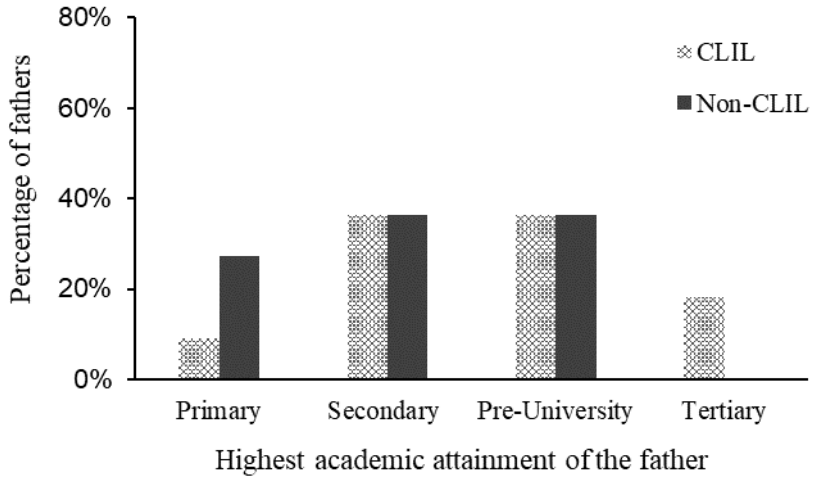


Figure 1. Highest academic achievement of the fathers in the CLIL and non-CLIL groups.

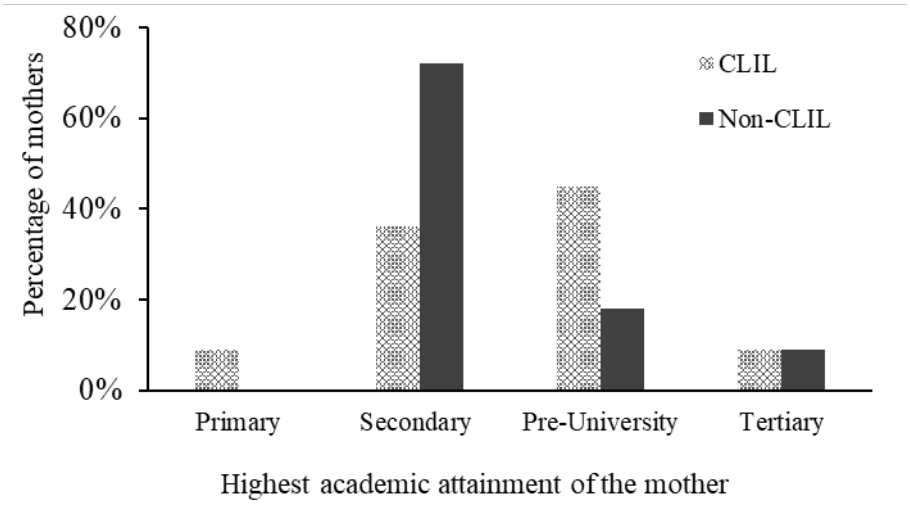


Figure 2. Highest academic achievement of the fathers in the CLIL and non-CLIL groups.

5.2. CLIL effects on syntactic development

In keeping with the design, grammatical language knowledge was tested through an EI task with the experimental design explained above. Students in the CLIL group surpassed students in the non-CLIL group for each type of sentence (Table 3). As predicted, students repeated more sentences with postposed clauses and forward anaphora than sentences with preposed clauses and backward anaphora. Three was the maximum score students could obtain for each type of sentence, and nine the maximum overall score (Table 3).

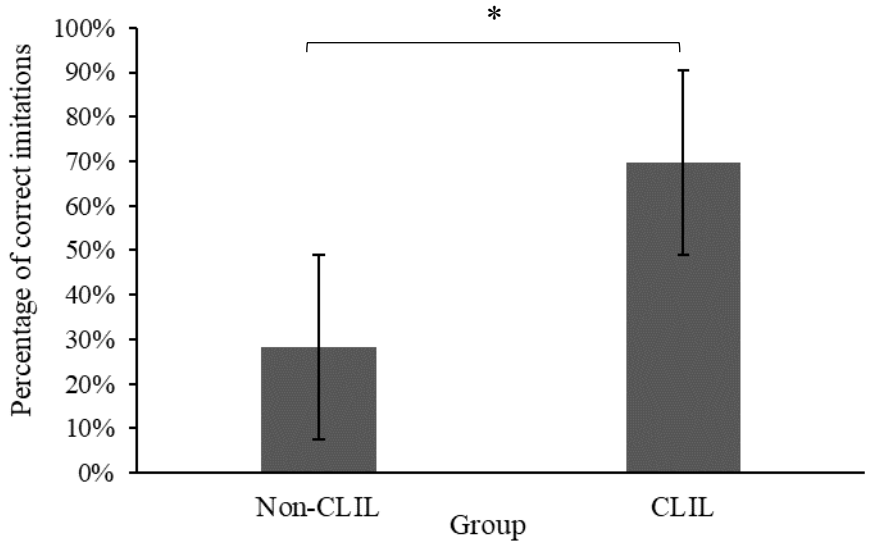
GROUP	Preposed backward Mean (SD)	Preposed forward Mean (SD)	Postposed forward Mean (SD)	Total Mean (SD)
Non-CLIL	0.45 (0.66)	0.91 (0.79)	1.18 (0.94)	2.54 (1.78)
CLIL	2 (1.04)	2.09 (1)	2.18 (0.94)	6.27 (2.6)

Total	1.23 (1.17)	1.5 (1.07)	1.68 (1.06)	4.41 (2.9)

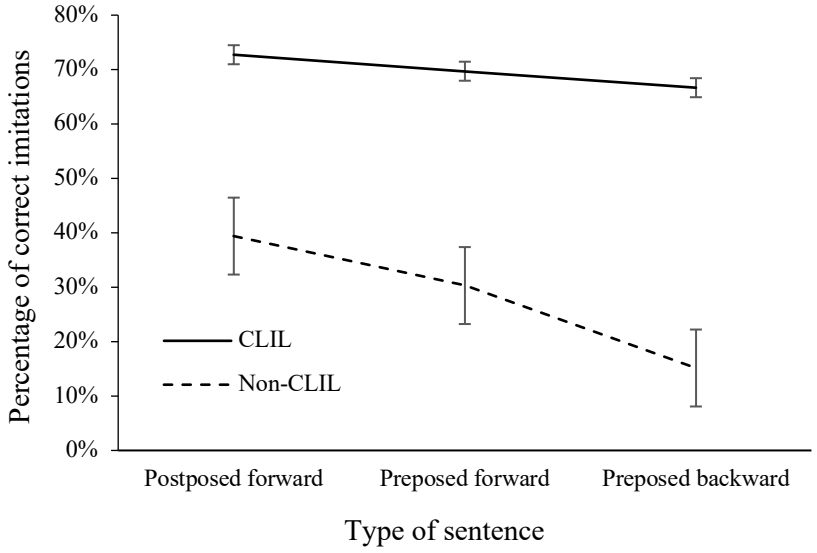
Table 3. Mean score obtained by students of both groups in each type of the EI task and total score.

A generalized Analysis of Covariance (ANCOVA) indicated a significant effect of the intervention on the total score of the EI task ($F(1, 17) = 12.15, p = 0.003$), Fig. 3. However, there was no effect of gender ($F(1, 17) = 0.29, p = 0.6$), GPA ($F(1, 17) = 0, p = 0.98$) or SES ($F(1, 17) = 0.04, p = 0.85$).

Figure 3. Percentage of total sentences correctly imitated by students in the CLIL and non-CLIL groups.



The number of correct imitations for each type of sentence was analyzed using a generalized mixed model where gender, GPA, SES and type of sentence were included as fixed effects, and student ID as a random effect. Overall, the scores of both groups were higher for the sentences that are acquired earlier, i.e. sentences with postposed clauses and forward anaphora in right-branching languages, Fig. 4. A marginal effect of type ($F(2,$



$42) = 2.63, p = 0.08$) and a significant difference between preposed backward and postposed forward sentences ($t(42) = 2.28, p = 0.03$) were found.

Figure 4. Percentage of each type of sentence correctly imitated by the CLIL and the non-CLIL groups.

In addition, even though the intervention by type interaction was not significant ($F(2, 40) = 0.97$, $p = 0.39$), a pairwise comparison was performed to test the hypothesis that the difference between complex sentences with preposed clauses (backward anaphora) and complex sentences with postposed clauses (forward anaphora) would only be significant for the non-CLIL group because they are at an earlier stage in the development of the language. As predicted, only students in the non-CLIL group differed significantly in the number of correct imitations for the two types of sentence ($t(40) = -2.58$, $p = 0.04$). The fact that no significant differences were found between the two types of sentences with preposed subordinate clauses ($t(42) = 1.37$, $p = 0.18$) or the two types of sentences with forward anaphora ($t(42) = 0.91$, $p = 0.37$), confirms that anaphora direction or clause position alone do not have an effect on the number of correct imitations.

In sum, students in the CLIL group produced significantly more correct imitations in the EI task than their counterparts in the non-CLIL class. In addition, students in the non-CLIL group produced significantly fewer sentences with preposed clauses and backward anaphora than sentences with postposed clauses and forward anaphora. This difference was not observed for students in the CLIL group, which confirms our prediction that CLIL students were able to produce all types of sentences because their English grammar is more advanced.

5.3. CLIL effects on proficiency

GROUP	Reading (max 3) Mean (SD)	Use of English (max 4) Mean (SD)	Writing (max 3) Mean (SD)	Overall (max 10) Mean (SD)
Non-CLIL	1.57 (0.93)	1.27 (0.58)	0.84 (0.55)	3.70 (1.56)
CLIL	2.52 (0.58)	3.05 (0.58)	2.30 (0.53)	8.00 (1.26)
Total	2.05 (0.91)	2.16 (1.06)	1.57 (0.91)	5.85 (2.57)

Table 4. Mean score obtained by students of both groups in each part of the proficiency test and total score

Table 4 shows that students in the CLIL group outscored their peers in the non-CLIL group in all the parts of the proficiency test. A generalized ANCOVA was conducted to determine a statistically significant difference between CLIL and non-CLIL students on the overall score of the proficiency test and scores of every part of the proficiency test including gender, GPA, and SES as covariates. There was a statistically significant main effect of CLIL on the overall score of the proficiency test, $F(1, 17) = 44.28$, $p < .001$, Fig. 5. Neither gender, GPA nor SES had a significant effect on the proficiency scores (all *ps*. < 0.05).

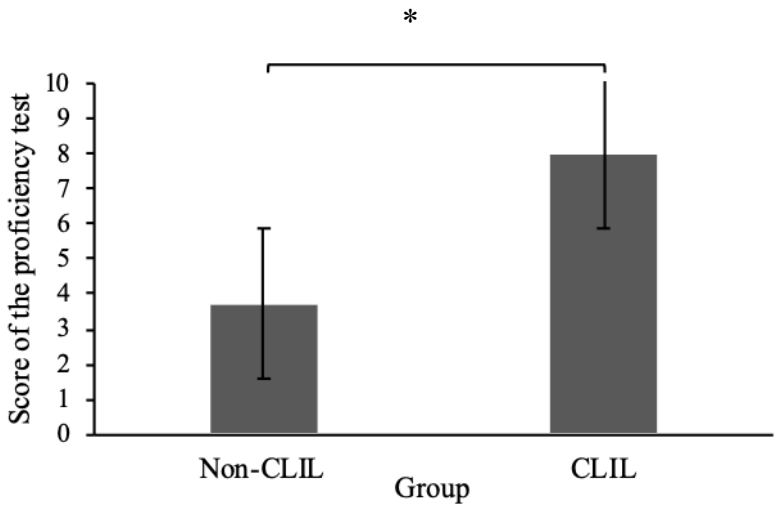


Figure 5. Mean score obtained by students in the CLIL and non-CLIL groups in the proficiency test.

A significant effect of intervention was found for reading comprehension ($F(1, 17) = 7.55$, $p = 0.01$), Use of English ($F(1, 17) = 42.91$, $p < 0.001$) and written production ($F(1, 17) = 44.06$, $p < 0.001$. In sum, the CLIL group scored significantly above the non-CLIL group in the three parts of the test (reading, writing and Use of English).

5.4. Correlation between proficiency and syntactic development

To assess the relationship between the scores obtained by students in both tasks, a Pearson product-moment correlation coefficient was obtained to measure the strength of the linear association between the proficiency scores and the EI scores. There was a strong positive correlation between the results of the proficiency test and the results of the overall EI task for the CLIL group, $r(9) = 0.71$, $p < 0.01$. However, the correlation was not significant for the non-CLIL students ($r(9) = 0.28$, $p = 0.39$).

5.5. Overall difference between CLIL and non-CLIL students

The students' responses to the linguistic and demographic questionnaire did not reveal significant differences between groups; the participants were native speakers of Spanish who started learning English at the same age in school. Students did not have significantly different grades in the first semester of high school. There were no significant differences between groups in SES or years in extracurricular English classes prior to enrollment in the program. There was a statistically significant effect of the intervention (CLIL) on the overall score of the proficiency test and the total number of correct imitations (EI task). Students in the CLIL group outscored their counterparts in the non-CLIL group on the total score of the proficiency test and each of its parts (reading comprehension, Use of English and written production). In addition, whereas students in the CLIL group were able to repeat a similar rate of sentences with preposed clauses (backward anaphora) and sentences with postposed clauses (forward anaphora), students in the non-CLIL group were weaker at sentences with preposed clauses (backward anaphora). Collectively, these results imply that the CLIL participation has a positive effect on the students' proficiency and grammatical development. Finally, a strong positive correlation was found between the scores of both tests for the CLIL group, suggesting that both are measuring the students' knowledge of the English language.

6. Discussion

The goal of the present research was to test the hypothesis that the CLIL approach enhances the students' level of English as an L2. As predicted, we have provided evidence that students enrolled in the CLIL group outscore their counterparts in the non-CLIL group both in a proficiency test which measures performance in reading comprehension, Use of English, and written production, and in an EI task that taps into grammatical knowledge more directly. Simultaneously, we have advanced a methodology to assess the language grammar of students enrolled in a CLIL group by designing an EI task based on the acquisition of adverbial subordinate clauses in English.

This is the first study to successfully use the EI task based on a syntactic developmental pattern as an assessment method. Järvinen (2005) used an EI in which CLIL and non-CLIL students had to repeat a set of complex sentences with relative clauses of varying levels of difficulty. Although CLIL groups produced significantly longer and more complex sentences than control groups, the mean score obtained by the two groups of students in each type of sentence did not cohere with the hypothesized difficulty of the sentences. In the present experiment, the constraints of PBD on the acquisition of the L2 syntax have provided a useful pattern to measure the acquisition of syntax by Spanish students. The students in the CLIL group produced a higher total of correct imitations and showed a more advanced knowledge of the English syntax than their peers in the non-CLIL group. It is possible that an intervention by type interaction was not found due to the sample size. However, a pairwise t-test revealed a significant difference between sentences with preposed clauses (backward anaphora) and sentences with postposed clauses (forward anaphora) only for students in the non-CLIL group.

Proficiency tends to be influenced by extra-linguistic factors, such as pedagogical intervention, and so it may vary across different CLIL programs. In turn, an assessment of the grammar with the EI task produces generalizable results about the context-free effect of the CLIL approach on the L2 acquisition. The low level of students in the non-CLIL group could explain the non-significant correlation between the two tasks. Given that on average non-CLIL students did not pass the proficiency test and only repeated 40% of the sentences in the EI task, these tests are likely too advanced to adequately evaluate the English level of these students. Either complemented with a proficiency test or on its own, the EI task can be used to test language knowledge by consulting developmental patterns uncovered by studies on L2 acquisition. Future research could evaluate different aspects of the students' grammar by developing batteries of sentences that adequately measure L2 acquisition. Differences between the L1, the L2, or even the L3 grammar would need to be taken into account in the creation of the stimuli. In sum, assessing the acquisition of a foreign language should become an interdisciplinary endeavor that integrates L2 teaching, linguistics, and experimental psychology to broaden teaching and learning effectiveness.

Results of this experiment have limitations and raise the necessity for further research. Having conducted the study only in one high school necessarily limits the generalizability of the results. Although there is a possibility that the differences are due to specific characteristics of this particular high school, the fact that the regional government has established the same guidelines and provided the same materials for all schools in Andalusia makes our results relevant for the most populated region in the country. Thus, we have shown the context-specific gains of the approach for English proficiency and competence in Andalusia, but future studies should extend this experiment to other CLIL schools located in different areas of the Spanish territory (other cities and towns) with students of different SES.

Additionally, this study shares the shortcomings of any quasi-experimental design. By assuming the absence of additional confounders, we run the risk of ignoring a common cause that is responsible both for the students' decision to enter a CLIL program and the results obtained. L2 proficiency does not only depend on the students' contact with the language (as we measured in the questionnaire), but on other factors such as motivation (Clément, 1980). If motivation had influenced the students' decision to enroll in the CLIL program and their outcomes in the test, it would become a potential confounder in our experiment. An ideal longitudinal study would have allowed us to obtain pre-intervention scores in English proficiency and motivation for L2 learning.

Notwithstanding these caveats, the results of this study have implications for the educational system in Spain. Given the low position of Spain in the European language rankings (TNS Opinion & Social, 2012), we need to find ways of improving the system of foreign language teaching. The study shows the benefits for the CLIL group, but there is cause for concern about the non-CLIL group. At the time of testing, almost none of the students in the non-CLIL group would have passed the English test that is part of the Spanish University entrance examination. The EI task and the experimental design have shown that the difference between groups may be partially attributed to the students' command of English grammar. This research may serve as a wake-up call for policymakers, teachers and researchers to look into the impact of CLIL not only on the students who enroll in the program, but also on the students who remain in non-CLIL classes.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

About the author

Mar Gutiérrez-Ortiz holds a BA in English (2013) and M. Ed. in Teaching English as a Second Language (2015) from the University of Seville. After doing research as a non-degree Ph.D. student in the English Department at Cornell University (2013-2014), she worked with the Cornell Language Acquisition Lab from 2014 to 2016. She earned a Fulbright-funded M.A. in Developmental Psychology from Cornell University in 2016. She co-authored the translation of Iris Murdoch's novel *Nuns and Soldiers*, which was published by Impedimenta in 2019. She is currently working as a High School teacher and pursuing her Ph.D. degree in the Department of English and American Literature at the University of Seville.

Acknowledgements

Thanks, first of all, to my advisor at Cornell University, Prof. Barbara C. Lust, for giving me a chance to

work on this project under her guidance. I must also thank my committee member Prof. Qi Wang for her insightful comments and suggestions. I gratefully acknowledge the effort of teachers and students at I.E.S. Mariana Pineda for facilitating data collection.

References

Bruton, A. (2011a). 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.

Bruton, A. (2011b). Is CLIL so beneficial, or just selective? Re-evaluating some of the research. *System*, 39(4), 523–532.

Bruton, A. (2013). CLIL: Some of the reasons why … and why not. *System*, 41(3), 587–597.

Bruton, A. (2015). CLIL: Detail matters in the whole picture. More than a reply to J. Hüttner and U. Smit (2014). *System*, 53, 119–128.

Chomsky, N. (1980). Rules and Representations. New York: Columbia University Press.

Dalton-Puffer, C. (2008). Outcomes and Processes in Content and Language Integrated Learning (CLIL): Current Research from Europe. In L. V. W. Delanoy (Ed.), Future Perspectives for English Language Teaching (pp. 139–157). Heidelberg: Carl Winter.

Clément, R., C, R., & Smythe, P. C. (1980). Social and individual factors in second language acquisition. *Canadian Journal of Behavioural Science/Revue Canadienne Des Sciences Du Comportement*, 12(4), 293–302.

Epstein, S. D., Flynn, S., & Martohardjono, G. (1996). Universal Grammar and second language acquisition: The null hypothesis. *Behavioral and Brain Sciences*, 19(4), 746–758.

Flynn, S. (1983). A Study of the Effects of Principal Branching Direction in Second Language Acquisition: the Generalization of a Parameter of Universal Grammar from First to Second Language Acquisition. Cornell University, Ithaca.

Flynn, S. (1986). Production vs. Comprehension: Differences in Underlying Competences. *Studies in Second Language Acquisition*, 8(2), 135–164.

Flynn, S., & Espinal, I. (1985). Head-initial/head-final parameter in adult Chinese L2 acquisition of English. *Second Language Research*, 1(2), 93–117.

Flynn, S., Foley, C., & Vinnitskaya, I. (2004). The Cumulative-Enhancement Model for Language Acquisition: Comparing Adults' and Children's Patterns of Development in First, Second and Third Language Acquisition of Relative Clauses. *International Journal of Multilingualism*, 1(1), 3–16.

Gaillard, S., & Tremblay, A. (2016). Linguistic Proficiency Assessment in Second Language Acquisition Research: The Elicited Imitation Task. Language Learning, 66(2), 419-447.

Ganzeboom, H. B. G., De Graaf, P. M., & Treiman, D. J. (1992). A standard international socio-economic index of occupational status. *Social Science Research*, 21(1), 1–56.

Gené-Gil, M., Juan-Garau, M., & Salazar-Noguera, J. (2015). Writing development under CLIL provision. In M. Juan-Garau & J. Salazar-Noguera (Eds.), Content-based language learning in multilingual educational environments (pp. 139–161). Berlin: Springer.

Housen, A., & Kuiken, F. (2009). Complexity, Accuracy, and Fluency in Second Language Acquisition. *Applied Linguistics*, 30(4), 461–473.

Järvinen, H. M. (2005). Language Learning in Content Based Instruction. In A. Housen & M. Pierrard (Eds.), Investigations in Instructed Second Language Acquisition. Berlin/Boston: De Gruyter Mouton.

Jiménez-Catalán, R. M., & Ruiz de Zarobe, Y. (2009). The receptive vocabulary of EFL learners in two instructional contexts: CLIL vs. Non-CLIL. In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (Eds.), Content and language integrated learning: Evidence from research in Europe. (pp. 81–92). Bristol: Multilingual Matters.

Kim, A., Park, A., & Lust, B. (2018). Simultaneous vs. successive bilingualism among preschool-aged children: a study of four-year-old Korean–English bilinguals in the USA. *International Journal of Bilingual Education and Bilingualism*, 21(2), 164-178.

Lasagabaster, D. (2008). Foreign Language Competence in Content and Language Integrated. *The Open Applied Linguistics Journal*, 1(1), 30–41.

Lázaro Ibarrola, A. (2012). Faster and further morphosyntactic development of CLIL vs. EFL Basque-Spanish bilinguals learning English in high-school. *International Journal of English Studies*, 12(1), 79+.

Lorenzo, F., Casal, S., & Moore, P. (2010). 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.

Lust, B. (1981). Constraints on anaphora in child language: A prediction for a universal. In S. Tavakolian (Ed.), Language acquisition and linguistic theory (pp. 74–96). Cambridge, Mass.: M.I.T. Press.

Lust, B. (2011). Acquisition of Language. In P. Hogan (Ed.), Cambridge Encyclopedia of the Language Sciences. (pp. 56–64). Cambridge, England: Cambridge University Press.

Lust, B. (2012). Tracking universals requires grammatical mapping. In K. K. Grohmann, A. Shelkovaya & D. Zoumpalidis (Eds.), Linguists of Tomorrow: Selected Papers from the 1st Cyprus Postgraduate Student Conference in Theoretical and Applied Linguistics (pp. 105-130). Newcastle upon Tyne: Cambridge Scholars Publishing.

Lust, B., Bhatia, T., Gair, J., Sharma, V., & Khare, J. (1995). Children's acquisition of Hindi anaphora: A parameter-setting paradox. In V. Ghambir (Ed.), Teaching and acquisition of South Asian languages (pp. 172–189). Philadelphia: University of Pennsylvania Press.

Lust, B., Flynn, S., & Foley, C. (1996). What Children Know about What They Say: Elicited Imitation as a Research Method for Assessing Children's Syntax. In D. McDaniel, C. McKee, & H. Smith Cairns (Eds.), Methods for Assessing Children's Syntax (pp. 55–76). Cambridge, Mass.: M.I.T. Press.

Lust, B. and Blume, M. (2016). Research Methods in Language Acquisition: Principles, Procedures, and Practices. Berlin/Boston: De Gruyter, Inc.

Martínez Adrián, M., & Gutiérrez Mangado, J. (2009). The acquisition of English syntax by CLIL learners in the Basque Country. In Y. Ruiz de Zarobe & R. M. Jiménez-Catalán (Eds.), Content and language integrated learning: Evidence from research in Europe (pp. 176–196). Bristol: Multilingual Matters.

Moreno Espinosa, S. (2009). Young learners' L2 word association responses in two different learning contexts. In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (Eds.), Content and language integrated learning: Evidence from research in Europe (pp. 93–111). Bristol/Tonawanda/Ontario: Multilingual Matters.

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, & J. M. Gallardo del Puerto (Eds.), Content and foreign language integrated learning: Contributions to multilingualism in European contexts (pp. 155–186). Bern: Peter Lang.

Navés, T., & Victori, M. (2010). CLIL in Catalonia: an overview of research studies. In D. Lasagabaster & Y. Ruiz de Zarobe (Eds.), CLIL in Spain: Implementation, results and teacher training (pp. 30–54). Newcastle upon Tyne: Cambridge Scholars Publishing.

Pérez Cañado, M. L. (2016). From the CLIL craze to the CLIL conundrum: Addressing the current CLIL controversy. Bellaterra Journal *of Teaching & Learning Language & Literature*, 9(1), 9-31.

Pérez Cañado, M.L. (2018) 'CLIL and Educational Level: A Longitudinal Study on the Impact of CLIL on Language Outcomes. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras*, 29, 51-70.

Pérez-Vidal, C., & Roquet, H. (2015). CLIL in context: Profiling language abilities. In M. Juan-Garau & J. Salazar-Noguera (Eds.), Content-based language learning in multilingual educational environments (pp. 237–255). Berlin: Springer.

Prieto-Arranz, J. I., Rallo Fabra, L., Calafat-Ripoll, C., & Catrain-González, M. (2015). Testing progress on receptive skills in CLIL and non-CLIL contexts. In M. Juan-Garau & J. Salazar-Noguera (Eds.), Content-based language learning in multilingual educational environments (pp. 123–137). Berlin: Springer.

Ruiz de Zarobe, Y. (2008). CLIL and Foreign Language Learning: A Longitudinal Study in the Basque Country. *International CLIL Research Journal*, 1(1), 60–73.

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/Philadelphia: John Benjamins.

Ruiz de Zarobe Y. (2015) The Effects of Implementing CLIL in Education. In: Juan-Garau M., Salazar-Noguera J. (Eds), Content-based Language Learning in Multilingual Educational Environments (pp. 51-68). Berlin: Springer.

Ruiz de Zarobe, Y., & Zenotz, V. (2012). Estrategias de lectura en CLIL. Presented at the International Symposium on Practical Approaches in CLIL, Pamplona, Spain.

Ruiz de Zarobe, Y., & Zenotz, V. (2015). Reading strategies and CLIL: the effect of training in formal instruction. *The Language Learning Journal*, 43(3), 319–333.

TNS Opinion & Social (2012). Europeans and their Languages. Special Eurobarometer 386. European Commission.

Villarreal, I., & García Mayo, M. P. (2009). Tense and agreement morphology in the interlanguage of Basque/Spanish Bilinguals. In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (Eds.), Content and language integrated learning: Evidence from research in Europe (pp. 157–175). Bristol: Multilingual Matters.

APPENDIX I

Elicited Imitation task

Preposed/backward anaphora:

When he entered the office, the professor questioned the man.
(1) When he delivered the message, the man questioned the lawyer.
When he prepared the breakfast, the doctor called the professor.

Preposed/forward anaphora:

When the doctor received the results, he called the gentleman.
(2) When the lawyer delivered the plans, he answered the worker.
When the professor opened the package, he answered the man.

Postposed/forward anaphora:

The man answered the boss when he installed the television.
(3) The mayor questioned the president when he entered the room.
The man introduced the actor when he delivered the plans.