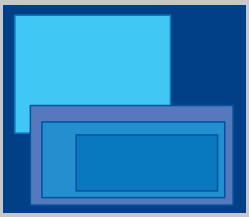


VIAL

VIGO INTERNATIONAL JOURNAL of APPLIED LINGUISTICS

Number 18-2021

Universidade Vigo

Servizo
de Publicacións

VIAL

Vigo International Journal
of Applied Linguistics

Universidade Vigo

VIAL. Vigo International Journal of Applied Linguistics.

VIAL is an international and peer-reviewed journal which is indexed and abstracted in CINDOC (Centro de Información y Documentación Científica), EBSCO, ERIH PLUS, JCR, Web of Knowledge (Arts and Humanities Citation Index, Social Sciences Citation Index), Linguistics Abstracts, Linguistics and Language Behavior Abstracts, Latindex, MLA Directory of Periodicals, PsychoInfo, Scimago Journal Rank Scopus, Ulrich's Database

Webpage: www.webs.uvigo.es/vialjournal/

Printed in Spain - Impreso en España
I.S.S.N.: 1697-0381 - Electronic ISSN: 2660-504X - Depósito Legal: VG-935-2003
Imprime e maqueta: Tórculo Comunicación Gráfica, S.A. Deseño da portada: Elena Gómez.

VIAL

Vigo International Journal of Applied Linguistics

Number 18 - 2021

Contents:

Classroom Enjoyment and Anxiety among Saudi Undergraduate EFL Students: Does Gender Matter?

Elias Bensalem
Department of Languages and Translation
Northern Border University, Saudi Arabia
elias.bensalem@nbu.edu.sa

Abstract

The current study was motivated by recent interest in the effect of positive and negative emotions in the context of foreign language learning resulting from the rise of the positive psychology movement (Dewaele & MacIntyre, 2016; MacIntyre & Mercer, 2014). It examines the construct of foreign language enjoyment (FLE) and its relationship with foreign language classroom anxiety (FLCA) among a group of 487 English as a foreign language (EFL) students (340 females, 147 males) enrolled in public universities in Saudi Arabia. A measure of FLE based on Likert scale ratings of ten items (Dewaele & MacIntyre, 2014), and a measure of FLCA based on eight items extracted from the FLCAS (Horwitz et al., 1986) were used. Male and female students had the same levels of FLE and FLCA. Correlation analysis showed that the relationship between students' FLE and FLCA was significantly negative. Qualitative analysis of the participants' learning experiences revealed the causes of FLCA and FLE among Saudi EFL learners.

Keywords: Classroom language anxiety, EFL learners, gender, foreign language enjoyment, positive psychology.

Resumen

El estudio actual está motivado por el reciente interés sobre el efecto de las emociones positivas y negativas en la adquisición de un idioma extranjero, como resultado del incremento del movimiento psicológico positivo (Dewaele & MacIntyre, 2016; MacIntyre & Mercer, 2014). Examina el constructo del disfrute de la lengua extranjera (FEL) y su relación con la ansiedad en la clase de lengua extranjera (FLCA) entre un grupo de 487 estudiantes de inglés como lengua extranjera (EFL) (340

mujeres, 147 hombres) matriculados en universidades públicas en Arabia Saudí. Se utilizó una medida de FLE, basada en los rangos de la escala Likert de diez ítems (Dewaele & MacIntyre, 2014) y una medida de FLCA basada en 8 ítems extraídos de los FLCA (Horwitz et al., 1986). Los hombres y las mujeres mostraron los mismos niveles de FLE y FLCA. El análisis de correlación mostró que la relación entre la FLE de los estudiantes y la FLCA fue significativamente negativa. El análisis cualitativo de participantes en la experiencia de aprendizaje reveló las causa de FLCA y de FLE entre los aprendices de EFL.

Palabras clave: Ansiedad en aula, aprendices de inglés como lengua extranjera, género, disfrute del idioma extranjero, psicología positiva.

1. Introduction

The role of emotion in the field of second language acquisition (SLA) used to be a neglected area of research. Swain (2013) argues that "emotions are the elephants in the room-poorly studied, poorly understood, seen as inferior to rational thought" (p. 11). Dewaele (2005) claims that only a small number of studies examined the potential impact of emotions on language learning. However, until recently, the literature has focused mostly on negative emotions such as anxiety, while positive emotions have not received enough attention (Bown & White, 2010; Fredrickson, 2001). The bulk of studies on negative emotions focused on the construct of foreign language anxiety (e.g., Dewaele et al., 2019a; Jin & Dewaele, 2018; Horwitz, 2001; MacIntyre, 2017). With the introduction of the concept of Positive Psychology into the field of SLA by MacIntyre and Gregersen (2012), many scholars have started to examine both negative and positive emotions as essential elements in the process of language acquisition (e.g., De Smet et al., 2018; Dewaele & MacIntyre, 2014; Dewaele & MacIntyre, 2016; Dewaele & Dewaele, 2017; Elahi Shirvan & Taherian, 2018). Furthermore, many studies have explored the potential role played by different variables in learners' experience of FLE and FLCA such as age (e.g., Dewaele et al., 2018; Dewaele & MacIntyre, 2014; Dewaele & MacIntyre, 2019), the number of languages studied (e.g., Dewaele et al., 2018; Dewaele & MacIntyre, 2014; Dewaele & MacIntyre, 2019), and teacher friendliness (e.g., Dewaele & MacIntyre, 2019; Dewaele et al., 2019a; Dewaele et al., 2019b). The role of gender was also another variable explored by a few studies (Dewaele & MacIntyre, 2014; Dewaele et al., 2019a; Jiang & Dewaele, 2019) primarily in an international context, in China and the United Kingdom. Gender differences merit exhaustive examination (Jiang & Dewaele, 2019).

Based on the above, it would be useful to examine whether female learners who study EFL in a different educational and linguistic setting experience more enjoyment

than their male counterparts. The understanding is that sociocultural factors, especially the status of women in society, determine the role of gender in foreign language (FL) learners' experience of FLCA (Bensalem, 2018; Park & French, 2013; Song, 2018). Denies and Janssen (2016) argued that gender differences may also vary from one country to another.

The present study seeks to explore this line of research through learners' FLE and FLCA in the Saudi EFL tertiary level context. More specifically, this study sets out to explore whether gender affects FLE and FLCA of EFL learners. In addition to investigating the impact of the gender variable on anxiety and enjoyment, the study seeks to identify and detect the sources of FLE and FLCA among male and female students in the underexplored Saudi educational EFL setting.

2. Literature review

The inclusion of positive emotions in SLA is based on the premise that positive emotions may help facilitate the language learning process (MacIntyre & Gregersen, 2012). It is not sufficient to stimulate students by alleviating negative emotions (Dewaele & Alfawzan, 2018). There is a need to recognize enjoyment as a powerful motivator in SLA (Pavelescu & Petri, 2018; Piniel & Albert, 2018). For Boudreau et al. (2018), foreign language enjoyment is a "complex and stable emotion" that is entirely separate from the "more superficial experience of pleasure" (p.153).

2.1. Studies on foreign language enjoyment

A large-scale study on FLE was conducted by Dewaele & MacIntyre (2014) involving 1.740 FL learners from childhood to adulthood and from all corners of the globe. The authors developed a 21-item FLE scale to measure positive emotions towards the learning experience, peers, and teacher. Eight additional items extracted from Foreign Language Classroom Anxiety Scale (FLACS; Horwitz et al., 1986) were added to measure participants' FLCA. Statistical analysis showed that there was a negative correlation between FLE and FLCA. The authors argued that even though the two constructs are interrelated, they constitute, nevertheless, separate dimensions. Another interesting finding is that students who exhibited significantly higher levels of FLE and lower levels of FLCA had a higher level of multilingualism and a more advanced level of proficiency; they felt they outperformed their peers in the language class at the college level and beyond.

The same dataset was used by Dewaele & MacIntyre (2016) to examine the underlying dimensions of FLE. Results of factor analysis generated three dimensions:

FLCA, social FLE, and private FLE. Together, these accounted for 45% of the variance. The authors found that the social FLE dimension was independent of the private FLE dimension.

In a pseudo-longitudinal study, Dewaele & Dewaele (2017) examined the development of FLE and FLCA among 189 foreign language students from two top-performing schools in the UK. Participants were divided into three age groups: 12-13 year olds (age group 1), 14-15 year olds (age group 2), and 16-18 year olds (age group 3). A Pearson correlation analysis revealed a significant negative correlation between FLE and FLCA only in age group 2; participants who had higher levels of FLE experienced lower levels of FLCA. Results showed slight variation in FLCA and a small increase in FLE. A repeated measures analysis of variance revealed little variation in FLCA and a slight increase in FLE among the participants. Furthermore, the data showed that the causes of positive and negative emotions are not static and tend to change over time.

Using the same data set of the previous study, Dewaele et al. (2018) investigated the effect of learner-internal and learner-external variables on levels of FLCA and FLE among 189 secondary school students who were mostly taking French, German, or Spanish courses. Results showed that participants had significantly higher levels of FLE than FLCA. Furthermore, a weak negative relationship between FLE and FLCA was reported, which corroborates the results reported in Dewaele and MacIntyre (2014). Finally, gender played a role in participants' experience of FLE and FLCA since female students had higher levels of FLE and FLCA than male students.

In the Canadian context, Boudreau et al. (2018) examined the dynamic relationship between FLE and FLCA in a second-by-second analysis of a group of ten college-level English-speaking students learning French as an L2. Participants had to complete oral tasks while being video recorded. First, they were instructed to describe a photograph of something they thought was enjoyable for a few minutes. Then, they were given five oral interview style questions. Idiodynamic software was used to rate students' anxiety and enjoyment while they watched their recorded tasks. The computer software generated a graph of the participant's ratings. The researcher and the participant examined the graph and discussed spikes and dips in ratings. The researchers examined the correlation between FLE and FLCA for each participant; similarly, correlation was examined for the photo description and the oral interview. This approach was used to measure the fluctuating relationships between FLE and FLCA. Data analysis showed a complex correlation between FLE and FLAC. In certain situations, FLE and FLCA moved closer to each other, while in others, they were further apart. In some situations, there was no relationship between the two, and both acted in separate ways.

The role of the target language in influencing learners' FLE and FLCA has also been a topic of interest. De Smet et al. (2018) undertook a study that compared anxiety and enjoyment among students learning two different target languages (English and Dutch). A group of 896 pupils from elementary and secondary education schools in French-speaking Belgium, located in two educational contexts, namely content and language integrated learning (CLIL) and non-language integrated learning (non-CLIL), participated in the study. Results of data analysis revealed that CLIL pupils exhibited lower levels of anxiety than non-CLIL pupils did. Furthermore, Dutch learners experienced more anxiety and less enjoyment than English learners. These findings suggest that emotional engagement may be a function of the target language.

Attempting to examine potential difference between FLE and FLCA among learners from different linguistic settings, Jiang & Dewaele (2019) compared FLE and FLCA of English as a foreign language (EFL) students in China to learners in other countries. The authors found that participants reported significantly higher levels of enjoyment than anxiety in their English classes. This finding is consistent with previous research (Dewaele et al., 2017; Dewaele & Dewaele, 2017; Dewaele & MacIntyre, 2014; Khajavy et al., 2018). The mean of FLE was higher than the mean in the Dewaele & MacIntyre's (2014) sample, which included learners from all over the world. The mean of FLCA, however, was much higher than the mean of the international sample. The researchers attributed Chinese EFL learners' tendency to experience higher levels of anxiety than their overseas peers to the educational system in China, which does not provide ample opportunities for EFL learners to practice English (see Shi, 2008).

Finally, studying FLE and FLCA in Saudi Arabia has barely started to draw the interest of researchers, and so far, only one study has been carried out by Dewaele & Alfawzan (2018). The authors examined the effect of FLE and FLCA on English language performance among a group of 152 Saudi learners and users of English. The majority of participants had completed their undergraduate education, while the remaining participants were undergraduate English students. The authors reported that participants' higher levels of FLE were correlated with significantly higher English proficiency scores. Conversely, participants' higher levels of FLCA were linked to lower English proficiency scores. The qualitative data suggested interactions between participants' experience of FLE and FLCA in their English classes. The participants' experiences of FLE and FLCA were shaped by their perception of the teacher and teachers' pedagogical practices in the classroom. These findings are consistent with the study outcomes reported by Mierzwa (2019a). Participants in this study cited teacher-related variables such as teacher attitude, level of support provided to students, teaching strategies, and classroom activities as the primary sources of FLE.

2.2. Self-perceived language proficiency and foreign language enjoyment and foreign language classroom anxiety

Very few studies explored the relationships between FLE, FLCA, and learners' self-perceived language proficiency. Dewaele & MacIntyre reported self-perceived foreign proficiency as one of the significant predictors of both FLE and FLCA. Similarly, Li et al. (2018) found that for EFL Chinese students at different levels, both FLCA and FLE are significant predictors of their self-perceived proficiency in English.

2.3. Gender differences in foreign language enjoyment and foreign language classroom anxiety

Studies examining the role of gender in FL learners' experience of FLE yielded inconsistent results. Dewaele et al. (2016) conducted a study on the same dataset used by Dewaele & MacIntyre (2014) to investigate potential gender differences at item level. The data revealed that female students (n = 1,287) experienced more enjoyment in foreign language lessons, actively agreed they had acquired interesting knowledge, and were more appreciative of their foreign language prowess than their male counterparts (n = 449). The female students also reported more excitement in the foreign language class than their male peers. They enjoyed being able to express themselves and learning something new. Unlike their male counterparts, they felt anxious about making mistakes and were unsure of themselves at times. In a more recent study involving 750 FL learners from around the world, Dewaele & MacIntyre (2019) confirmed females' tendency to experience higher levels of FLCA but did not find any evidence supporting the view that female learners report more enjoyment than male learners. A different role for gender was reported by Dewaele et al. (2019b). In their study which involved 592 learners of Turkish as a foreign language in Kazakhstan they found that no gender differences existed for FLE, while male participants experienced higher levels of FLCA than female participants (Dewaele et al., 2019b). The absence of the role of gender for FLE was also reported by Mierzwa (2019b). In her study, students enrolled in English Philology classes exhibited the same levels of FLE regardless of their gender or proficiency level.

In their seminal article, De Smet et al. (2018) found that different populations of learners studying two different foreign languages, namely English and Dutch, experienced varying levels of FLCA and FLE. They attributed this difference to how each foreign language is perceived, to the level at which it is studied, and the manner in which instruction is delivered: content and language integrated learning (CLIL) on the one hand, and non-CLIL on the other.

3. The present study

So far, only a limited number of studies have examined the role of gender in FL learners' experience of FLE. No single study has been conducted yet in male-dominated societies, such as the ones located in the Arabian Gulf region. Previous research has shown that learners' experience of FLCA is determined by sociocultural factors. What remains to be further explored is whether these same factors would affect learners' FLE. So far, only one study on FLE was conducted with Arab EFL students. This study involved participants who had already completed their undergraduate education. However, it did not control for the fact that over half of the participants had travelled outside Saudi Arabia, where they had studied or used English. Therefore, its outcomes cannot be generalized. Thus, the puzzle of FLE and its relationship with FLCA among Saudi students remains incomplete.

There are three motivations for the present research. The first to enrich the emerging literature on the newfound interest in the effect of positive and negative emotions on foreign language learning in terms what De Smet et al. (2018) call emotional engagement, FLCA, and FLE. The second motivation is to investigate the effect of context on how learners perceive and experience these feelings. The context under scrutiny here is Saudi Arabia, where an empirical investigation is called for to unpack the complex relationship between FLE, FLCA and self-perceived FL proficiency in English. All the participants were Arabic first language users studying EFL and with no experience studying or staying abroad. In order to fine-tune the outcomes of this study, the gender variable is added in hopes of finding out whether Saudi female learners who study EFL experience more enjoyment than their male counterparts. This third motivation is driven by the prevailing understanding that gender as a sociocultural construct impacts how FL learners in different countries experience FLE and FLCA and may even attain different FL proficiency levels. The following research questions will guide the present study:

1. What are the levels of FLE and FLCA of Saudi EFL learners?

2. Is there a relationship between Saudi EFL learners' levels of FLE and FLCA?

3. Are there any differences between male and female Saudi learners' levels of FLE and FLCA?

4. What is the effect of self-perceived proficiency in English and year of study on FLE and FLCA of Saudi EFL learners?

5. What are Saudi learners' views on enjoyable and anxious experiences in the EFL class?

4. Method

4.1. Participants

A total of 487 EFL Saudi students (340 females, 147 males) participated in the study. They are all native speakers of Arabic. Their ages ranged from 18 to 33 (*M* = 21, *SD* = 2.45). About 33.7% are freshman, 24.2% are sophomore, 18.1 % are junior, and 24% are senior (see Table 1). All students were required to enroll in a preparatory English program prior to enrollment at the university. It is an intensive five-course (half-semester) program which aims at developing the English proficiency required in a university. Students who finish the program should have at least an intermediate level of English. None of the participants had study abroad experience. Since permission to measure students' levels of proficiency through a proficiency test was not granted, the researcher decided to use a self-rated proficiency measure as an alternative method. Participants were instructed to rate their proficiency in English on a scale from 1 to 10 for listening, speaking, reading and writing are reported. The same scale was used in previous studies (e.g., Santos et al., 2017; Thompson & Lee, 2013). Participants' English self-perceived proficiency levels are reported in Table 2.

Table 1: Participants' background information

Variable	Category	Frequency	Percentage
Gender	Female	340	69.8%
	Male	147	30.2%
Year of study	Freshman	164	33.7%
	Sophomore	118	24.2%
	Junior	88	18.1%
	Senior	117	24%

Table 2: English self-perceived proficiency levels

	M	*SD*
Speaking (max = 10)	6.41	1.96
Listening (max = 10)	7.29	2.43
Reading (max = 10)	7.28	2.11
Writing (max = 10)	6.43	2.28

4.2. Instruments

Data were collected through an Arabic version of Dewaele & MacIntyre's (2014) self-report questionnaire measuring students' anxiety and enjoyment in the classroom, along with background information. The questionnaire was used in recent studies (Jiang & Dewaele, 2019; Li et al., 2018). The first section of the survey included questions that elicited participants' sociobiographical information related to their age, gender, and year of study. The second section included 10 items from Dewaele & Dewaele (2017) that were extracted from the original 21-item FLE questionnaire (Dewaele & MacIntyre, 2014). The items, which were all positively phrased, were based on five-point Likert scales (1= strongly disagree, 2 = disagree, 3 = neutral, 4 = agree, 5 = strongly agree). In terms of internal consistency, the Cronbach's alpha for the FLE scale was. 703, suggesting acceptable internal consistency. The third section of the survey consisted of eight items extracted from the FLCAS (Horwitz et al., 1986) and used in Dewaele & MacIntyre (2014). They were used to determine participants' anxiety levels and were based on five-point Likert scales (1= strongly agree, 2 = agree, 3 = neutral, 4 = disagree, 5 = strongly disagree). Six items were positively phrased. The two remaining items were reverse coded since they were negatively phrased. High scores indicate high levels of anxiety. The scale had a good internal reliability (Cronbach alpha =. 781). The questionnaire ended with two open questions. Participants were asked to describe one memorable enjoyable learning experience of enjoyment and one memorable anxious learning experience in the EFL class.

4.3. Procedure

After obtaining the Human Research Ethics Committee's approval, a number of teaching staff from target institutions helped recruit participants for the study. The participants were informed of the purpose of the research and were assured that the collected data would be kept confidential. All survey items were formulated in Arabic, the native language of students in order to ensure that participants fully understood the survey items. An Arabic version would increase the participation rate as students were more comfortable with Arabic. Participants were provided with a link to a Google form containing the survey, which was completed anonymously. The survey took about 15 minutes to complete. Data collected from Google form were coded and analyzed using the Statistical Package for Social Sciences (SPSS).

4.4. Data analysis

A data-validation variant of a convergent parallel mixed-methods design (Creswell & Plano Clark, 2011) was adopted. The quantitative data (closed questions) aims

at measuring participants' levels of FLE and FLCA among males and females and identifying the link between FLE and FLCA. The qualitative data (open questions) is used to analyze participants' enjoyable and anxious experiences. Creswell & Plano Clark (2011) argue that mixed method research design enables a greater degree of understanding than if a single approach was adopted. They argue that qualitative items "provide the researcher with emergent themes and interesting quotes that can be used to validate and embellish the quantitative survey findings" (p.81).

The data analysis was performed in two stages. In the first stage, descriptive statistics (i.e., means and standard deviations) were used to summarize participants' responses. As Kolmogorov-Smirnov test revealed that the distribution was not normal, $D(487) = 0.084$, $p < .001$, a Mann-Whitney U test was carried out to examine whether there were differences in FLE and FLCA between male and female participants. Furthermore, Spearman Rho correlation analysis was performed to examine any statistically significant correlation between FLE and FLCA. Furthermore, Linear regression analysis was used to assess the effect of the learners' self-perceived proficiency and year of study and determine their relative contribution to the prediction of FLE and FLCA.

In the second stage, content analysis was used to analyze participants' enjoyable and anxious experiences in the EFL class following Dewaele & MacIntyre (2014). The data was coded into main themes. Many respondents chose to respond to the open-ended questions in Arabic or sometimes in a mix of Arabic and English, which made it difficult to use Nvivo software. Therefore, the data was coded manually. Data extracts were selected from the open questions guided by the principle that they are representative of a particular topic, concise, and interesting.

5. Results

The findings of the study will be presented in two sections. The first section will present quantitative findings related to levels of FLE and FLCA among male and female participants, and the relationship between FLE and FLCA. Furthermore, statistical results regarding the effect of self-perceived proficiency in English and year of study on the participants' FLE and FLCA will be presented. The second section will present the quantitative findings. It reports on participants' views on enjoyable and anxious experiences in the EFL class.

5.1. Quantitative findings

5.1.1. Levels of foreign language enjoyment and foreign language classroom anxiety and gender differences

Average scores on the 5-point scale were calculated for FLE (M = 3.61, *SD* = 0.59) and for FLCA (M = 3.70, *SD* = 0.70). Whilst the levels of FLE are comparable to the values reported in previous studies ranging from 3.4 (Dewaele & Alfawazan, 2018) to 3.9 (Jiang & Dewaele, 2019), the levels of FLCA are reported higher by recent studies with a range between 2.4 (Dewaele et al., 2017) and 3.1 (Jiang & Dewaele, 2019). A Mann-Whitney *U* test was conducted to determine whether there was a difference in learners' levels of FLE and FLCA. The results indicated no significant difference between levels of FLE among male and female learners, $z = -1.87$, $p > .05$. Female learners had an average rank of 251.83, while male learners had an average rank of 225.90. Similarly, there was no significant difference between levels of FLCA amongst male and female learners, $z = -1.87$, $p > .05$. Female learners had an average rank of 236.17, while male learners had an average rank of 262.10. In terms of FLE individual items, compared to their male peers, the female participants felt significantly more that they were worthy members of the English language class. Female learners had an average rank of 261.62, while male learners had an average rank of 203.24, ($U = 18998.5$, $p < .001$). Similarly, female students agreed more strongly that it is "cool" to know English ($U = 22125.50$, $p = .037$). Regarding individual FLCA items, the female participants felt significantly more anxious in the English class even if they felt they were well prepared ($U = 21718.50$, $p = .018$). Furthermore, the female participants reported that they started to panic when they had to speak without preparation in the English class ($U = 21085.50$, $p = .005$).

5.1.2. Relationship between foreign language enjoyment and foreign language classroom anxiety

Spearman Rho correlation analysis revealed a perfect negative significant correlation between FLE and FLCA ($Rho = .19$, $p < 0.001$, $r2 = .036$). In other words, participants who had higher levels of FLE tended to have lower levels of FLCA. The effect size for FLE (0.036) can be described as "small" (Cohen, 1992), which means that the FLE and FLCA are not strongly correlated.

5.1.3. Effect of self-perceived English proficiency and year of study on the participants' foreign language enjoyment and foreign language classroom anxiety

In order to examine the simultaneous effect of self-perceived proficiency in English and year of study (freshmen, sophomore, junior, and senior) on the participants' FLE and FLCA, multiple regression analysis was conducted. Results show that the presented model is significant as there was a significant relationship between self-perceived proficiency in English and FLE as well as FLCA (p = .03; $r2$ = 0.084; r = 0.290). The R value for the regression model indicates a small linear relationship between English self-perceived proficiency and FLE. The same linear relationship was found between self-perceived proficiency and FLCA. The $r2$ was 0.28, which English self-perceived proficiency explains 8% of the variance in participants' FLE and FLCA. According to Cohen's (1988) criteria for assessing the predictive power of independent variables, this indicates a small effect size. Year of study was not a predictor of FLE and FLCA.

5.2. Qualitative findings

This section reports on participants' views on enjoyable and anxious experiences in the EFL class. A close reading of the participants' responses to the open questions generated the following themes: classroom activities, instructor skills, personal success/failure in the classroom, assessment, and other (see Table 3). The revealed sources of anxiety and enjoyment are applicable for both males and females.

Table 3: Main themes in the feedback of participants' enjoyable and anxious experiences in the EFL class

Theme	FLE		FLCA	
	Frequency	%	Frequency	%
Classroom activities	126	25.8	87	17.9
Instructor skills	143	29.5	119	24.4
Personal success/failure	73	14.9	81	16.7
Assessment	57	11.8	94	19.2
Other	88	18	106	21.8

5.2.1. Classroom activities

Some classroom activities are enjoyable, while others can cause anxiety among students. Many participants mentioned group activities as a source of joy for them. Participant 13 explained how she enjoyed task based activities:

> *During one of my English classes the teacher set up a group activity which involved everyone in the classroom. Each group member was handed a piece of paper with a specific task to complete. Students' tasks were rated by their fellow group members. Ratings were in the form of the number of stars. By the end of the task, each group would share the number of stars generated by group members. I really enjoyed this activity (P13, female).*

Male participants share their appreciation of group activities. Participant 8 reported the joy of working with his classmates:

> *I enjoy participating in group activities which involve discussions among group members with the instructor interacting with us (P8, Male).*

Some classroom activities were also sources of anxiety for many participants. Speaking activities involving students standing and addressing their classmates were cited as the most anxiety triggering events in the classrooms as Participant 25 remembers:

> *One day the instructor asked me to stand up and share my answer with the rest of classmates. I was not prepared so I became nervous and struggled to deliver. All students laughed and some even made fun of me. It was very embarrassing (P25, female).*

Male students experienced similar anxious situations related to speaking. Participant 53 recalls:

> *The English teacher brought me to his desk and asked me to address my classmates. I was forced to read a passage aloud. This scenario made me nervous (P53, Male).*

Learners appear to enjoy group activities in the classroom over being made to stand in front of the class alone.

5.2.2. Instructor skills

Instructor skills, which included pedagogical practices, were another source of both FLE and FLCA. Some participants expressed their satisfaction of the instructor's ability to create a pleasant learning atmosphere, which resulted in experiencing many enjoyable scenarios in the classroom. For example, Participant 7 enjoyed the instructor's work to engage students in the learning process:

I enjoyed my teacher's way of teaching. She designs fun activities and uses music as a tool to energize students. I love the way she interacts with us (P7, female).

Students also seemed to experience enjoyable episodes when they received support and encouragement from their teacher:

It felt good when my teacher acknowledged my hard work. He thought I was one of his best students. (P4, male).

Other students mentioned particular teachers who bring joy to the classroom because of their willingness to assist students:

The presence of wonderful, caring professors such as MS Sonia makes us enjoy our English classes. I appreciate all the help and dedication she shows to students. She is very approachable. (P1, female).

The instructor attitude was a source of anxiety for some participants. Participant 60 remembers a painful episode that caused him frustration:

Each time I make a mistake the instructor gets upset. Students are supposed to make mistakes. This instructor was not supportive (P60, male).

Teachers who resort to intimidation techniques trigger anxiety among students as mentioned by Participant 3:

One of my professors started warning us that very few students would be able to pass the course and that no student would get an A. I understand that the professor wanted to push us to work harder but it created a very uncomfortable learning atmosphere (P3, female).

It is not surprising that teachers who build rapport with students and who show care and support manage to create a positive learning environment where learners enjoy their language classes. Conversely, instructors who use intimidation tactics as if they are in the military will most likely see their strategies backfire. This will result in students experiencing more negative feelings, including anxiety and loss of interest, rather than experiencing the joy of learning.

5.2.3. Personal success/failure

The extent to which students enjoy their course can be influenced by their perceptions of success and failure. Participant 55, for example, remembers feeling very happy when he managed to communicate in English:

Last semester I was full of joy when I was able to communicate perfectly with my instructor without a single problem. Reaching a certain level of proficiency is very rewarding (P55, male).

However, Participant 71 experienced a sense of failure when struggling to keep up with the teacher along with the following emotion:

> *It was very frustrating when I fail to understand certain concepts. Sometimes I can't follow the instructor because of my lack of understanding (P71, male).*

Students that perceive success are more likely to enjoy their course, whereas the enjoyment of those that experience failure is inhibited.

5.2.4. Evaluation

Many participants especially males, like Participant 24 cited teacher assessment as a source of enjoyment or anxiety:

> *When I was freshmen, my anxiety levels went up because of my low grades. It was struggle as I started to question my ability to pass my courses. I even considered changing my major. However, when my grades improved during the subsequent semesters I felt good. It was exciting to get high grades in translation courses, for example (P11, male).*

Another participant had a similar experience:

> *The first year at university was quite tough. Fear of negative evaluation crippled my ability to do well in exams. Later, I as I worked harder. I was able to overcome my anxiety, and my grades improved (F15, female).*

Grades seem to have an impact on academic motivation. Good grades enhanced enjoyment and self-confidence in one's ability to succeed among participants. In contrast, bad grades seem to have the potential to thwart basic psychological needs and academic motivation, and trigger anxiety.

5.2.5. Other

This category covers participants' statements that did not point out to a specific source of anxiety or enjoyment. For example, one participant stated that he had a good learning experience:

> *I enjoyed everything about my English classes" (P258, male). Another participant expressed her dislike of English classes: "I don't like English and I find classes boring" (P344, female).*

There is a tendency among groups of participants when summarizing their experiences about whether they like their English classes or not: their experience of enjoyment or anxiety depends on how they feel overall about the course.

6. Discussion

The first research question investigated levels of FLE and FLCA of Saudi EFL learners. The mean for FLE in the present study (3.61) is lower than the mean reported by the international sample of Dewaele & MacIntyre (2014), which was 3.82 and the mean reported by Jiang & Dewaele (2019), which was 3.94. The mean for FLCA (3.70) is much higher than the mean reported by Jiang & Dewaele (2019) which was 3.14 and the mean reported in Dewaele & MacIntyre's (2014) study, which was 2.75. This outcome corroborates the findings reported in previous studies, which state that Saudi EFL learners tend to experience higher levels of anxiety compared to students in other countries (e.g., Alrabai, 2015; Hamouda, 2012). Students' experience of high anxiety levels could be ascribed to the educational context in Saudi where students are not exposed enough to English, since instructors tend to use Arabic extensively in the classroom (Alrabai, 2016). EFL teachers' lack of qualifications (e.g., Al-Hazmi, 2003) could have contributed to the creation of a classroom environment that triggered anxiety among learners.

Our second research question addressed the relationship between FLE and FLCA. A negative correlation was found between students' FLE and FLCA, which means that participants who experience higher levels of FLE tend to have lower levels of FLCA. The results are in line with Dewaele & MacIntyre's (2014) finding which reported a small effect size (Cohen, 1992), since FLE and FLCA shared 12.9% of their variance. In the current study, FLE and FLCA shared only 3.6% of their variance. It can be argued that enjoyment and anxiety are opposite ends of the spectrum but as MacIntyre & Legatto (2011) observed, it may be possible to witness learners who are enjoying their English classes, but at the same time, are experiencing episodes of anxiety." (Dewaele & MacIntyre, 2014).

The potential impact of gender on FLE and FLCA was addressed in our third research question. The results indicated no significant difference between levels of FLE among male (*Mdn* = 35) and female learners (*Mdn* = 36), *U* = 22329, p .05). Similarly, there was no significant difference between levels of FLCA amongst male (*Mdn* = 30) and female learners (*Mdn* = 29), *U* = 22329, *p* .05). The difference between female and male respondents was only significant for two items of FLE, with females feeling like they were worthy members of their English classes and that it is "cool" to know English. Similarly, the difference between male and female participants was only significant for two items of FLCA, with females showing more anxiety despite feeling well prepared for their English classes, and experiencing panic when they had to speak without preparation in the English class. The insignificant role of gender reported in the current study corroborates the results found by Jiang & Dewaele (2019), but contradicts previous studies where female students reported both more FLE and

FLCA than male students (Dewaele & MacIntyre 2014; Dewaele et al. 2016). The lack of gender effect on FLCA revealed by the current study is an unexpected finding since previous research has shown that female Saudi students tend to experience higher levels of FLCA than male students (Bensalem, 2018, Bensalem, 2019). Some argue that this is because in Saudi Arabia cultural norms place expectations on women to be reserved and even may discourage women from engagement in academic activities (Song, 2018). As Dewaele (2018) argued, one possible explanation of the lack of gender effect on enjoyment and anxiety is that students who choose to participate in this survey are typically good learners who are pleased with their English classes, rather than weak students. The sample may have included students who rejected the crippling effects of social norms. Therefore, the effect of gender may have been neutralized because of the profile of the participants. In other words, if the participants had been mostly struggling students the outcomes may have been different.

The fourth research question focused on the effect of self-perceived proficiency in English and year of study on the participants' FLE and FLCA. Results showed that year of study was not a predictor of FLE and FLCA, which suggests the fact that participants experienced the same levels of enjoyment and anxiety regardless of their year of study. Conversely, self-perceived proficiency in English predicted both FLCA and FLE. This outcome echoes the findings in Dewaele & MacIntyre (2014) and Li et al. (2018). This suggests that learners who experience a high level of FLE and a low level of FLCA tend to have more confidence and more optimistic self-assessment of foreign language proficiency. This result is aligned with Horwitz et al.'s (1986) claim that there is a relationship between FLCA and self-perceptions in foreign language.

Finally, participants were asked to share their enjoyable and anxious episodes in the EFL class. Content analysis of the participants' responses generated four major themes that shed light on sources of enjoyment and anxiety: classroom activities, instructor skills, personal success/failure in the classroom, and assessment. Three of these themes were similar to those related to FLE reported by Dewaele & MacIntyre (2014), namely classroom activities, instructor skills, and personal success/failure in the classroom. However, only one theme was similar to the ones found by Dewaele & Alfawzan (2018), whose study was conducted in the Saudi context. One possible explanation is the profile of participants in Dewaele & Alfawzan's (2018) study was different. They were mostly users of English who had already graduated. Students who are still involved in the learning process have different experiences and perceptions than those who have successfully completed their studies. One of the unique themes revealed by the current study is related to assessment. Fear of negative evaluation was widely discussed as a main source of anxiety for FL learners (e.g., Aida 1994; Horwitz & Young, 1991; Sellers 2008; Zhang & Zhong, 2012), but assessment as a source of enjoyment had not been discussed in previous research on FLE. It was interesting

to read how the majority of participants cited fear of negative evaluation when they described anxious classroom experiences, while receiving positive evaluation as the most memorable, enjoyable experiences in EFL classes. This could be explained by cultural norms in Saudi Arabia. Getting bad grades can cause embarrassment and "loss of face". According to Saudi culture, "loss of face" may result in loss of honor and respect of others (Al-Saraj, 2014). Conversely, getting good grades will bring pride and respect, even to parents and the whole family. Furthermore, the opportunities that come with good grades include full scholarships and good jobs since some employers check transcripts before hiring potential candidates in Saudi Arabia.

7. Conclusions

The current study examined the construct of FLE and its relationship with FLCA among 487 EFL college level Saudi students. Participants reported similar FLE levels to previous studies, but FLCA levels were higher than those reported in recent research. Male and female students had the same levels of FLE and FLCA. Correlation analysis revealed a negative significant relationship between FLE and FLCA. However, in terms of individual items, female participants felt significantly stronger than their male peers that they were more worthy members of the English language class and agreed more strongly that it is "cool" to know English. Similarly, female students felt more anxious in the English class and experienced panic situations when they had to speak without preparation in the English class. The study confirmed the assumption that self-perceived proficiency in foreign language could be a predictor for FLE and FLCA, while of study had no effect on participants' FLE and FLCA. Qualitative material collected from participants revealed that classroom activities, instructor skills, personal success/failure in the classroom, and assessment were reported to be the main causes of anxiety and enjoyment for male and female participants. Further research could focus on the role of other variables that my combine with gender to influence FLCA and FLE (Dewaele et al., 2016), such as socio-economic status, proficiency in additional languages, and experience abroad. It would be interesting to examine whether the gender patterns in FLE and FLCA are similar for Saudi students enrolled in international schools to those enrolled in public schools.

A number of pedagogical implications can be drawn from this study. Participants cited teacher classroom practices as one of several sources of anxiety and enjoyment. Previous research has documented that learner attributes such as personality and personal experience could be sources of FLE and FLCA not related to the classroom. However, teachers could still play a major role in alleviating anxiety and boosting enjoyment among students. Teachers can turn the classroom into a safe learning environment by being friendly, supportive, humorous, and considerate. Consequently,

learners will be able to experience higher levels of enjoyment (Dewaele & MacIntyre, 2014; Oxford, 2017). Dewaele et al., (2019a) pointed out to the crucial role by FLE in achieving a successful L2 classroom learning experience. Students with a positive attitude tend to be more involved in the learning process and make the best out of learning opportunities. Furthermore, teachers should carefully design activities that are enjoyable and involve all students regardless of their proficiency levels and interests. Such activities can certainly boost students' positive emotions and help them achieve better performance in the FL class (Dewaele & Alfawzan, 2018).

Participants of the current study report fear of failure as a source of anxiety. In this regard, Oxford (2017) suggests teachers should take initiative to strengthen an anxious learner's ability by helping them focus and visualize a positive or interesting fact of the language activity or text, and removing any negative thoughts of failing or difficulty. Teachers should also aid them in releasing any emotional icebergs and grudges they may still hold.

The current study has limitations, which make the findings hard to generalize even though the sample size is relatively large. First, the participants were self-selected. As Dewaele (2018) argued, learners who choose to take the time to fill out a questionnaire are most likely those who are having a good learning experience, rather than learners who are not satisfied with their language performance. Another limitation is the data itself were collected mainly from two universities out of over 30 universities located in Saudi Arabia. Therefore, students from other universities with different academic setting, including teaching staff, may have different perceptions even though the cultural norms are applicable to all students. Thirdly, several learner variables were not accounted for, such as language proficiency, study abroad experience, or knowledge of a third knowledge. These variables, along with motivation, could have influenced the outcomes of the study.

Acknowledgement

The author wishes to acknowledge the approval and the support of this research study by the grant no. 7718-EAR-2018-3-9-F from the Deanship of Scientific Research at Northern Border University, Arar, K.S.A

8. References

Aida, Y. (1994). Examination of Horwitz, Horwitz, and Cope's construct of foreign language anxiety: The case of students of Japanese. *The Modern Language Journal, 78*(2), 155-168.

Al-Hazmi, S. H. (2003). EFL teacher preparation programs in Saudi Arabia: Trends and challenges. *TESOL Quarterly, 37*(2), 341-344

Alrabai, F. (2015). The influence of teachers' anxiety-reducing strategies on learners' foreign language anxiety. *Innovation in Language Learning and Teaching, 9*(2), 163-190.

Alrabai, F. (2016). Factors underlying low achievement of Saudi EFL learners. *International Journal of English Linguistics, 6*(3), 21-37

Al-Saraj, T. M. (2014). Revisiting the Foreign Language Classroom Anxiety Scale (FLCAS): The Anxiety of Female English Language Learners in Saudi Arabia. *L2 Journal, 6*(1), 50-76.

Bensalem, E. (2018). Foreign Language Anxiety of EFL Students: Examining the effect of self-efficacy, self-perceived proficiency and sociobiographical variables. *Arab World English Journal, 9*(2), 38-55Bensalem, E. (2019). Multilingualism and Foreign Language Anxiety: the case of Saudi EFL Learners. *Learning and Teaching in Higher Education: Gulf Perspectives*, *15*(2), 1-14.

Boudreau, C., MacIntyre, P. D., & Dewaele, J.-M. (2018). Enjoyment and anxiety in second language communication: an idiodynamic approach. *Studies in Second Language Learning and Teaching, 8,* 149-170

Bown, J., & White, C. J. (2010). Affect in a self-regulatory framework for language learning. *System, 38*(3), 432-443. Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science, 1*(3), 98-101. Oxford: Blackwell Publishing.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York: Routledge Academic

Creswell, J. W., & Plano Clark, V. L. (2011). *Designing and Conducting Mixed Methods Research* (2nd ed.). Los Angeles: Sage Publications Ltd.

Denies, K, Janssen, R (2016) Country and gender differences in the functioning of CEFR-based can-do statements as a tool for self-assessing English proficiency. *Language Assessment Quarterly 13*(3), 251-76.

De Smet, A., Mettewie, L., Galand, B., Hiligsmann, Ph., & Van Mensel, L. (2018). Classroom anxiety and enjoyment in CLIL and non-CLIL: Does the target language matter? *Second Language Learning and Teaching, 8*(1), 47-72

Dewaele, J.-M. (2005). Investigating the psychological and the emotional dimensions in instructed language learning: Obstacles and possibilities. *The Modern Language Journal, 89*(3), 367-380.

Dewaele, J.-M. (2018). Online questionnaires. In A. Phakiti, P. De Costa, L. Plonsky & S. Starfield (Eds.), *The Palgrave Handbook of Applied Linguistics Research Methodology* (pp. 269-286). Basingstoke: Palgrave Macmillan.

Dewaele, J.-M. & Alfawzan, M. (2018). Does the effect of enjoyment outweigh that of anxiety in foreign language performance? *Studies in Second Language Learning and Teaching, 8*, 21-45.

Dewaele, J.-M. & Dewaele, L. (2017). The dynamic interactions in foreign language classroom anxiety and foreign language enjoyment of pupils aged 12 to 18. A pseudo-longitudinal investigation. *Journal of the European Second Language Association, 1*, 12-22.

Dewaele, J.-M., & MacIntyre, P.D. (2014). The two faces of Janus? Anxiety and enjoyment in the foreign language classroom. *Studies in Second Language Learning and Teaching, 4*, 237-274

Dewaele, J.-M., & MacIntyre, P.D. (2016). Foreign language enjoyment and foreign language classroom anxiety: The right and left feet of the language learner? In P. D. MacIntyre, T. Gregersen, & S. Mercer (Eds.), *Positive psychology in SLA* (pp. 215–236). Multilingual Matters.

Dewaele, J.-M., & MacIntyre, P.D. (2019). The predictive power of multicultural personality traits, learner and teacher variables on foreign language enjoyment and anxiety. In M. Sato & S. Loewen (Eds.), *Evidence-based second language pedagogy: A collection of Instructed Second Language Acquisition Studies*. London: Routledge.

Dewaele, J.-M., MacIntyre, P. D., Boudreau, C., & Dewaele, L. (2016). Do girls have all the fun? Anxiety and enjoyment in the foreign language classroom. *Theory and Practice of Second Language Acquisition, 2*(1), 41–63.

Dewaele, J.-M., Franco Magdalena, A., & Saito, K. (2019a). The effect of perception of teacher characteristics on Spanish EFL learners' anxiety and enjoyment. *Modern Language Journal, 103*, 412-427.

Dewaele, J.-M., Özdemir, C., Karci, D., Uysal, S., Özdemir, E. D., & Balta, N. (2019b). How distinctive is the foreign language enjoyment and foreign language classroom anxiety of Kazakh learners of Turkish? *Applied Linguistics Review, 1,* 1-23.

Dewaele, J.-M., Witney, J., Saito, K., and Dewaele, L. (2018). Foreign language enjoyment and anxiety in the FL classroom: the effect of teacher and learner variables. *Language Teaching Research. 22,* 676-697.

Elahi Shirvan, M. & T. Taherian. (2018). Longitudinal examination of university students' foreign language enjoyment and foreign language classroom anxiety in the course of General English: Latent growth curve modelling. *International Journal of Bilingual Education and Bilingualism, 30, 23-41.*

Fredrickson, B. L. (2001). The role of positive emotions in positive psychology: The broaden-and-build theory of positive emotion. *American Psychologist, 56*, 218-226.

Hamouda, A. (2012). An exploration of causes of Saudi students' reluctance to participate in the English language classroom. *International Journal of English Language Education, 1*(1), 1-34.

Horwitz, E., Horwitz, M., & Cope, J. (1986). Foreign language classroom anxiety. *The Modern Language Journal, 70*, 125-132.

Horwitz, E. K. & Young, D.J. (Eds) (1991). *Language anxiety: From theory and research to classroom implications*. Englewood Cliffs, NJ: Prentice Hall.

Horwitz, E. K. (2001). Language anxiety and achievement. *Annual Review of Applied Linguistics*, *21*(1), 112-126.

Jiang, Y., and Dewaele, J. M. (2019). How unique is the foreign language classroom enjoyment and anxiety of Chinese EFL learners? *System, 82*, 13-25.

Jin, Y. X., & Dewaele, J.-M. (2018). The effect of positive orientation and perceived social support on foreign language classroom anxiety. *System, 74*, 149-157.

Khajavy, G. H., MacIntyre, P. D. & E. Barabadi (2018). Role of the emotions and classroom environment in willingness to communicate: Applying doubly latent multilevel analysis in second language acquisition research. *Studies in Second Language Acquisition, 40*, 605-624.

Li, C., Jiang, G., & Dewaele, J.-M. (2018). Understanding Chinese high school students' foreign language enjoyment: Validation of the Chinese version of the Foreign Language Enjoyment Scale. *System 76*, 183-196.

MacIntyre, P. D. (2017). An overview of language anxiety research and trends in its development. In C. Gkonou, M. Daubney, & J.-M. Dewaele (Eds.), *New insights into language anxiety: Theory, research and educational implications* (pp. 11-30). Bristol: Multilingual Matters.

MacIntyre, P.D., & 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.

MacIntyre, P. D., & Legatto, J. J. (2011). A dynamic system approach to willingness to communicate: Developing an idiodynamic method to capture rapidly changing affect. *Applied Linguistics, 32*, 149-171.

MacIntyre, P.D., & Mercer, S. (2014). Introducing positive psychology to SLA. *Studies in Second Language Learning and Teaching, 4*, 153-172

Mierzwa, E. (2019a). Foreign Language learning and teaching Enjoyment: Teachers' Perspectives. Journal of Education *Culture and Society, 2*(10), 170-188.

Mierzwa, E. (2019b). Foreign language enjoyment among English Philology students: what do students enjoy while learning English as a FL? *Theory and Practice in English Studies, 8*(1), 7-21

Oxford, R. (2017). Anxious language learners can change their minds: Ideas and strategies from traditional psychology and positive psychology. In C. Gkonou, M. Daubney, & J.–M. Dewaele (Eds.), *New insights into language anxiety: Theory, research and educational implications* (pp. 179-199). Multilingual Matters.

Park, G. P., & French, B. F. (2013). Gender differences in the foreign language classroom anxiety scale. *System, 41*, 462-471.

Pavelescu, L. M. & Petri, B. (2018). Love and enjoyment in context: Four case studies of adolescent EFL learners. *Studies in Second Language Learning and Teaching, 8*, 73-101.

Piniel, K. & Albert, A. (2018). Advanced learners' foreign language-related emotions across the four skills. *Studies in Second Language Learning and Teaching, 8*, 127-147.

Santos, A., Cenoz, J., & Gorter, D. (2017). Communicative anxiety in English as a third language. *International Journal of Bilingualism and Bilingual Education, 14*(1), 23-37.

Sellers, V. (2008). Anxiety and reading comprehension in Spanish as a foreign language. *Foreign Language Annals, 33*, 512-521.

Shi, L. (2008). The successors to Confucianism or a new generation? A questionnaire study on Chinese students' culture of learning. *Language, Culture and Curriculum, 19*, 122-147.

Song, J. (2018). "She Needs to Be Shy!": Gender, culture, and nonparticipation among Saudi Arabian female students. *TESOL Quarterly, 53*(2), 405-429.

Swain M. (2013). The inseparability of cognition and emotion in second language learning. *Language Teaching, 46*, 195-207.

Thompson, A. S., & Lee, J. (2013). Anxiety and EFL: Does Multilingualism Matter? *International Journal of Bilingual Education and Bilingualism, 16*, 730-749.

Zhang, R., & Zhong, J. (2012). The hindrance of doubt: causes of language anxiety. *International Journal of English Linguistics, 2*(3), 27-33.

Appendix: FLE and FLCA Questionnaire

I. Background Information

1. Gender:
2. Age:
3. Year of study:
4. Have you visited an English speaking country?

   Yes.

   No.

5. Have participated in as study abroad program in an English speaking country such as England or the US?

   Yes.

   No.

6. Self-perceived proficiency in English:

On a scale from zero to ten, please select your level of proficiency in speaking English

1 2 3 4 5 6 7 8 9 10

On a scale from zero to ten, please select your level of proficiency in listening in English

1 2 3 4 5 6 7 8 9 10

On a scale from zero to ten, please select your level of proficiency in reading in English

1 2 3 4 5 6 7 8 9 10

On a scale from zero to ten, please select your level of proficiency in writing in English

1 2 3 4 5 6 7 8 9 10

II. To what extent do you agree with the following statements? Strongly disagree/ Disagree /Undecided/ Agree /Strongly agree

A. Foreign Language Enjoyment Scale

1. I don't get bored
2. I enjoy it
3. I'm a worthy member of the Foreign language class
4. In class, I feel proud of my accomplishments
5. It's a positive environment
6. It's cool to know a Foreign language
7. it's fun
8. The peers are nice
9. There is a good atmosphere
10. We laugh a lot

B. Foreign Language Classroom Anxiety

1. Even if I am well prepared for Foreign language class, I feel anxious about it
2. I always feel that the other students speak the Foreign language better than I do
3. I can feel my heart pounding when I'm going to be called on in Foreign language class
4. I don't worry about making mistakes in Foreign language class (reverse)
5. I feel confident when I speak in Foreign language class (reverse)
6. I get nervous and confused when I am speaking in my Foreign language class

7. I start to panic when I have to speak without preparation in Foreign language class

8. It embarrasses me to volunteer answers in my Foreign language class

C. Open questions

a. Describe one specific event or episode in your Foreign language class that you really enjoyed, and describe your feeling in as much detail as you can.

b. Describe one specific event or episode in your Foreign language class that made you really anxious, and describe your feeling in as much detail as you can.

The Language in Language and Thinking

Vivian Cook
Newcastle University
Vivian.Cook@ncl.ac.uk

Abstract

The relationship of language and thinking is starting to receive considerable attention in the field of SLA research under the name of Bilingual Cognition. This paper argues that it needs to be underpinned by a proper foundation in the language side of the relationship: it is dangerous to take language for granted. First it argues for researchers to clearly spell out what they mean by *language*, whether as the general property of human beings, in an abstract sense, as an external reality, as mental knowledge, as social community or as action, each of which has different implications for the relationship of language and cognition. Then it argues for the Language Commitment to an adequate theory and description of language as a basis for research, claiming that current emphasis is too much on isolate semantic categories rather than syntactic categories and on word-referent mapping rather than the full complexity of lexical meaning.

Keywords: Language and thinking, linguistic relativity, SLA research, Linguistic Commitment, thinking for language

Resumen

La relación del lenguaje y el pensamiento está comenzando a recibir una atención considerable en el campo de la investigación de segundas lenguas bajo el nombre de Cognición Bilingüe. Este artículo argumenta que debe estar respaldado por una base adecuada en lo que respecta a la lengua en esta relación: es peligroso dar por sentada la lengua. Primero, defiende que los investigadores expliquen claramente qué entienden por lengua, ya sea como propiedad general de los seres humanos, en un sentido

abstracto, como realidad externa, como conocimiento mental, como comunidad social o como acción, cada uno de los cuales tiene diferentes implicaciones para la relación de la lengua y la cognición. Posteriormente, defiende el Compromiso de la Lengua con una teoría adecuada y una descripción del lenguaje como base para la investigación, alegando que el énfasis actual está demasiado centrado en las categorías semánticas aisladas en lugar de en las categorías sintácticas y en el mapeo de palabra-referencia en lugar de en la complejidad total del significado léxico.

Palabras clave: lengua y pensamiento, relativismo lingüístico, investigación en ASL, Compromiso Lingüístico, pensar para la lengua

1. Introduction

The 1990s heralded a renaissance of research into the relationship between language and thinking, alias linguistic relativity, basically establishing that speakers of different first languages behave differently on a variety of cognitive tasks such as spatial orientation (Levinson 1996), classification of objects (Lucy 1992) and perception of colours (Davidoff, Davies & Roberson 1999). Recently this has started to affect second language acquisition (SLA) research through a spate of books (De Groot 2010; Han & Cadierno 2010; Pavlenko 2011; Cook & Bassetti 2011), coming to be called the field of bilingual cognition. The current paper is an attempt to look at some of the underlying issues of the relationship between language and thinking in the context of second language acquisition. A survey of the bilingual cognition research itself is available in Bassetti and Cook (2011).

The discussion of the relationship between language and thinking inevitably hinges on what the words *language* and *thinking* are taken to mean. There are doubtless as many linguistic models of language as there are psychological theories of thinking. Investigating the relationship between language and thinking therefore means being explicit about the nature of both language and thinking and being aware of the alternative versions of both. Any language and thinking research in effect tests whether an aspect of cognition relative to a particular psychological theory correlates with an aspect of language relative to a particular linguistic theory.

But there are obvious dangers in such marriages of convenience. The two partners need enough in common to have a relationship: the theory and description of cognition and the theory and description of language have to be compatible; the research has to choose relevant cross-disciplinary aspects of language and cognition out of the many possibilities that present themselves. Yet they also need to be sufficiently different for each to make their own contribution; psychological theories that deny a distinctive

role to language have little relevance as they deny the relationship in advance. Equally linguistic theories that put impermeable barriers between language and cognition have little to contribute.

Linguists and psychologists are both primarily responsible for their own side of the garden fence and may deny any duty to look at it from their neighbours' side. Both linguists' views of thinking and psychologists' views of language are undoubtedly regarded as naïve by their counterparts in the other discipline. Yet research into language and cognition depends on balancing both. To be credible, the language and thinking debate has to be couched in terms that are sound for both sides. Evans (2011: 71) talks of the Cognitive Commitment to use 'a characterization of language that accords with what is known about the mind and brain from other disciplines'. This paper argues the parallel need to pin down the language side of the relationship, called the Language Commitment in Cook (2011). Though this is central to the structure-centred approach to language and thinking (Lucy 1997), a working concept of language is still necessary whichever approach is adopted. The paper tries to make two simple points: language is many things to many people; research connecting language to thinking should use an adequate analysis of language.

2. Meanings of *language*

A starting point is the English word *language*, virtually taken for granted in most discussion. Yet *language* is not a word with a single unambiguous interpretation: it means what a particular theory says it means. Nor are the meanings of the English word *language* necessarily found in other languages: compare say the French distinction between *langue, langage* and *parole* (de Saussure 1916/1976). Similar problems occur with the English word *cognition*, which has no equivalent in Polish (Wierzbicka 2011). A fatal trap in language or cognition studies is to assume that English can be the neutral language for describing either, prevalent for example in Whorf's habit of translating Hopi notions into English (Whorf, 1941a/1956).

Cook (2007; 2010) has distinguished different meanings of *language* in English, seen in the table below, which will be used to organise the discussion here. These are working definitions rather than watertight boxes and doubtless overlap and contradict each other in various ways; they are discussed more fully in Cook (2010).

Table 1. Six meanings of language (based on Cook, 2007; 2010)

$Lang_1$	a human representation system	'human language'
$Lang_2$	an abstract external entity	'the English language'
$Lang_3$	a set of actual or potential sentences	'the language of Shakespeare'
$Lang_4$	the possession of a community	'the language of English people'
$Lang_5$	the knowledge in the mind of an individual	'I know English'
$Lang_6$	a form of action	'language is doing'

— The **$Lang_1$** sense of *language* as a representation system treats it as a possession of human beings: 'a species-unique format for cognitive representation' (Tomasello 2003: 13). $Lang_1$ *language* is an uncountable noun; you have $Lang_1$ language or you don't but you don't have *a* $Lang_1$ language. Here the language/thinking relationship is at its most general: does human language itself have a connection to human thinking rather than any particular language? Lucy (1997: 292) describes this as a semiotic level at which 'speaking any natural language at all may influence thinking'. In this sense, human beings themselves are incapable of examining the links between their language and their thinking: only a non-human intelligence not bound by human language and thinking might be able to detect them.

In the multi-competence perspective, most of the human race are seen as possessing *languages*, not *language*; all human beings have the initial potential to acquire more than one language (Cook 2009). It cannot be assumed that a typical human being knows only a single language with a single grammar, a single mental lexicon, and so forth. The question is not so much how knowing a single language relates to our thinking as how knowing two or more languages relates to thinking. Perhaps the general links of $Lang_1$ to thinking occur regardless of how many languages the person knows; perhaps, however, the second language (L2) user connects language to thinking in ways that monolinguals are not capable of.

— The **$Lang_2$** sense 'an abstract external entity' sees *language* as existing in the world of abstractions; it is described in rulebooks such as the dictionary and the grammar. So the English language is codified in grammars such as the *Comprehensive Grammar of the English Language* (Quirk et al 1972) and dictionaries such as the *Oxford English Dictionary* (OED 1997); seven governments agreed in 1990 on a protocol for reforming the spelling of

Portuguese. Anything that has ever occurred in the language anywhere is part of $Lang_2$: every recorded word since 1150 AD is in principle present in the pages of the OED.

$Lang_2$ is dangerously confusable with the individual's knowledge of language, described below as $Lang_5$. For $Lang_2$ is not the same as individual knowledge in scope – who could actually know the 430 meanings of the word *set* included in the *OED* or the contents of all 1779 pages of the *Comprehensive Grammar?* Nor does it correspond to any individual's actual usage; at best this will comprise a small subset of the words and grammatical rules in the $Lang_2$ language. In addition, it sweeps dialect speakers under the carpet in favour of a single standard, usually the variety spoken by a status group from one area or one class. Nor does the form of the grammarian's description have a necessary connection to the knowledge stored in the brain; it is highly unlikely that the rules of the grammar-book correspond directly to the systems and processes in individuals' minds, let alone to the way they are stored in their brains.

$Lang_2$ is *language* as a countable noun; there is *the English language*, *the French language* ... up to the 6,909 living languages in *Ethnologue* (Lewis 2009). Hence the question can be asked whether users of $Lang_2$ A think differently from users of $Lang_2$ B, just as it might be asked whether people in countries with Common Law legal systems behave differently from those in countries with the Napoleonic Code. This is a matter of correspondence between an idealised language object and idealised thinking, typical of early suggestions from Whorf about Hopi and Algonkian languages (Whorf 1941a/1956).

The prime source of insights for one type of $Lang_2$ analysis is the mind of the investigator, supplemented by observations of people, texts or experiments, as in the great English grammars of Jespersen (1933) and Zandvoort (1957), rather than large corpora of texts. An outstanding modern exponent is Talmy, whose analyses of language deal with an ideal object and usually with more than one language; the evidence in Talmy (2005) consists of thought experiments describing boards lying across streams and personal communications about Dutch and Makah rather than descriptive grammars or sentences collected in a corpus. Such analysis depends on the brilliance of its instigator, as in Talmy's (1985) seminal observation of the distinction between verb-framed languages like Spanish that express path and motion separately *entra caminando* 'he enters walking' and satellite-framed languages like English that express it though particles *he walked in.* This has become a pillar in the language/thinking debate, as we see from many of the papers in Pavlenko (2011) and Han and Cadierno (2010). It is clearly an insight about a $Lang_2$ abstract entity.

$Lang_2$ entities have always been seen as single languages rather than in-between languages like pidgins, since the eighteenth century often identified with a nation-state (Anderson 1983). Corpora and grammars are now starting to emerge for varieties of English such as English as Lingua Franca (ELF) that do not have native speakers (Seidlhofer 2004). However, these are descriptions of native-speaker-less languages, not authoritative statements of the $Lang_2$ of a multilingual nation or group, and have indeed been attacked for undermining the 'standards' of such national communities. The $Lang_2$ sense is equally remote from the L2 user; an individual L2 user no more knows a $Lang_2$ in either language than a monolingual.

– **$Lang_3$** is language as 'a set of actual or potential sentences': 'the totality of utterances that can be made in a speech-community' (Bloomfield 1926/1957: 26): a language is a corpus of sentences that have been spoken or written. This approximates to the sense of *language* in usage-based connectionist and emergentist studies in psychology and Conversation Analysis. This can be opposed to the Chomskyan notion that the goal of a grammar is to describe all the sentences that could be spoken or written – the creative aspect of language use (Chomsky 1972: 100) – i.e. the potential sentences rather than the actual sentences that happen to have been spoken. $Lang_3$ descriptive grammars such as Biber et al (1999) reflect the properties of actual corpora; the COBUILD dictionary (1995) reports meanings from actual usage rather than the complete definitions found in the OED. The question of whether the set of sentences constituting language A goes with different cognition from the set constituting language B is, however, virtually unanswerable as language is being treated as a objective external object, not as an internal mental reality, and so the possibility of thinking as such does not arise without involving other senses.

– The **$Lang_4$** sense is *language* as 'the possession of a community': 'The *mental individuality* of a people and the *shape of its language* are so intimately fused with one another, that if one were given, the other would have to be completely derivable from it' (Humboldt 1836/1999: 46). $Lang_4$ is shared among a group of speakers such as 'the English-speaking world' or 'native speakers of Chinese'. A language community is often equated with a nation – people born in Japan tend to speak Japanese. Nevertheless, Sapir (1921: 179) insists 'It is impossible to show that the form of a language has the slightest connection with national temperament'. A language community can also be a virtual community unconstrained by political borders (Anderson 1983): Kurdish is spoken in Iraq, Turkey and Iran despite the lack of a country of

Kurdistan. A community may also be multilingual, using several languages for different functions in everyday life. In India for example everyone has to know Hindi and English, plus the local state language if the local state language is neither Hindi nor English, known as the 'Three Language Formula' 3±1 system (Laitin 2000).

The question for thinking and language research is then whether the social interaction of $Lang_4$ and the sense of identity it promotes in its speakers link in some way to their thinking: 'languages reflect cultural preoccupations and ecological interests that are a direct and important part of the adaptive character of language and culture' (Evans & Levinson 2009: 436). It might be comparatively easy to demonstrate this in terms of social interaction. The terms of respect in Japanese or the alternative pronouns for social status in Thai undoubtedly go with particular social values. Language is bound to mirror the way the society functions. But does language lead or follow? The disappearance of the Old English word *sweostersunu* ('sister's son', i.e. nephew) attests to the decreased importance of the uncle/nephew bond but is hardly responsible for it.

Extending $Lang_4$ to L2 users raises a long contested issue of how community goes with language. Does the individual L2 user effectively belong to two communities or do they belong to a different community specifically of L2 users, as in the case of ELF? On the one hand Mackey (1972: 554) claims 'An individual's use of two languages supposes the existence of two different language communities; it does not suppose the existence of a bilingual community'; on the other Brutt-Griffler (2002) proposes 'the multi-competence of the community' and Canagarajah (2007: 930) insists 'Linguistic diversity is at the heart of multilingual communities.' The $Lang_4$ link between language and thinking needs to take multilingual communities into account, not just those which employ a single language.

— The **$Lang_5$** sense of *language* refers to an individual's mental knowledge: 'a language is a state of the faculty of language, an Ilanguage, in technical usage' (Chomsky 2005: 2). The classic competence/performance distinction (Chomsky 1965; 1995) in part distinguishes $Lang_5$ mental knowledge from $Lang_3$ sets of sentences in so far as performance refers to 'the actual use of language in concrete situations'.

This sense of *language* is perhaps the one most used in language/thinking research, hardly surprisingly since it involves a concept of mind: the crucial connection is between individual mental knowledge of language and

individuals' thinking, as seen in say 'adult L2 acquisition is a process of establishing form-meaning connections' (Ekiert 2010: 125). In methodological terms, a description of $Lang_5$ individual knowledge is neither the same as the general rules of the language described in $Lang_2$ nor as the pattern regularities in the individual's $Lang_3$. The valid relationship is that between the individual's linguistic competence and the individual's thinking.

There is often therefore a tension between *language* as an abstract codified system, as a community possession and as knowledge in an individual mind. Many see the community $Lang_4$ and the individual $Lang_5$ meanings as two sides of the same coin: 'although languages are thus the work of *nations* ... they still remain the self-creations of *individuals* ' (Humboldt 1836/1999: 44). The slogan of interactionist psychology was B=f(P, E) (Lewin 1936; Cook 1981) – behaviour is a function of Person and Environment, translatable into 'language is both internal $Lang_5$ and external $Lang_4$'. Yet whole theories of linguistics have based themselves exclusively on the internal $Lang_5$ meaning, just as whole theories of psychology have based themselves on the external $Lang_3$ sense, whether behaviourism, connectionism or emergentism.

The importance for the language-cognition debate is the relationship between the two language systems in the same mind. Cook (2003) talked of an integration continuum between the poles of total separation and total integration of the two languages, related to Weinreich (1953)'s distinction between coordinate and compound bilingualism. Neither pole is completely achievable: total separation is impossible since the two languages coexist in the same mind; total integration is impossible as the speaker would be unable to control which language they were speaking. The point on the continuum at a given moment varies according to the aspect of language involved, say phonology, syntax or semantics, and the choice of bilingual or monolingual mode (Grosjean, 1998). So, establishing the $Lang_5$ of an individual may be trickier when the more complex language system of the L2 user is involved. While it would be attractive to take the well-known Chomskyan metaphor that we need to study the pure water of monolingualism rather than sample the polluted waters of the River Charles of bilingualism, this does not work if bilingualism is everyone's potential state and if bilingual minds are more common and more typical than monolingual ones.

– $Lang_6$ is 'language as action', as expressed say in Schegloff et al (2002: 5), 'People use language and concomitant forms of conduct to *do* things, not only to transmit information', long part of the British school of linguistics (Malinowski 1923; Halliday 1978) and an important constituent of Vygotskyan

theory in that language development consists of the child internalising external social action (Vygotsky 1934/62).

Do the things that people do with language differ from one language to another? A major assumption of 1970s communicative language teaching was that people had to be taught to complain or to argue rather than taught the vocabulary and grammar for complaining or arguing. This survives in the Common European Framework (2001) for language teaching devised by the Council of Europe, which measures language proficiency through 'can-do' statements such as 'I can order a meal' and 'I can recognise familiar names'. One interpretation of $Lang_6$ for language/thinking research would be to see how different ways of speaking linked with different ways of acting, corresponding to the behaviour-centred approach (Lucy 1997).

The relevance of L2 users for the language/thinking debate again concerns whether we think monolingual action or L2 user action is the norm. Businessmen like the Arab and Danish L2 users of English described in Firth (2009) are not using English in the same way as their monolingual colleagues; they carrying out L2 acts that do not necessarily have monolingual counterparts and which they might well be incapable of doing in their first language. Similarly, codeswitching is a resource available to L2 users that enables them to make subtle points about social status and topic by changing language. Before linking L2 users' language to their actions and their thinking, we need to know what is distinctive about them that is lacking from those of monolinguals.

To sum up, the links between language and thinking vary according to the sense of *language* being used. A claim about abstract standard $Lang_2$ is not the same as one about social community $Lang_4$ or individual mental $Lang_5$, particularly if L2 users and their communities are taken into account. For example, Sapir says 'we see and hear and otherwise experience very largely as we do because the language habits of our community predispose certain choices of interpretation' (Sapir, cited in Whorf (1941a/1956: 134)). Perhaps language/thinking research needs to take all of these aspects of language into account; perhaps it needs to concentrate on one or two. But it cannot duck the responsibility to specify exactly which meaning or meanings of *language* it is utilizing and whether one language is involved or more than one.

Take the idea of grammar. The grammar of the $Lang_2$ grammar book is not actually known to individual speakers. Firstly, it is a generalization, based either on large numbers of sentences and people or on the insights of gifted

grammarians; secondly, it includes all the rules of 'the language' rather than the subset of the rules any individual might use; thirdly, it is unlikely to be in a form that actually corresponds to the mental $Lang_5$ knowledge of syntax in an individual's mind; fourthly, being based on monolinguals, it does not include complex language systems known and used by L2 users. Many examples over the years have shown how out of step these different 'grammar's can be; the $Lang_2$ book grammar of English will insist on the object case in non-subject positions yet a recent UK Foreign Secretary said *people like you and I*; such crucial Chomskyan test-cases of syntax as the differences between *eager/easy to please* and the subjacency principle in **Who is that she met worrying?* turn out not to apply to many of the native English-speaking population: the standard institutional $Lang_2$ grammar reflects the $Lang_5$ knowledge of individual speakers imperfectly. Any syntactic property that is tested for fit against one person's thinking needs to come out of the $Lang_5$ mental grammar of that person, not out of a general $Lang_2$ abstract grammar.

3. Aspects of language in relationship to thinking

The word *language* has many facets, reflected in the spectrum of linguistic disciplines from phonology to syntax to sociolinguistics, and a dozen more. None of these specialities boast a single uncontested theory or an agreed set of descriptive terms and categories. Phonology encompasses theories as diverse as optimality theory and lexical phonology, some relying on terms such as *phoneme, distinctive feature* or *constraint ranking,* some rejecting them out of hand. The relationship of language to thinking depends not only upon the meaning of *language* employed but also on the specific aspect of language involved and the methods of analyzing it. The links between optimality phonology and thinking are certainly likely to be different from those between componential semantics and thinking. Out of the thousands of possible aspects of language, one or two are bound to favour or deny a link to some equally arbitrary aspect of thinking by chance, unless the choice is severely constrained by the theory or area of language. Research into language and thinking needs careful specification of the aspects of language that are used.

3.1. Syntactic categories and grammar

According to Whorf, the categories of noun and verb make Standard Average European thinking divide the world into things and actions, unlike the thinking of Hopi speakers; 'English and similar tongues lead us to think of the universe as a collection of rather distinct objects and events corresponding to words... as goes

our segmentation of the face of nature, so goes our Physics of the Cosmos' (Whorf, 1941b/1956: 241). His nouns and verbs are defined in almost the traditional terms of school grammar, represented, say, by Cobbett (1819: 12-15), 'Nouns are the names of persons and things. .. Verbs express all the different actions and movements of all creatures and of all things, whether alive or dead'.

Linguists used to inveigh against such semantic definitions of parts of speech on the grounds of their woolliness (the noun *fire* is hardly an object, the verb *seem* hardly an action), and their overlap (*request* is both a noun and a verb, *up* can be a preposition *up the hill*, a noun *ups and downs* and a verb *Up the grant!*). No two people would agree on whether many words were nouns or verbs by these subjective definitions; at best there is a statistical tendency that a noun will refer to an object, a verb to an action (Lyons 1966). Instead linguists have preferred to define syntactic categories in terms of formal structural properties; Fries (1952) defines words that can be inflected for number and can be preceded by a determiner as Word Class 1; i.e. nouns are defined structurally rather than notionally. Other linguists have treated noun and verb as primitive terms – 'certain fixed categories (Noun, Verb, etc.) can be found in the syntactic representation of the sentences of any language' (Chomsky 1965: 28); these are then substantive universals – 'items of a particular kind in any language must be drawn from a fixed class of items' (Chomsky 1965: 28) – a built-in aspect of the human mind. Even Sapir (1921: 96) has a universal leaning: 'There must be something to talk about and something must be said about this subject of discourse... No language wholly fails to distinguish noun and verb, though in particular cases the nature of the distinction may be an elusive one'. Oddly enough many of those who reject the Chomskyan idea of innate universals of language are quite happy to use nouns and verbs (and indeed word) as self-evident universal categories. But even Chomsky's definition refers to syntactic properties; it is an empirical question whether there are two groups of word-classes that can be isolated in all languages on the grounds of syntactic properties (Robins 1952).

The view that nouns and verbs relate to things and actions respectively derives more from the semantics-based school tradition of grammar-teaching than from the linguistics tradition, in which the connection between parts of speech, i.e. syntactic categories, and thinking starts from their syntactic nature; as Sapir puts it, 'A part of speech outside of the limitations of syntactic form is but a will o' the wisp' (Sapir 1921: 96). Linking syntactic categories to thinking would involve first demonstrating that categories such as noun and verb are not universal, as implied say in Chomskyan theory but refuted by their absence in Straits Salish (Jelinek (1995) cited in Evans & Levinson (2009)) or by Nootka studied by Sapir himself (Robins 1952), and secondly correlating thinking with syntactically defined categories: what does syntactic rather

than semantic nouniness correspond to in thinking? Researching the relationship of a syntactic category to thinking is not the same as investigating the effects of a semantic feature of a group of nouns. When Evans & Levinson (2009: 000) say 'each word class we add to the purported universal inventory would then need its own accompanying set of syntactic constraints', they are putting the cart before the horse: a word class is precisely the label for a set of words that have the same syntactic constraints, not a concept in search of its syntactic characteristics.

Much language/thinking research with nouns has concentrated on the article system, alias determiners. In English the written and spoken forms of the articles are: *the* /ðɪ~ðə/, *a~an* /eɪ~ɘ~æn~ɘn/, and Ø (zero article, i.e. the lack of an article), which precede the noun. Put simply, the main English article system is a complex intertwining of:

— definite/indefinite *the man/a man* (singular), *Ø the men/men* (plural),

— countable/uncountable *a man/Ø man,*

— singular/plural *a man/the man/the men/Ø men*

— first/second mention *a man came in/the man spoke.*

Some uncountables can be quantified with classifier-like phrases, *a cup of tea* or *a pile of sugar*. Subtle everyday uses abound, say the function/place difference between *he went to hospital* and *he went to the hospital,* or the Ø article before certain professions; *he's professor of English at UCL.* And phonology also enters in through the meaning difference between stressed and unstressed forms; *John is the /ðə/ man to watch* versus *John is the /ðɪ:/ man to watch.* A more extensive discussion of English articles in a language and cognition context can be found in Ekiert (2010).

The English article system is notoriously hard for speakers of Chinese and Japanese to acquire, examples include mistakes such as *a research* and *an evidence* familiar to all university teachers. But it is not just mass versus count nouns that gives students problems nor the absence of articles from their first languages so much as the sheer complexity of the English article system, attested by the vast literature on the L2 acquisition of English articles, say Butler (2002), Master (1994) and Thomas (1989). Evans (2011) makes a three-way distinction between *a* introducing 'a referent which the hearer is held to be unable to readily identify', *a* designating 'a unitary instantiation of the referent' and *the* introducing 'a referent which the hearer is held to be able to readily identify'. This takes in the discourse context but does not mention the zero article, equally part of the article system, i.e. the plural form of *a* with unidentifiable nouns is zero *a man came in/Ø men came in.* And it does not include the other uses of *a*.

It is important then that, in any language and cognition research, the linguistic elements are properly described. In a typical example Imai and Gentner (1997) looked at how speakers of classifier languages like Japanese classify objects differently from speakers of non-classifier languages like English. The crucial point is 'the grammatical distinction between count nouns and mass nouns', i.e. countability, which exists in English but not in Japanese. This distinction is signalled by the presence of *a* before *book*, *a book*, and its absence before *sand*, **a sand*, and the plural form *books* but not **sands*. Syntactically this analysis reduces a complex syntactic meaning-related formal system to an either/or semantic choice. The Linguistic Commitment requires full syntactic analyses for the determiner systems of both languages, as is for instance carried out in Lucy (1992). A similar reduction of complex syntactic matters to an either/or choice is seen in Talmy's satellite/verb-framed distinction, which appears to be a matter of frequency of occurrence within a language rather than an either/or distinction across languages (Croft,2010).

So the analysis in language and thinking research usually establishes two classes of nouns, count and mass, on semantic grounds rather than as syntactic features of nouns that vary continuously (Bollinger 1969). It is always possible to create countable forms for mass nouns with specialised meanings, *the sands of time, the waters of the Nile*, or uncountable forms for count nouns *there is too much book in school*. Most language and cognition research is based on the meanings associated with nouns rather than on their formal syntactic properties. Where Imai & Gentner (1997) claim to be talking about 'count/mass syntax', they are actually talking about count/mass semantics: the syntax is far more complex – whichever syntactic model might be used.

Nor are nouns the only syntactic category to be treated semantically rather than syntactically in this research. Coventry et al (2011) for example explored how English/Spanish bilinguals express spatial relationships through prepositions such as *in* glossed as 'containment' and *on* glossed as 'having support'. The linguists Leech & Svartvik (1975: 83-85) talk, not of a pair of prepositions *in/on*, but of '*in*-type' prepositions that include *into, in, out of, through* involving an area or volume, and '*on*-type' prepositions such as *onto, on, off, across, along* involving a line or a surface; volume and surface contrast in *He was living on a desert island* and *He was born in Cuba*. As with nouns, the linguist sees a complex network of meanings, usually finding it difficult to make hard and fast divisions between pairs such as *in* and *on*. Testing out the meanings of pairs of prepositions against thinking reduces prepositions to isolated atoms of meaning rather than complex molecules; prepositions contrast with many other prepositions not just with one, *in a plane, on a plane, inside a plane, with a plane, inside a plane, by plane*, and so on, each with a meaning that contrasts with the others to a greater or lesser extent. Most uses of *in* and *on* are in fact temporal, *in the afternoon* and *on Tuesday*, not

spatial. Other uses are directional, *arrive in* and *on your left.* The containment/support distinction is thus one small facet of the use of *in* and *on*. Our knowledge of *in* and *on* is many-facetted; even beginners in the language need to know multiple meanings for *in* and *on*. Do 'container *in*' and 'surface *on*' stand out sufficiently from this web of meanings to be testable cross-linguistically? Levinson (2003: 38) feels 'in general it is hard to find any pair of spatial descriptors with the same denotation across languages'.

Primarily to linguists, preposition is a structural category, like noun, defined by its appearance before noun phrases – *in a moment, on the train*; 'Typically, prepositions function as the first constituent of a prepositional phrase' (Greenbaum 1996: 159). Comparing two languages is not just taking in certain core meanings of the prepositions but also looking at how they behave syntactically, at the extreme with the postpositions of Japanese that occur after the noun phrase *nihon ni* (Japan in) rather than before *in Japan*. Prepositions have a complex set of syntactic behaviours in English, going with particular verbs *come on/come in* and adjectives *dependent on/ disappointed in.* Prepositions are not independent units like content words but have a syntactic function like articles; they are partly like closed-class function words, partly like open-class content words. Unless cross-linguistic comparison takes such basic information on board, research results may be skewed by the other syntactic and semantic features of preposition oppositions like *in* and *on* across languages. Again language and cognition research tends to deal with them as semantic units rather than syntactic categories.

The major syntactic differences between languages have, oddly enough, seldom featured in the language/thinking debate, for example the contrast between configurational languages which have phrase structure and non-configurational languages which do not (Hale 1983) and the preposition/postposition word order difference seen in English and Japanese. Whorf made the interesting claim that 'Hopi can and does have verbs without subjects, a fact which may give that tongue potentialities, probably never to be developed, as a logical system for understanding some aspects of the universe' (Whorf, 1941b/1956: 243). Since Chomsky (1981), this has been known as the prodrop or null subject parameter; 97% of languages allegedly have verbs without surface Subjects, the well-known exceptions being English, German and modern written French. Yet no-one appears to have correlated such a well-attested syntactic difference with differences in thinking. One exception where a syntactic difference per se was indeed correlated with a thinking difference was Bloom (1981)'s controversial attempt to link the presence of *if*-clauses in English such as *If he had not gone to the forum that day, Caesar would have lived longer* with the ability to reason counterfactually. And indeed, crucially for the Linguistic Commitment, we need to establish an overall syntactic theory within which the languages can be compared on

an equal footing (Stringer 2010), rather than using English as an unacknowledged universal system into which other languages can be converted (Wierzbicka, 2011).

The relationship between language and thinking has then often been conceived in terms of semantics, not of syntax, of the meanings of words and sentences, not of their form, as if syntax were an unimportant appendix to language. Syntax is seldom treated in its own right as having syntactic form and meaning but seen as a vehicle conveying a basic set of semantic meanings; gender for instance is not seen as syntactic agreement between elements in the sentence but as semantic and arbitrary meanings attached to nouns. Yet the way that we think could be as influenced by the way we construct our sentences as by what we mean by our words, as the writing-direction research of Tversky et al. (1991) suggests. It is one thing to deduce a relationship between syntax and thinking based on a formal feature of syntax, whether the effects of SVO, the syntactic categories of noun and verb or the presence or absence of articles; it is another to deduce a relationship from a semantic idea incorporated in a complex web of syntactic devices, such as gender and mass/count. What is being asked for is then in the spirit of Brown (1986: 482): 'Relativity is the view that the cognitive processes of a human being – perception, memory, inference, deduction – vary with the structural characteristics – lexicon, morphology, syntax – of the language he speaks'. This is not the same as saying that semantic notions expressed in syntax correlate with concepts in cognition – of course they do or language would be impossible.

3.2. Lexical items and sets

The bridge between language and thinking has usually been perceived as the lexicon. Famously Whorf (1940/1956: 216) based arguments for linguistic relativity on the many words for snow in Inuit, the scarcity of 'snow' words in English and the reduction of 'cold', 'ice' and 'snow' to one word in Aztec. Yet Sapir (1921: 181) pointed out 'the linguistic student should never make the mistake of identifying a language with a dictionary'.

The area of colour research investigates how the set of colour terms in the lexicon of a language relates to 'objective' measures of colour such as hue and saturation, as seen in Athanasopoulos (2011). As with grammar, there are a range of linguistic theories and tools for handling lexis and its importance varies according to one's theory. For instance, while Chomsky's theory is often seen as syntactic, its instantiation in the Minimalist Program treats syntactic structure as a projection from the properties of lexical items, which are then Merged (Chomsky 1995).

Overall, however, there is a major slippage between the way that many psychologists see vocabulary and that adopted by most linguists. One preliminary

difference is already seen in the term *lexical item* used in the heading. The category of word prevalent in psychology is extremely hard to define except in the writing-based sense of letters surrounded by spaces; what is a word in a language like *Illujuaraalummuulaursimannginamalittauq* 'But also, because I never went to the really big house'? (Dorais 1988: 8, cited in Genesee 2003) The unit preferred by linguists is more often lexical item or lexical entry, which may well be more than one word *look up*, say, or *in front of*, or lemma in which the various alternate forms of the word are reduced to one; a lexical entry contains far more information than a word's reference, such as argument structure and collocability. Mental $Lang_5$ lexicons are organised in lexical items or entries, not in the words typically found in $Lang_2$ dictionaries.

Psychologists by and large regard words as having single core meanings: *dog* means , not 'to follow someone closely' or 'a complete flop': one word, one meaning. Levinson (2003: 35) sums this up as the assumption that 'corresponding to a lexical item is a single holistic concept'. *Red* and *blue* are taken to refer to particular colours/hues/saturation on some visual scale or field rather than say to the Labour and Conservative parties, snooker balls or York City Football Club. Evans (2011: 75) for instance distinguishes *red ink* from *red squirrel*. This tells us rather little about the wider set of colour terms that *red* belongs to: in the case of squirrels, the opposition is between *red* and *grey*, in ink probably between *red, blue, black* and *green*, in roulette between *red* and *black*, in traffic lights between *red*, *green* and *amber*, and so on. A word has many meanings in different contexts and relates to the meanings of many other words.

More crucially treating *red* as a unique visual perception only covers its denotative reference to physical colour rather than its many other uses; it assumes that there is a core aspect of the physical world denoted by the colour term however much this may vary cross-linguistically. Defining the language component of language and thinking is taken to mean enumerating which physical shade on the colour or hue scales etc. the words actually refer to before relating them to different thinking. The central belief is that the purpose of words is refer to things and actions; hence the emphasis in psychologists' research is on content nouns with clear (and usually drawable or at least visible) physical reference, as opposed to structure words like *to* with syntactic meaning or words with abstract meanings like *nation*. Imai and Gentner (1997) for example frequently refer to 'names' and 'nouns' as if they were the same – 'substance names and object names', 'children initially learn object names rather than names for relations of properties', 'extending their names on the basis of shape'; discussions of colours frequently talk of 'colour naming' (Regier & Kay 2009); Malt and Ameel (2010: 175) talk of naming patterns as 'patterns of application of words to objects', Pavlenko (2010: 200) of 'word-to-referent mapping'. To most linguists the word-object link is only one of the many aspects of meaning that go with a word; there is no simple bond between an object in the world and the noun that refers to it: most nouns are

not names; most nouns have multiple meanings; most meanings do not have clear denotations in the physical word; most meanings relate to other meanings rather than being independent – the basic structuralist claim summarised as 'lexical concepts are defined relative to other items in the lexicon' (Stringer 2010: 102). Indeed in English actual names have their own distinctive grammar with no determiners *London* apart from rare exceptions like *The Hague* and their own spelling conventions such as *Cooke, Hogg* and *Smythe*.

To linguists, not only do words with unambiguous visual referents form a small subset of the lexicon but also reference is only one aspect of meaning. Take the word *blue*. One aspect of *blue* is indeed the peculiar physical shade it refers to. But *blue* exists in a network of other information. For an English person, its five most common associations are *sky, black, green, red, white, sea, colour, yellow, eyes, aristocracy* (Edinburgh Word Association Thesaurus 2011). While this confirms a relationship with other colours on the physical scale, it also relates *blue* to the superordinate level *colour* in Roschian terms (Rosch, 1977) and to adjective-noun collocations, *blue sky/sea/eyes/ blood,* the syntagmatic level of association (Deese 1966). *Blue* is not only an adjective but also a noun, as in *Klein blue* or *Oxford Blue,* or, in the plural, a kind of music *the blues* or the nick-name of Birmingham City and Chelsea football teams. Meanings can be deconstructed into features – *blue* has the feature [+colour] etc; looked at in networks (Cruse 1986) – *blue* is the opposite of *red* in political talk; studied in collocations – *blue joke, blue funk, turn blue*, *black and blue*; and counted – *blue* is 7987th in frequency on the British National Corpus, *red* 11605th.

Doubtless the counter-argument is that *blue* has a central meaning from which all the others derive: see how people relate this meaning to thinking and that's all that needs to be done. But some linguists precisely deny that words have central meanings: polysemy is the norm. Many of the everyday meanings of *blue* do not relate to its supposedly core referential meaning; why should a joke be *blue* (it's yellow in Chinese)? An intellectual woman *a blue stocking*? Rather the word *blue* has multiple meanings and connections, of which its denotative link to a quality in the physical world forms a small part. Relating referential meanings of limited sets like colour words to thinking takes in one small aspect of the multitude of possible relationships between language and thinking. It is not that it is unusual or untypical of words to have such a range of senses: cognitive blends are the norm (Fauconnier, 2003). Perhaps only the simulated situations psychological model encompasses the many aspects of a single word: 'perceptual symbols are multimodal, originating in all modes of perceived experience, and they are distributed widely throughout the modality-specific areas of the brain' (Barsalou 1999: 583) The overall point once again is that any attempt to relate words to cognition needs to be grounded on a rich model of words, not only on single denotative meanings.

4. Conclusion

The intention here was to point out that, despite much exciting and novel work into language and thinking, we are getting nowhere until we have spelled out what *language* means in the context of the particular piece of research. This is not to say that one interpretation of language or one descriptive view of syntax should be preferred; many alternative avenues can be explored. But it is dangerous to take language for granted; lack of explicitness in the discussion often means it falls back by default on folklore and common-sense rather than the scientific study of language practiced in the twenty-first century.

5. References

Anderson, B. (1983) *Imagined communities*. New York: Verso.

Athanasopoulos, P. (2011). Colour and bilingual cognition. In Cook, V. J. & Bassetti, B. (eds.) (2011) Language *and bilingual cognition.* (pp. 241-262) New York, NY: Psychology Press.

Barsalou, L.W. (1999) Perceptual symbol systems. *Behavioral and Brain Sciences,* 22 (4), 577-660.

Bassetti, B. & Cook. V. J. (2011). Relating language and cognition: the second language user. In Cook, V. J. & Bassetti, B. (eds.) (2011) Language *and bilingual cognition.* (pp. 143-190) New York, NY: Psychology Press.

Biber, D., Finegan, E., Johansson, S., Conrad, S. & Leech, G. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.

Bloom, A. H. (1981) *The linguistic shaping of thought: a study in the impact of language on thinking in China and the West.* Hillsdale, NJ: Erlbaum Associates.

Bloomfield, L. (1926/1957) A set of postulates for the science of language. *Language* 2, 15364. Reprinted in M. Joos (ed.), *Readings in linguistics I*. Chicago: University of Chicago Press, 1957.

Bollinger, D. (1969) Categories, features, attributes. *Brno Studies in English*, 8, 37-41.

Brown, R. (1986) *Social psychology*. New York: Free Press. 2nd edition.

Brutt-Griffler, J. (2002) *World English: a study of its development.* Clevedon: Multilingual Matters.

Butler, Y. G. (2002) Second language learners' theories on use of English articles. *Studies in Second Language Acquisition,* 24, 451-480.

Canagarajah, S. (2007) Lingua Franca English, multilingual communities and language acquisition. *Modern Language Journal*, 91, 923-939.

Carroll, J. B. (Ed.) (1956) *Language, thought, and reality: selected writings of Benjamin Lee Whorf* Cambridge, MA: MIT Press.

Chomsky, N. (1965) *Aspects of the theory of syntax*. Boston, Mass.: MIT Press.

Chomsky, N. (1972) *Language and mind*. Enlarged edition. New York, NY: Harcourt and Brace.

Chomsky, N. (1981) *Lectures on Government and Binding*. Dordrecht, Foris.

Chomsky, N. (1995) *The Minimalist Program*. Boston, Mass.: MIT Press New Horizons.

Chomsky, N. (2005) Three factors in language design. *Linguistic Inquiry*. 36, 1, 1-22.

Cobbett, W. (1819) *A grammar of the English language*. Reprinted Oxford University Press 1984.

COBUILD dictionary (1995) *Collins COBUILD English Dictionary*. Harper Collins.

Common European framework of reference for languages (2001). Strasburg: Council of Europe, Cambridge University Press

Cook, V. J. (1981) Second language acquisition from an interactionist viewpoint. *Interlanguage Studies Bulletin–Utrecht*, 6, 93-111.

Cook, V. J. (2003) Introduction: the changing L1 in the L2 user's mind. In Cook, V.J. (ed.), *Effects of the second language on the first*. (pp.1-18) Clevedon: Multilingual Matters.

Cook, V. J. (2007) The nature of the L2 user. In L. Roberts, A. Gurel & L. Marti (eds.) *EUROSLA Yearbook*, 7, 205-220. Reprinted in L. Wei (ed.), *The Routledge applied linguistics reader* (pp. 77-89) London: Routledge.

Cook, V. J. (2009) Multilingual Universal Grammar as the norm. In I. Leung (ed.), *Third language acquisition and universal grammar*. (pp. 55-70) Bristol: Multilingual Matters.

Cook, V. J. (2010) Prolegomena to second language learning. In P. Seedhouse & S. Walsh (eds), *Conceptualising language learning*. (pp. 6-22) Basingstoke: Palgrave MacMillan.

Cook, V. J. (2011) Relating language and cognition: the speaker of one language. In Cook, V. J. & Bassetti, B. (eds.) (2011) *Language and bilingual cognition*. (pp. 3-22) New York, NY: Psychology Press.

Cook, V. J. & Bassetti, B. (eds.) 2011. *Language and bilingual cognition.* New York, NY: Psychology Press.

Coventry, K., Guijaro-Fuentes, P. & Valdes, B. (2011). Spatial language and second language acquisition research. In Cook, V. J. & Bassetti, B. (eds.) (2011*) Language and bilingual cognition.* (pp. 315-340) New York, NY: Psychology Press.

Croft, W. (2010) Relativity, linguistic variation and language universals. *Cognitextes*, 5 http://cognitextes.revues.org/.

Cruse, D.A. (1986) *Lexical semantics.* Cambridge: Cambridge University Press.

Davidoff, J., Davies, I., & Roberson, D. (1999) Color categories in a stone-age tribe. *Nature, 398,* 203-204.

Deese, J. (1966). *The structure of associations in language and thought.* Baltimore: John Hopkins Press.

De Groot, A. M. B. (2010). *Bilingual cognition: an introduction.* New York, NY: Psychology Press.

de Saussure, F. (1916/1976) *Cours de linguistique générale*. Bally, C. & A. Sechehaye (eds.) 1915. Critical edition by T. de Maurio. Paris: Payothèque, Payot.

Dorais, L.-J. 1988. *Tukilik: an Inuktitut grammar for all.* Quebec, QC: Association Inuksiutiit Katimajiit. Cited in Genesee, F). (2003) *Portrait of the bilingual child*. In V. Cook (ed.), *Portraits of the L2 user*. (pp. 170-196) Bristol: Multilingual Matters.

Edinburgh Word Association Thesaurus (2011), http://www.eat.rl.ac.uk/.

Ekiert, M. (2010) The linguistic effects on thinking for writing: The case of articles in L2 English. In Han & Cadierno. *Linguistic Relativity in Second Language Acquisition: Thinking for speaking* (pp. 125-153) Clevedon: Multilingual Matters.

Evans, N. & Levinson, S. (2009) The myth of language universals: language diversity and its importance for cognitive science. *Behavioral and Brain Sciences,* 32, 429-448.

Evans, V. 2011. Language and cognition; the view from cognitive linguistics. Cook, V. J. & Bassetti, B. (eds.) 2011. *Language and bilingual cognition* (pp. 69-108) New York, NY: Psychology Press.

Fauconnier, G. (2003) *The way we think.* New York: Basic Books

Firth, A. (2009) Doing *not* being a foreign language learner: English as a *lingua franca* in the workplace and (some) implications for SLA. *International Review of Applied Linguistics,* 47, 127-156

Fries, C. C. (1952). *The structure of English.* New York: Harcourt Brace.

Genesee, F. (2003). Portrait of the bilingual child. In V. Cook (ed.), *Portraits of the L2 user*. (pp. 170-196) Bristol: Multilingual Matters.

Gentner, D. & Goldin-Meadow, S. (Eds.) 2003. *Language in mind: advances in the study of language and Thought*. Cambridge, MA: The MIT Press.

Goldin-Meadow, S., So, W.-C., Ozyurek, A., & Mylander, C. (2008) The natural order of events: How speakers of different languages represent events nonverbally. *Proceedings of the National Academy of Sciences, 105*, 9163-9168.

Greenbaum, S. (1996) *Oxford English grammar*. Oxford: Oxford University Press.

Grosjean, F. (1998). Transfer and language mode. *Bilingualism: Language and Cognition*, 1, 3, 175-176.

Hale, K. (1983) Warlipiri and the grammar of non-configurational languages. *Natural Language and Linguistic Theory*, 1, 5-47.

Halliday, M. A. K. (1978) *Language as social semiotic*. London: Edward Arnold.

Han, Z. & Cadierno, T. (eds.) (2010) *Linguistic relativity in SLA: thinking for speaking*. Bristol: Multilingual Matters.

Humboldt, W. von (1836/1999) *On language*. Translated by P. Heath. Cambridge: Cambridge University Press.

Imai, M. & Gentner, D. (1997) A cross-linguistic study of early word meaning: universal ontology and linguistic influence. *Cognition*, 62, 169-200.

Jarvis, S. & Pavlenko, A. (2009) *Crosslinguistic influence in language and cognition*. Abingdon: Routledge.

Jelinek, E. (1995) Quantification in Straits Salish. In E. Bach, E. Jelinek, A. Kratzer & B. Partee (eds.), *Quantification in natural languages*. (pp. 487-540) Kluwer.

Jespersen, O. (1933) *Essentials of English grammar*. London: Allen & Unwin.

Laitin, D. D. (2000) What is a language community? *American Journal of Political Science*, 44, 1, 142-145.

Leech, G. & Svartvik, J. (1975) *A communicative grammar of English*. London: Longman.

Levinson, S. (1996) Relativity in spatial conception and description. In J. J. Gumperz & S. C. Levinson (eds.). *Rethinking linguistic relativity*. (pp. 177-202) Cambridge: Cambridge University Press.

Levinson, S. C. (2003) Language and mind: let's get the issue right. In Gentner and Goldin-Meadow (eds.), *Language in mind*. (pp. 25-46) Cambridge, Mass.: MIT Press.

Lewin, K.(1935) *A dynamic theory of personality*. New York, N.Y.: McGraw-Hill.

Lewis, M. Paul (ed.) (2009) *Ethnologue: languages of the world*. Sixteenth edition. Dallas, Tex.: SIL International. Online version: http://www.ethnologue.com/.

Lucy, J. A. (1992) *Grammatical categories and cognition: a case study of the linguistic relativity hypothesis*. Cambridge: Cambridge University Press.

Lucy, J. A. (1997) Linguistic relativity. *Annual Review of Anthropology*, 26, 291-392.

Lyons, J. (1966) Towards a "notional" theory of the "parts of speech". *Journal of Linguistics*, 2, 2, 209-236.

Mackey, W. F. (1972) The description of bilingualism. In Fishman, J. A. (ed.), *Readings in the sociology of language.* (pp. 554-584) The Hague: Mouton.

Malinowski, B. (1923) The problem of meaning in primitive languages. In C. K. Ogden & I. A. Richards, *The meaning of meaning* (pp. 296-336) London: Routledge Kegan Paul.

Malt, B. C. & Ameel, E. (2011). The art and science of bilingual object naming. In Pavlenko (ed.) *Thinking and speaking in two languages* (pp. 170-197) Bristol: Multilingual Matters.

Master, P. (1994) The effect of systematic instruction on learning the English article system. In T. Odlin (Ed.), *Perspectives on pedagogical grammar.* (pp. 229-252) Cambridge: Cambridge University Press.

Oxford English Dictionary (1997) 2nd revised edition. Oxford: Oxford University Press. Online at http://www.oed.com/.

Pavlenko, A. (ed.) 2011. *Thinking and speaking in two languages*. Bristol: Multilingual Matters.

Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1972) *A grammar of contemporary English.* London: Longman.

Regier, T. & Kay, P. (2009) Language, thought and color: Whorf was half right. *Trends in Cognitive Sciences*, 13, 10, 439-446.

Robins, R. H. (1952) Noun and verb in universal grammar. *Language*, 28, 289-298.

Rosch, E. (1977) Human categorisation. In N. Warren (ed.), *Advances in cross-cultural psychology,* 1 (pp. 1-72) London: Academic Press.

Sapir, E. (1921) *Language: an introduction to the study of speech.* Reprinted no date by: Harcourt Brace and Co, New York.

Schegloff, E. A., Koshik, I., Jacoby, S. & Olsher, A. (2002) Conversation analysis and applied linguistics. *Annual Review of Applied Linguistics,* 22, 3-31.

Seidlhofer, B. (2004) Research perspectives on teaching English as a lingua franca. *Annual Review of Applied Linguistics,* 24, 209-239.

Sera, M. D., Forbes, J., Burch, M. C. & Rodriquez, W. (2002) When language affects cognition and when it does not: An analysis of grammatical gender and classification. *Journal of Experimental Psychology: General,* 131, 377-397.

Slobin, D.I. (2004) The many ways to search for a frog: linguistic typology and the expression of motion events. In Strömqvist, S. & Verhoeven, L. (eds.), *Relating events in narrative Vol 2*, (pp. 219-257) Mahwah, N.J.: Lawrence Erlbaum.

Stringer, D. (2010) The gloss trap. In Han & Cadierno (eds.*) Linguistic Relativity in Second language Acquisition: Thinking for speaking.* (pp. 102-124) Clevedon: Multilingual Matters.

Talmy, L. (1985) Lexicalization patterns: Semantic structure in lexical form. In T. Shopen (ed.), *Language typology and lexical description: Vol. 3. Grammatical categories and the lexicon. (pp. 36-149)* Cambridge: Cambridge University Press.

Talmy, L. (2005) The fundamental system of spatial schemas in language. In B. Hamp (ed.), *From perception to meaning: image schemas in cognitive linguistics*. (pp. 199-243) Berlin and New York: Mouton de Gruyter.

Thomas, M. (1989) The acquisition of English articles by first- and second-language learners. *Applied Psycholinguistics*, 10, 335-355.

Tomasello, M. (2003) *Constructing a language*. Cambridge, Mass.: Harvard University Press.

Tversky, B., Kugelmass, S. & Winter, A. (1991) Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, 23, 515-557.

Vygotsky, L. S. (1934/1962) *Thought and language*. Cambridge, Mass.: MIT Press.

Weinreich, U. (1953) *Languages in contact.* The Hague: Mouton.

Whorf, B. L. (1940/1956) Science and linguistics. *Technology Review*, 42, 8, 229-231, 247-248. Reprinted in Carroll (1956*) Language, thought and reality: Selected writings by Benjamin Lee Whorf*. Cambridge (pp. 207-219) MA: Technology Press of Massachusetts Institute of Technology.

Whorf, B. L. (1941a/1956) The relation of habitual thought and behavior to language. In L. Spier (ed.) Language, culture and personality. Reprinted in Carroll (1956*) Language, thought and reality: Selected writings by Benjamin Lee Whorf.* Cambridge (pp. 134-159) MA: Technology Press of Massachusetts Institute of Technology.

Whorf, B. L. (1941b/1956). Languages and logic. Reprinted in Carroll (1956) *Language, thought and reality: Selected writings by Benjamin Lee Whorf.* Cambridge (pp. 233-245) MA: Technology Press of Massachusetts Institute of Technology.

Wierzbicka, A. (1996) *Semantics: primes and universals.* Oxford: Oxford University Press.

Wierzbicka, A. (2011) Bilingualism and cognition: the perspective from semantics. Cook, V. J. & Bassetti, B. (eds.) 2011. *Language and bilingual cognition.* (pp. 191-218) New York, NY: Psychology Press.

Yeh, D. & Gentner, D. (2005) Reasoning counterfactually in Chinese: picking up the pieces. *Proceedings of the 27th Meeting of the Cognitive Science Society*, 2410-2415.

Zandvoort, R. W. (1957) *A handbook of English grammar*. London: Longman.

Agreement morphology errors and null subjects in young (non-)CLIL learners

Yolanda Fernández-Pena
Departamento de Filología
Universidad de Cantabria, Spain
yolanda.fernandezpena@unican.es

Francisco Gallardo-del-Puerto
Departamento de Filología
Universidad de Cantabria, Spain
francisco.gallardo@unican.es

Abstract

There is a wealth of studies on L2 English acquisition in CLIL contexts in Spain, but most have underexplored the potential impact of CLIL in the longer run on the morphosyntax of earlier starters from monolingual regions. This paper fills this gap by exploring agreement morphology errors and subject omission in the oral production of Primary Education English learners from the Spanish monolingual community of Cantabria. The sample investigated consists of the individual narration of a story by learners in two age-matched (11-12 year-olds) groups, one CLIL (*n*=28) and one non-CLIL (*n*=35). The results show no statistically significant differences between both groups for the provision of specific linguistic features at a younger age, though some evidence also points to a subtle effect of additional CLIL exposure. Both groups show moderately low rates of null subjects; they omit affixal morphology (**he eat*) significantly more frequently than suppletive inflection (**he _ eating*) and they seldom produce commission errors (**they eats*). Interestingly, non-CLIL learners show far greater rates of omission with auxiliary *be* than copula *be* and frequently use the placeholder *is* (**he is eat*), which evinces an earlier stage of acquisition than that of CLIL learners.

Keywords: CLIL, L2 English, Primary Education, inflectional morpheme error, null subject

Resumen

En estudios previos sobre la adquisición de inglés como L2 en contextos AICLE en España no se ha explorado en profundidad el impacto potencial de AICLE en la morfosintaxis en estudiantes con una exposición temprana y prolongada a la lengua

meta y en regiones monolingües. Nuestra investigación contribuye a esta línea de investigación explorando los errores de morfología flexiva y la omisión del sujeto en la producción oral de aprendices de inglés de Educación Primaria de la comunidad española monolingüe de Cantabria. La muestra investigada consiste en la narración individual de una historia por aprendices de edades similares de un grupo CLIL (*n*=28) y uno no CLIL (*n*=35). Los resultados no muestran diferencias significativas entre ambos grupos en lo que respecta a la provisión de rasgos lingüísticos específicos a edades tempranas, aunque sí evidencian un ligero efecto de la exposición adicional a AICLE. Ambos grupos muestran tasas relativamente bajas de sujetos nulos; omiten la morfología afijal (**he eat_*) con una frecuencia significativamente mayor que la supletiva (**he _ eating*) y raramente emplean morfemas flexivos de forma errónea (**they eats*). Sin embargo, los estudiantes de programas tradicionales omiten más el auxiliar *be* que la cópula *be* y utilizan más frecuentemente el 'comodín' *is* (**he is eat*), lo que evidencia una etapa de adquisición más temprana que la de los estudiantes en programas AICLE.

Palabras clave: AICLE, inglés L2, Educación Primari,; error de morfología flexiva, sujeto nulo

1. Introduction

Research on the acquisition of L2 English has reported great difficulties and problems with the acquisition of properties that pertain to various domains of language, as is the case of the properties related to the syntax-morphology interface (Gutiérrez-Mangado & Martínez-Adrián, 2018; Montrul, 2011; Slabakova, 2008; White, 2003a). Content and Language Integrated Learning (henceforth CLIL) programmes have been recently implemented in Spain in an attempt to promote and increase learners' proficiency in English through additional exposure to the target language in the curriculum. The main asset of this educational approach is not only the increase in the hours of exposure to the target language but also the input that CLIL learners receive. It is more natural and communicatively more meaningful and authentic, as language is used for interactional purposes (see Gutiérrez-Mangado & Martínez-Adrián, 2018; Lázaro & García Mayo, 2012; Martínez Adrián & Gutiérrez Mangado, 2015b). Research hitherto, however, has confirmed that, while CLIL programmes prove to exert a positive impact on the learners' average proficiency in English (e.g. Jiménez Catalán et al., 2006 and Navés & Victori, 2010 for Primary Education; Lasagabaster, 2008; Martínez Adrián & Gutiérrez Mangado, 2015a; Ruiz de Zarobe, 2008, among others, for Secondary Education), some more specific aspects of the language such as pronunciation (Gallardo del Puerto et al., 2009; Ruiz de Zarobe, 2007) and morphosyntax (Lázaro Ibarrola, 2012; Martínez Adrián & Gutiérrez Mangado, 2009, 2015a; Villarreal Olaizola, 2011) do not seem to benefit so clearly

from CLIL approaches (for a discussion, see Gallardo del Puerto & Martínez Adrián, 2013; Dalton-Puffer, 2008; Martínez Adrián, 2011 and Ruiz de Zarobe, 2011). Two cases in point are the production of agreement morphology errors and null subjects illustrated in (1) to (3) (here 'error' is used for non-native-like forms in consistency with prior generative approaches to the acquisition of L2 morphosyntax (e.g. García Mayo & Villarreal Olaizola, 2009; Villarreal Olaizola, 2011 and Villarreal Olaizola & García Mayo, 2009). For these types of errors, the evidence provided thus far precludes categorical conclusions on a potential advantage of additional CLIL exposure and on a subsequently more target-like performance of CLIL learners over learners receiving only traditional English as a Foreign Language (EFL) lessons.

(1) a. *the mother **prepare** food to the dog [CLIL I: subject 43]

b. *Tim and parent eeh **is** welcomed the house [non-CLIL: subject 169]

(2) *the dad Ø running to the bathroom [CLIL I: subject 50]

(3) a. *because Ø see raining [CLIL II: subject 80]

b. *because Ø is raining [non-CLIL: subject 186]

Our study delves into this issue with a view to shed more light into the potential impact of additional CLIL exposure on the oral production of these types of errors by young (11- and 12-year-old) Spanish learners of L2 English from three Primary Schools in Cantabria, two of them implementing CLIL programmes and the third one providing just traditional EFL instruction. The paper is organised as follows: first, Section 2 reviews previous literature and studies on the acquisition of the English agreement morphology and the obligatory use of overt subjects, particularly in CLIL and non-CLIL contexts. Section 3 describes the aims and the research questions of our study and Section 4, the specifics of the methodology. Section 5 reports and discusses the data obtained from the analysis of both agreement morphology errors and null subjects. Finally, Section 6 presents the main conclusions and the limitations of the study, and suggests some lines for further research.

2. Theoretical background

2.1. The acquisition of agreement morphology and obligatory subjects

Omission of inflectional morphology, as in (1a) above, is not exclusive to L2 English learners. In the early grammar of children acquiring English as an L1 there is also a stage at which they omit agreement morphemes (Brown, 1973; Rizzi, 1993;

see also Ionin & Wexler, 2002 for a review). Although English falls into the category of languages which require an overt grammatical or lexical subject or, in Generativist terms, it is a 'non-pro-drop' language (Chomsky, 1981), this early grammar is also characterised by the production of null referential and expletive subjects, as in (3a) and (3b) above, respectively (Hyams, 1989). Some authors maintain that the L1 learners' consistent production of finiteness markers on verbal forms has been found to correlate with the consistent provision of overt subjects (Guilfoyle, 1984), and some others note that for that developmental stage it is essential to have acquired expletive subjects first, as it is the child's awareness of the need to provide these purely grammatical subjects that facilitates their systematic production and, subsequently, that of referential subjects (Hyams, 1989; see also Ruiz de Zarobe, 1997, 1998, 2000). As regards L1 English inflectional morphology, little difficulties have been claimed, except for its possible omission, but no evidence has been found for the incorrect production of agreement morphology errors (i.e. 'commission errors') as in (1b) above (see Brown, 1973; Ionin & Wexler, 2002; Zobl & Liceras, 1994). Finally, unlike in L2 English, research has shown that in the acquisition of L1 English both suppletive and affixal finiteness morphemes "cluster close together in development" (Ionin & Wexler, 2002, p. 102; Zobl & Liceras, 1994).

In the acquisition of L2 English, which is the focus of this paper, it has been widely attested that adult learners of English with 'pro-drop' L1s (i.e. with an L1 which accepts the omission of the subject), such as Spanish, transfer certain aspects of the null subject parameter in their L1 to their L2 grammar (White, 1986, 1989, 2003b). Subject omission – as in (4) and (5) – is to be expected particularly at earlier stages of L2 English acquisition (see Phinney, 1987 on adult L2 grammar; cf. Haznedar, 2001; Ionin & Wexler, 2002 on child L2 English learners and White, 2003a on one adult learner), though expletive subjects (4) have been found to be problematic even for advanced adult L2 learners (Ruiz de Zarobe, 1998).

(4) *in the city Ø is cloudy and sunny [non-CLIL: subject 162]

(5) *because Ø is in city in the morning [non-CLIL: subject 169]

Acquiring the obligatory use of overt subjects in L2 English entails certain difficulty for learners with pro-drop L1s, as there is a greater cost involved in resetting the 'unmarked' (i.e. pro-drop) value of the null subject parameter in languages like Spanish to the 'marked' non-pro-drop English parameter (Phinney, 1987). An additional factor which complicates the resetting of this parameter and the Spanish learners' acquisition of obligatory subjects in L2 English is the extensive use of purely grammatical and semantically empty expletive subjects in English, unlike in Spanish, which do not have to meet the interpretative identification requirement criteria of null

subjects (6) (e.g. on adult learners, see Judy, 2011; Judy & Rothman, 2010; Phinney, 1987; Ruiz de Zarobe, 1997, 1998, 2000). Unlike in L1 English, where it is expletive subjects that have a leading role in the readjustment of the pro-drop parameter, the acquisition of expletive subjects by Spanish learners of English is quite late. In fact, with adult learners it has been observed to be delayed until referential subjects and auxiliary and modal verbs are acquired, with the acquisition of the progressive auxiliary being the actual trigger of the readjustment of the pro-drop parameter in L2 English (Ruiz de Zarobe, 1998).

(6) It is cloudy and sunny

*ello está nublado y soleado

As concerns the acquisition of English inflectional morphology, L2 English learners have been found to have great difficulties with inflectional morphemes, especially in spoken production (Ionin, 2013). This is, in fact, a widespread and frequent error in the acquisition of English as a foreign language irrespectively of the learners' L1 (see García Mayo & Villarreal Olaizola, 2011). Previous research has shown that adult non-native speakers of English tend to be quite inconsistent in their production of English verbal inflection and often resort to uninflected forms as the default option (García Mayo & Villarreal Olaizola, 2011; Ionin, 2013), as illustrated by *protect* and *look* in (7) and (8). Variability in the production of agreement morphology is claimed to persist even if a high level of proficiency in the target language is achieved (Lardiere, 2000).

(7) *he **protect** the dog for the rain [non-CLIL: subject 162]

(8) *the person **look** at the dog [CLIL II: subject 75]

The asymmetry between the acquisition of suppletive and affixal morphology is another characteristic of the L2 learners' grammar that differs from the acquisition of L1 English and has attracted a great deal of the scholarly attention (García Mayo & Villarreal Olaizola, 2011; Ionin & Wexler, 2002; Villarreal Olaizola, 2011; Villarreal Olaizola & García Mayo, 2009; Zobl & Liceras, 1994). These studies have observed how suppletive inflection is provided with a greater frequency than affixal morphology, with an earlier emergence and mastery of copula *be* compared to auxiliary *be*.

From a very broad perspective, L2 learners' agreement morphology errors and null subjects can be explained under the lens of two main generativist proposals, both of them with a range of specific variants (see Ionin, 2013; Slabakova, 2008 and White, 2003b for a review). On the one hand, there are scholars who claim that errors result from a global or more local representational impairment in functional categories and

feature values (e.g. Goad et al., 2003; Hawkins & Casillas, 2008; Hawkins & Chan, 1997; Tsimpli & Dimitrakopoulo, 2007). Under this view, "learners are incapable of acquiring new features in the L2 that are not present in their L1" (Ionin, 2013: 507). Age seems to play a decisive role in this account, as it predicts that formal features that are unspecified in the L1 will not be accessible to adult learners and this will inevitably derive in syntactic deficits in the L2 (García Mayo & Villarreal Olaizola, 2011).

On the other hand, there are scholars who contend that the learners' L2 is not impaired and attribute variability and errors to different factors such as a mapping problem between abstract features and their corresponding morphological form or problems with the specifications and selection/reassembly of those features (e.g. Haznedar & Schwartz, 1997; Ionin & Wexler, 2002; Lardiere, 2008, 2009; Prévost & White, 2000; for a review of these approaches, see Villarreal Olaizola, 2011). Evidence for this stand comes, for instance, from the fact that both child and adult L2 English learners have proved to have the agreement category in their grammar: despite having difficulties supplying the correct morphological form, particularly in the case of affixal inflection, they do tend not to misuse verbal inflection unsystematically (García Mayo & Villarreal Olaizola, 2011; Ionin, 2013; Villarreal Olaizola, 2011). If their L2 grammar suffered from a representational impairment at the level of syntactic representation, commission errors, as in (1b) above, would be higher (Ionin & Wexler, 2002). Research on subject features also confirms that even adult L2 grammars are not impaired: although the L2 features that are not present in their L1 do not seem to be fully acquired, learners do show learning development with increased exposure when confronted with grammaticality judgments by rejecting null expletive subjects (as in (4)) to a larger extent (Pladevall Ballester, 2013).

2.2. CLIL in Spain and the acquisition of specific morphosyntactic features

As already mentioned, one approach that has been advocated and implemented in the Spanish context to improve L2 English learners' proficiency in the target language involves additional exposure to English through so-called 'CLIL programmes' (Marsh, 2002). Content and Language Integrated Learning or CLIL, coined in 1994 (Marsh, 1994), is an umbrella term that encompasses those approaches "in which a second language (a foreign, regional or minority language and/or another official state language) is used to teach certain subjects in the curriculum other than language lessons themselves" (Eurydice, 2006: 8). Accordingly, in CLIL approaches, while there may be support for the L1 and the classroom culture is that of the L1, it is mainly the target language that is used as a medium of instruction. The learners' knowledge of the target language is most commonly limited but teachers are assumed to be sufficiently competent, as the L2 curriculum has to parallel that of the L1 (Martínez Adrián, 2011).

CLIL is thus a dual-focused approach where the L2 is integrated in the curriculum to teach content classes while content is on some occasions integrated in the language classes (Martínez Adrián, 2011; Ruiz de Zarobe, 2011).

There is no doubt that CLIL has postulated itself as a very convenient, effective method to compensate for the frequently low number of hours of instruction in the target language in countries like Spain, where the popularity of CLIL has increased in the last decades to the extent that it is considered one of the European leaders in both its implementation and its research (Coyle, 2010; Martínez Adrián, 2011; Pérez-Cañado, 2012). In the case of CLIL in Spain, its implementation is characterised by, on the one hand, the wide range of CLIL programmes, which vary depending on the autonomous region, and, on the other, the integration of the target language in monolingual or bilingual communities (Galicia, Basque Country, Catalonia, Valencia and the Balearic Islands) (Martínez Adrián, 2011; Pérez-Cañado, 2012). The impact of CLIL on Spanish bilingual regions has attracted the attention of many scholars and a great bulk of the research on the topic (e.g. for the Basque Country, see e.g. García Mayo & Villarreal Olaizola, 2011; Gutiérrez-Mangado & Martínez-Adrián, 2018; Lázaro Ibarrola, 2012; Martínez Adrián & Gutiérrez Mangado, 2009, 2015a, 2015b; Ruiz de Zarobe & Lasagabaster, 2010 and Villarreal Olaizola, 2011; for Galicia, San Isidro, 2010; San Isidro & Lasagabaster, 2019a, 2019b, and for the Catalan territories, Aguilar & Muñoz, 2014; Juan-Garau & Pérez-Vidal, 2011; Navés & Victori, 2010; Pérez-Vidal, 2007, Pérez-Vidal & Juan-Garau, 2010, 2011; and Pérez-Vidal & Roquet, 2015). In contrast, monolingual communities such as the region explored in this investigation, Cantabria, have received less attention in the literature (e.g. Gutiérrez Martínez & Ruiz de Zarobe, 2017; Merino & Lasagabaster, 2018; for further references, see Fernández Fontecha, 2009 and Pérez-Cañado, 2012). In addition to this, another reason to explore the production of learners from monolingual regions concerns their disadvantageous position with respect to the greater metalinguistic awareness of L3 English learners from bilingual communities (Jessner, 2014).

Overall, research on CLIL contexts in Spain has demonstrated that the greater exposure to the target language that CLIL grants typically benefits the learners' overall proficiency in English. CLIL learners have been found to perform more target-like than non-CLIL groups in oral and written fluency, syntactic complexity, reading comprehension and receptive vocabulary, for instance (for Primary Education, see Jiménez Catalán et al., 2006 and Navés & Victori, 2010; for Secondary Education, Gutiérrez-Mangado & Martínez-Adrián, 2018; Lasagabaster, 2008; Ruiz de Zarobe, 2008). Nonetheless, for the acquisition of more specific areas of the target language the evidence obtained is less conclusive and the potential positive impact of CLIL more dubious. A case in point is the syntax-morphology interface. Some studies do attest an advantage in CLIL groups (compared to learners with only EFL instruction) as regards

the acquisition of certain morphosyntactic aspects such as irregular past forms (Lázaro Ibarrola, 2012), syntactic complexity and article use (Gutiérrez-Mangado & Martínez-Adrián, 2018), affixal compared to suppletive inflection (Villarreal Olaizola & García Mayo, 2009) and placeholders (Martínez Adrián & Gutiérrez Mangado, 2009). Despite these results, all based on data from teenage L2 learners, most of the research carried out hitherto has observed minimal differences between age-matched CLIL and non-CLIL groups in their rate of agreement morphology errors and subject omission (García Mayo & Villarreal Olaizola, 2011; Martínez Adrián & Gutiérrez Mangado, 2009, 2015a; Villarreal Olaizola, 2011). Some of these investigations, however, do find remarkable differences in the production of the different inflectional forms and subjects, as is commented in detail below.

Verb inflection is claimed to be the 'bottleneck' of L2 acquisition (see Slabakova, 2008). Omission of suppletive forms and, particularly, affixal verb morphology is a very frequent phenomenon in both child and adult L2 English grammars (Ionin, 2013). As commented above, affixal inflection (9) is not only highly variable but also more frequently omitted than suppletive verbal forms (10) because the latter are acquired at earlier stages of the acquisition of L2 English. This has been attested in both CLIL and non-CLIL groups at Secondary School (García Mayo & Villarreal Olaizola, 2011; cf. Villarreal Olaizola & García Mayo, 2009 on the more target-like performance of CLIL learners and Villarreal Olaizola, 2011 on the disappearance of that advantage a year after the CLIL programme is over). Non-CLIL learners, however, have been found to have greater problems with the production of not only affixal inflection but also auxiliary *be* compared to copula *be* (Villarreal Olaizola, 2011). This well-known dissociation between the acquisition of affixal and suppletive inflection seems to be rooted in the different process whereby auxiliary and copula *be* check their tense and agreement features: whereas lexical verbs features are checked covertly and, thus, their inflection "may or may not be expressed morphologically, depending on language-specific rules", the feature checking of *to be* is overt and is expressed morphologically, which entails fewer complications for L2 learners (see García Mayo & Villarreal Olaizola, 2011, pp. 132-134 for further discussion and references).

(9) *that the boy eeeh **take** the dog in his house [CLIL II: subject 77]

(10) *the dog eeeh **Ø** hungry [CLIL I: subject 39]

In contrast, commission errors such as (11) and (12) are most often negligible. The few studies carried out on CLIL and non-CLIL teenage learners of L2 English attest a very low incidence of this type of errors in both groups, thus claiming that they are not representative of these learners' interlanguage (García Mayo & Villarreal Olaizola, 2011; Villarreal Olaizola, 2011; Villarreal Olaizola & García Mayo, 2009).

(11) *her mother and her father eeeh **sees** happiness [CLIL II: subject 77]

(12) *in the book **are** there one dog in the town [CLIL I: subject 43]

The use of placeholders, that is, the use of a suppletive form – often *is* (13) or *he* (14) – before the (bare) main verb to hold its inflection is another characteristic of Spanish L2 English learners' interlanguage, particularly in their earlier stages of acquisition (also in L1 English, see Lázaro Ibarrola, 2002, 2012). This mechanism, which is assumed to result from L1 transfer of functional categories (see Martínez Adrián & Gutiérrez Mangado, 2009), is attested in the literature mainly in non-CLIL groups as a sign of their less advanced stage of acquisition of English (see Lázaro Ibarrola, 2012 and Martínez Adrián & Gutiérrez Mangado, 2009 for data from Secondary Education; for data from non-CLIL contexts, see García Mayo et al., 2005 and Lázaro Ibarrola, 2002).

(13) *the dog **is** walk in the city [non-CLIL: subject 179]

(14) *when the the child **he**'s coming [CLIL I: subject 41]

Finally, subject omission is expected to be found in Spanish learners of L2 English given the pro-drop nature of the Spanish language. In general terms, no significant differences between CLIL and (age-matched as well as older) non-CLIL teenage learners have been attested, although the latter have been found to produce more null subjects (Martínez Adrián & Gutiérrez Mangado, 2009, 2015a). Research in EFL contexts has observed in Spanish adult oral and written production that null subjects are more frequent in earlier stages of acquisition, with their omission and acceptance decreasing with increasing proficiency in the L2 (see Ortega Durán, 2016; Ruiz de Zarobe, 1997, 1998, 2000). Interestingly, some of the studies carried out in non-CLIL contexts have found that not all subjects are acquired at the same time. The evidence from adult L2 English written production provided in Judy (2011), Judy and Rothman (2010), Phinney (1987) and Ruiz de Zarobe (1997, 1998, 2000), for instance, points to an earlier acquisition of obligatory referential subjects (15), compared to obligatory expletive subjects (16). Based on these results, a high rate of null subjects and variability with the provision of the different types of subjects can be expected in the L2 grammar of young CLIL learners (see García Mayo, 2003 on the problems of 11- and 12-year-olds to identify sentences with null subjects as ungrammatical after approximately four years of exposure to English in an EFL setting).

(15) *because *no eh doesn't doesn't like raining* **Ø** doesn't like raining [non-CLIL: subject 169]

(16) *and *de repent* **Ø** is *ra* raining [non-CLIL: subject 185]

3. Aims and research questions

This study is not intended to corroborate or refute the theoretical approaches succinctly reviewed above but to delve into the production of agreement morphology errors and null subjects from a usage-based perspective. To this end, we will analyse the oral production of L2 English learners from three Primary Schools in Cantabria, two of them implementing CLIL programmes and the third one providing just traditional EFL instruction, with the aim of shedding more light into the potential impact of an early and long-term CLIL exposure, compared to an early traditional EFL instruction, on L2 English morphosyntax. In EFL contexts, an earlier start has been claimed not to be always an advantage (Cadierno et al., 2020; Muñoz, 2006), especially if the amount of exposure is not increased or used effectively (García Mayo, 2003; Muñoz, 2002). Still, Pladevall Ballester (2012) observes that Spanish 5 year-olds attending an immersion school for two years were sensitive to grammaticality judgments, and García Mayo (2003) with older subjects (11 to 17 year-olds), that an increased exposure in EFL instruction results in a more target-like performance. A positive impact of longer-term exposure has also been attested in CLIL contexts, with CLIL learners performing better than non-CLIL groups (e.g. Lasagabaster, 2008; and, for Primary Education, Jiménez Catalán et al., 2006) and older EFL groups (Navés & Victori, 2010; Ruiz de Zarobe, 2008) in aspects other than morphosyntax. The studies reviewed in Section 2.2, especially those investigating agreement morphology, were mostly conducted in Secondary Education (with 13 to 18 year-olds) in Spanish bilingual communities (mainly the Basque Country), and compared the oral production of CLIL and (age-matched or older) non-CLIL learners, with the former receiving an additional CLIL exposure to the target language of a maximum of three years. Their results suggest that the amount of exposure to the L2 may not be as relevant a factor as age when explaining morphological development: higher accuracy has been found in older learners, as the acquisition of the morphological system has been claimed to speed up at the age of 12-13 and to accelerate with the acquisition of the pronominal system at the age of 15 (see Lázaro Ibarrola, 2012; Lázaro & García Mayo, 2012; Martínez Adrián & Gutiérrez Mangado, 2015a; Villarreal Olaizola, 2011). Prior research has thus underexplored the potential impact of an earlier and longer-term CLIL exposure to the target language on the realisation of specific morphosyntactic aspects of younger Spanish learners of L2 English from monolingual communities, a gap that this investigation comes to bridge.

In particular, this paper aims at exploring agreement morphology errors and subject omission in Primary Education Grade 6 English learners from Cantabria with a view to further assess the potential benefit of additional CLIL exposure on L2 English morphosyntax in the longer run. Our investigation contributes to previous research by surveying not only two underresearched areas, 11- and 12-year-old learners

of English and a monolingual community in northern Spain, but also a group which has received a considerably large additional exposure to the target language through a CLIL programme of approximately six years. The main objective is thus to gauge the extent to which an earlier and longer-term additional CLIL exposure translates into a potentially greater advantage in the provision of specific morphosyntactic aspects in comparison to (age-matched) non-CLIL learners who have also been exposed to the L2 for six years but have received traditional English lessons only. In the second place, our research also aims at examining potential differences in the production of agreement morphology errors and null subjects within each of the two groups surveyed.

Based on prior research and findings on the acquisition of verb inflection and overt subjects in L2 English, this investigation seeks to answer the following research questions:

1. Are there any differences between CLIL and non-CLIL learners in the production of agreement morphology errors and null subjects?

2. Are there any intragroup differences in CLIL and non-CLIL groups as regards

   i. the omission of inflection compared to commission errors?

   ii. the omission of affixal inflection compared to the omission of suppletive inflection?

   iii. commission errors in affixal verbal morphology compared to commission errors in suppletive inflection?

   iv. null referential subjects compared to null expletive subjects?

4. Methodology

4.1. Participants

The participants in this study were sixty-three Primary Education (Grade 6) students learning L2 English in three schools in the Spanish monolingual region of Cantabria. They constitute a subsample of all the learners from *The Primary Education Learners' English Corpus* (PELEC) (see Blanco-Suárez et al., 2020). As Table 1 shows, the students in the sample are divided into two age-matched (11- and 12-year-old) groups: the CLIL group (*n*=28) and the non-CLIL group (*n*=35). The English onset age for

both the CLIL learners and the non-CLIL learners was around 5 and 6 years old. In the three Primary Schools, learners receive an estimated average of around 2.5 and 3.5 hours of English instruction per week, which amounts to a mean of 617.5 hours of English as a Foreign Language (EFL) lessons at the time the data were collected. Additionally, the students enrolled in CLIL programmes are taught several content classes – Arts and Crafts, Music, Natural Sciences, and Physical Education to be more precise – in the target language at an average of two or three hours a week. As a result, they had received a mean of 488 hours of additional exposure to English since Grade 1 until the time they participated in the study.

Table 1: Description of the sample

	CLIL	**non-CLIL**
Participants	28	35
Age at testing	11–12	11–12
English onset age	5–6	5–6
CLIL onset age	5-6	none
Hours of EFL per week	2.5–3.5	2.5–3.5
Mean hours of EFL at testing	617.5	617.5
Hours of CLIL per week	2–3	none
Mean hours of CLIL at testing	488	none

As for the participants' target language competence, both groups differ slightly in their level of English. Based on previous standardised tests that the Education Authorities administered at the end of Primary Education in the Autonomous Community of Cantabria, learners in CLIL schools were expected to reach, at best, the A2 level of the Common European Framework of Reference for Languages (CEFR) whereas schoolchildren in non-CLIL schools were examined for the A1 level. Learners in this study were administered some language tests (drawing on materials from the Cambridge English A1 Movers and A2 Flyers tests) prior to the collection of the data as part of a larger project that involves a greater number of schools and participants (see Blanco-Suárez et al., 2020 for further information). As for the subsample analysed in the present study, these tests revealed slightly superior mean scores of the CLIL group over their non-CLIL homologous in the listening comprehension, reading comprehension and, particularly, the Use of English modules.

4.2. Instruments and procedures

This investigation draws on data from *The Primary Education Learners' English Corpus* (PELEC) (Blanco-Suárez et al., 2020). PELEC was compiled as part of a larger project which involved collecting learner data by means of several instruments: from foreign language motivation, communication strategies and background questionnaires to various written and (interactive) oral tasks (writing a letter, telling a story, spot-the-differences) as well as listening comprehension, reading comprehension and Use of English (cloze) tests. The data reported in this paper focus on the oral production of the participants resulting from the oral task consisting of telling a story. In that task, the participants were asked to narrate individually a story based on the 8-vignette story illustrated in the Appendix. The task was recorded in the Primary Schools by one or two bilingual researchers, who, when necessary, guided the participants and answered their questions in English only. The participants' oral production was orthographically transcribed and analysed for the production of agreement morphology and obligatory subjects, as explained in detail below. The data were transcribed by two different researchers who had been previously trained to follow the CHAT conventions for the CLAN subprogramme in the Child Language Date Exchange System (CHILDES, MacWhinney, 2000). Two codifiers revised the transcriptions and identified the errors, showing a high degree of agreement. Any controversies or ambiguous cases were solved jointly on a case-by-case basis.

In keeping with Ionin and Wexler (2002), we considered as analysable utterances all those contexts where, in the case of agreement morphology, there was a finite or non-finite verb, where inflection was realised by either (i) a missing, correct or incorrect affix or (ii) a missing or overt suppletive form of the auxiliary or copula *be*. Regarding subject contexts, we analysed all the utterances where either a referential or an expletive subject was omitted. We excluded from the analysis instances involving:

(i) regular and irregular past tense lexical verbs: *the dog eeeh he see (2'') eeh he saw is raining* [CLIL I: subject 41]

(ii) autocorrection: *the boy eeh goes, no, go at your home* [non-CLIL: subject 179]

(iii) the repetition of the same verbal form: *his mom is cooking with (1'') is cooking (2'') is cooking.* [CLIL I: subject 34]

(iv) an ambiguous referent: *and the boy eeeh (3'') eeeh at look the dog and dog eeeh (2'') eeeh it's ha bueno is happy. and (6'') eeeh and go to (1'') to house*. [CLIL I: subject 39]

(v) an unclear form: *the dog names [name's?] eeh Tim is go going to the city* [non-CLIL: subject 169]

(vi) an unfinished utterance at the verb phrase level: *the boy and the dog (1'') going to (2'') to (5'') eeeh going to (12'') xxx eeeh going to (9'')* joer ahora no me sale nada [CLIL I: subject 39]

The type of utterances that we analysed in this investigation are exemplified in (17) to (21) below. In terms of verbal inflection, we considered instances which involve either the omission (17) or the commission (18) of the verbal morphology. More specifically, we analysed the omission and commission of both affixal (19) and suppletive inflection (20) in the present tense. In the case of suppletive forms, omission refers to the elision of *be*, as no uninflected forms of the auxiliary or the copula were attested in the sample surveyed. Finally, the use of placeholder *is* (21) as a resource to hold inflection was also investigated (see García Mayo et al., 2005; Lázaro Ibarrola, 2002, 2012; Martínez Adrián & Gutiérrez Mangado, 2009). The use of placeholders in the subject, as in *the wolf* ***he*** *opened the door*, is not discussed here (see García Mayo et al., 2005), owing to the fact that only two tokens were attested in the sample, one in each of the research groups. For this reason, only placeholder *is* is considered.

(17) *the boy **sleep** with the dog [CLIL I: subject 46]

(18) *the dog and the boy **goes** to the bed to sleep [non-CLIL: subject 194]

(19) *and the dog **eat** [non-CLIL: subject 191]

(20) *the dad **Ø** running to the bathroom [CLIL I: subject 50]

(21) *one man **is walk** for the street [non-CLIL: subject 170]

Concerning subject contexts, we explored not only subject omission in general but also the omission of referential (22) compared to expletive (23) subjects.

(22) *because **Ø** is in city [non-CLIL: subject 169]

(23) *because **Ø** is raining [non-CLIL: subject 186]

The results from the quantitative analysis, which are reported in the next section, were subjected to statistical testing with SPSS.22. Mean scores (expressed as the mean percentage of errors derived from the relative number of errors made by every child over the number of potential contexts for those errors to happen) and standard deviations were calculated for the different morphosyntactic features and learner

groups. Total number of contexts and mean number of contexts per learner are also reported. Kolmogorov-Smirnov tests were computed to explore the normality of the distribution of the samples. As the distribution of the samples was not normal, non-parametric procedures were used, namely Mann-Whitney U tests for the intergroup comparisons and Wilcoxon tests for the intragroup ones. With regard to statistical significance, alpha levels of .05 (*), .01 (**) and .001 (***) were used.

5. Results and discussion

This section is organised into three different parts. Firstly, to answer our first research question, we compare the rate of omission of inflection, commission errors and null subjects produced by CLIL learners with the results obtained in the non-CLIL group. Secondly, we explore intragroup differences with respect to the incidence of omission of inflection and commission errors as well as null subjects. In the final section, we provide an overall discussion of the results from both the inter- and the intragroup contrasts.

5.1. Intergroup comparisons

In this section, the results for affixal and suppletive inflection (both auxiliary *be* and copula *be*) are presented first, followed by those for subject contexts.

5.1.1. Affixal and suppletive inflection

Starting with affixal inflection, Table 2 reports the total number of contexts and the mean number of affixal contexts per learner, the total number of errors of each affixal type, the mean percent and the standard deviation (in brackets) of omission and commission errors as well as the use of placeholders by both CLIL and non-CLIL learners. Omission of inflection is by far the most common type of error in both groups, with CLIL learners omitting affixal inflection at a slightly higher rate (62.81%) than their non-CLIL counterparts (53.63%). Commission errors, in contrast, are very rare in general and more frequently produced by the non-CLIL group (1.78% vs CLIL: 0.22%). These results, however, can only be taken as tendencies, as the intergroup differences in omission and commission did not reach statistical significance.

The use of placeholder *is* is also reported in Table 2. Both CLIL and non-CLIL groups made use of it, particularly the latter, which proved to overly realise inflection via a placeholder twice as frequently (19.52%) as the CLIL group (9.79%). Despite the apparent difference in the frequency in use of placeholders, the statistical test found no statistical support.

Table 2: Absolute values, mean percent and standard deviation for omission and commission of affixal inflection and use of placeholders

Affixal inflection contexts	**CLIL**		**non-CLIL**	
Total number of contexts	179		180	
Mean number of contexts per learner	6.39		5.14	
	Total	Mean % (SD)	Total	Mean % (SD)
Omission **the boy **sleep** with the dog*	110	62.81% (31.69)	106	53.63% (32.52)
Commission **they **goes** to the bed*	1	0.22% (1.18)	5	1.78% (4.36)
Placeholders **the boy **is** sleep with the dog*	13	9.79% (21.03)	30	19.52% (30.40)

Moving on to suppletive inflection, Table 3 reports the omission and the use of an incorrect form of auxiliary *be*. The data show that the production of a wrong form of the auxiliary is very unlikely and that it is only CLIL learners that produce one error of this type (1.59%), which naturally renders the intergroup contrast non-significant. Likewise, the difference in the omission rate of the auxiliary between CLIL and non-CLIL learners did not reach statistical significance, although the latter were found to omit it at a greater extent (38.39%) than the former (23.38%).

Table 3: Absolute values, mean percent and standard deviation for omission and commission of auxiliary be

Auxiliary *be* contexts	**CLIL**		**non-CLIL**	
Total number of contexts	57		69	
Mean number of contexts per learner	2.04		1.97	
	Total	Mean % (SD)	Total	Mean % (SD)
Omission **boy **sleeping** the dog*	9	23.38% (39.08)	26	38.39% (42.99)
Commission **the dog **are** coming on the city*	1	1.59% (7.27)	0	0.00% (0.00)

As in the previous two contexts, commission errors of copula *be* are remarkably rare in comparison to omission of inflection. Still, commission errors of the copula were found to be slightly more common than those of auxiliary *be*. Table 4 shows that CLIL learners are less target-like in their production of copula *be* than their non-CLIL homologous, as the former produce both more errors of commission (10.91%) and a higher rate of omission (14.06%) of the copula than non-CLIL learners (commission: 3.33%; omission: 1.83%). Despite these observations, no support from inferential statistics was found for this intergroup divergence.

Table 4: Absolute values, mean percent and standard deviation for omission and commission of copula be

Copula *be* contexts	**CLIL**		**non-CLIL**	
Total number of contexts	75		52	
Mean number of contexts per learner	2.68		1.49	
	Total	Mean % (SD)	Total	Mean % (SD)
Omission **the dog **Ø** with a bedroom*	7	14.06% (30.05)	2	1.83% (5.67)
Commission **your eyes **is** brown*	8	10.91% (23.92)	2	3.33% (11.60)

On the whole, the data presented in this section reveal minimal differences between CLIL and non-CLIL learners. Both groups omit and incorrectly provide affixal and suppletive inflection at similar rates and the problems of CLIL learners to provide copula *be* correctly were also found not to differ significantly from the commission rate of their non-CLIL homologous. In general terms, these non-significant differences between the performance of the CLIL group and that of its (age-matched) non-CLIL counterpart are in keeping with most of the research on older (i.e. teenage) L2 learners of English (see, for instance, Martínez Adrián & Gutiérrez Mangado, 2009, 2015a; cf. Villarreal Olaizola, 2011 and Villarreal Olaizola & García Mayo, 2009). One trend that somehow deviates from previous research is the CLIL group's use of placeholder *is*. Given the younger age of the participants in this study compared to prior investigations, this finding can be taken as an indication of their still early stage of acquisition of the L2 compared to older CLIL groups in previous studies (see Lázaro Ibarrola, 2012; Martínez Adrián & Gutiérrez Mangado, 2009).

5.1.2. Subject contexts

The learners' rate of subject omission is presented in Table 5. Both the CLIL group and the non-CLIL group produce null subjects at a very similar rate, with the former's performance being slightly less target-like (10.40%) than the latter's (9.65%) in the overall rate of omission. By inspecting the data in detail, it can be observed that this difference, which in any case is not statistically significant, stems from the different performance of the groups with respect to the two possible types of subjects. Thus, whereas non-CLIL learners have greater difficulties to provide expletive subjects (63.46% vs CLIL: 56.52%), the rate of omission of referential subjects is slightly higher in the case of the CLIL group (3.53% vs non-CLIL 2.41%). As in the previous contexts under analysis, the differences between CLIL and non-CLIL learners did not reach statistical significance.

Table 5: Absolute values, mean percent and standard deviation for subject omission

Subject contexts	**CLIL**		**non-CLIL**	
Total number of subject (expletive/referential) contexts	311 (41/270)		301 (38/263)	
Mean number of subject (expletive/referential) contexts per learner	11.11 (1.46/9.64)		8.6 (1.09/7.51)	
	Total	Mean % (SD)	Total	Mean % (SD)
Overall subject omission	33	10.40% (10.14)	30	9.65% (9.82)
Omission expletive subjects **because Ø is raining*	22	56.52% (42.66)	23	63.46% (43.97)
Omission referential subjects **because Ø is in the city*	12	3.53% (6.61)	7	2.41% (6.94)

The overall moderately low rate of null subjects in both groups seems to point to an apparent mastery of obligatory subjects. This could be taken in principle as a sign of a more advanced acquisition stage, not being too far away from the low rate of null subjects produced by older learners in previous studies (Martínez Adrián & Gutiérrez Mangado, 2009, 2015a).

5.2. Intragroup comparisons

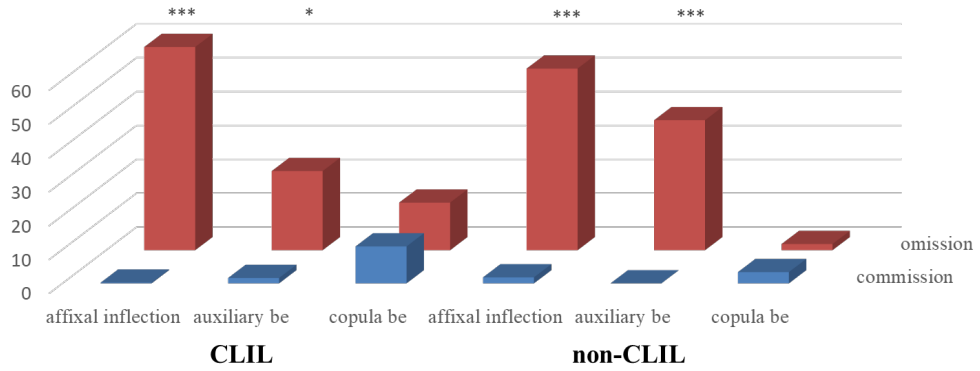
To answer our second research question, this section scrutinises the data in more detail by exploring the types of features under analysis within each of the two groups.

5.2.1. Omission of inflection vs commission errors

Figure 1 illustrates the intragroup differences between the mean percent of omission and commission errors in both groups. The data confirm that the incidence of commission errors is negligible in comparison to omission of inflection, as discussed above. With both auxiliary *be* and affixal inflection, the rate of omission is exceedingly greater (auxiliary *be*: CLIL 23.38%, non-CLIL 38.39%; affixal: CLIL 62.81%, non-CLIL 53.63%) than that of commission (auxiliary *be*: CLIL 1.59%, non-CLIL 0.00%;

affixal: CLIL 0.22%, non-CLIL 1.78%). Accordingly, the difference between these two trends is highly significant in both the CLIL (affixal inflection: z=-4.465, p<.001***; auxiliary *be*: z=-2.410, p=.016*) and the non-CLIL group (affixal inflection: z=-4.629, p<.001***; auxiliary *be*: z=-3.351, p=.001***).

Figure 1: Mean percent and Wilcoxon test results for omission of inflection vs commission errors



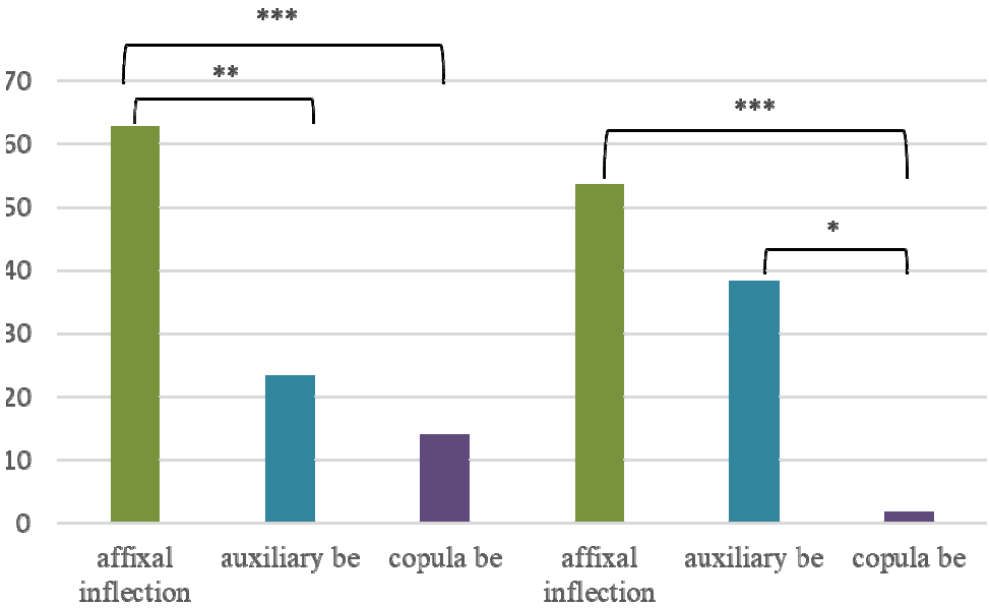
Interestingly, both groups show greater problems to supply the correct form of copula *be* than the rest of the forms investigated. The rate of commission errors with this suppletive form (3.33%) is in fact more frequent than that of omission of inflection (1.83%) in the non-CLIL group (in line with results from older L2 learners in Villarreal Olaizola & García Mayo, 2009), but this intragroup contrast found no statistical support. The contrast is also non-significant in the case of the CLIL group, although in this case CLIL learners were found to be more likely to omit the copula (14.06%) than produce an incorrect form (10.91%). All the same, the commission rate of copula *be* in this group is considerably higher than that of affixal inflection and auxiliary *be*.

The fact that the overall marginal commission error rate, particularly in both affixal inflection and auxiliary *be*, contrasts significantly with the higher frequency of omission of inflection in the L2 English learners in this study is in keeping with the main trends attested in prior literature for older learners (García Mayo & Villarreal Olaizola, 2011; Villarreal Olaizola & García Mayo, 2009; Villarreal Olaizola, 2011). The evident difference in frequency between affixal and suppletive inflection observed in Figure 1 is discussed in more detail in the next section.

5.2.2. Omission of inflection

Figure 2 elaborates on the previous results by focusing only on the omission rate in each of the three contexts analysed in both groups. The results for the CLIL group confirm the high frequency of omission of affixal inflection (62.81%) compared to that of suppletive forms: the contrast between the former and copula *be* (14.06%) was found to be highly significant (z=-3.672, *p*<.001***), while that between affixal inflection and auxiliary *be* (23.38%) is only slightly significant (z=-2.962, *p*=.003**). The evidence of the non-CLIL group reveals a very similar pattern, with the omission of copula *be* (1.83%) being significantly lower than the omission rate of affixal inflection (53.63%) (z=-3.705, *p*<.001***). Auxiliary *be* omission (38.39%) is significantly higher than that of copula *be* (z=-2.555, *p*=.011*) but not significantly different from affixal omission. CLIL learners, in contrast, were not found to omit auxiliary *be* and copula *be* at significantly different rates.

Figure 2: Mean percent and Wilcoxon test results for omission of inflection



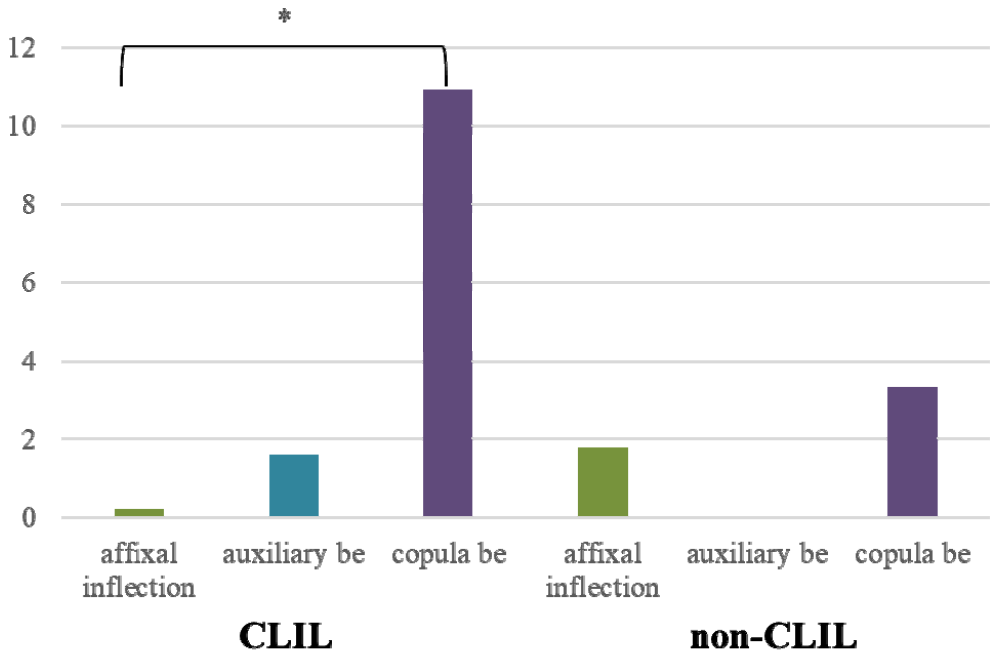
The results clearly indicate that affixal inflection is by far the most problematic context of the three: the omission of affixal inflection is the most common error made by both CLIL and non-CLIL groups, followed by the omission of the auxiliary *be* and, less commonly, the elision of copula *be* (in line with data from older learners in García Mayo & Villarreal Olaizola, 2011; Villarreal Olaizola, 2011; Villarreal Olaizola

& García Mayo, 2009). In the case of the CLIL group, despite the high omission of affixal inflection in lexical verbs, the significantly lower rate of omission of suppletive inflection points to a considerably better mastery of both copula and auxiliary *be*. The non-CLIL group, in contrast, still shows greater difficulties with auxiliary *be*, which they very frequently omit.

5.2.3. Commission errors

Commission errors are extremely rare in the sample investigated, as Figure 3 further illustrates. Even so, in the CLIL group, unlike with the omission trends presented above, it is copula *be* that is produced incorrectly at a significantly higher rate (10.91%) than affixal inflection (0.22%) (z=-2.207, p=.027*). Auxiliary *be* was found to be produced erroneously more often (1.59%) than the inflection in lexical verbs but remarkably less frequently than the inflection of copula *be*. None of these differences found statistical support, though. In the case of the non-CLIL group, although the commission error rate of the copula (3.33%) was higher than that of affixal inflection (1.78%), the difference is not statistically significant either. In this group, unlike in its CLIL homologous, commission errors of the auxiliary *be* were not attested.

Figure 3: Mean percent and Wilcoxon test results for commission errors

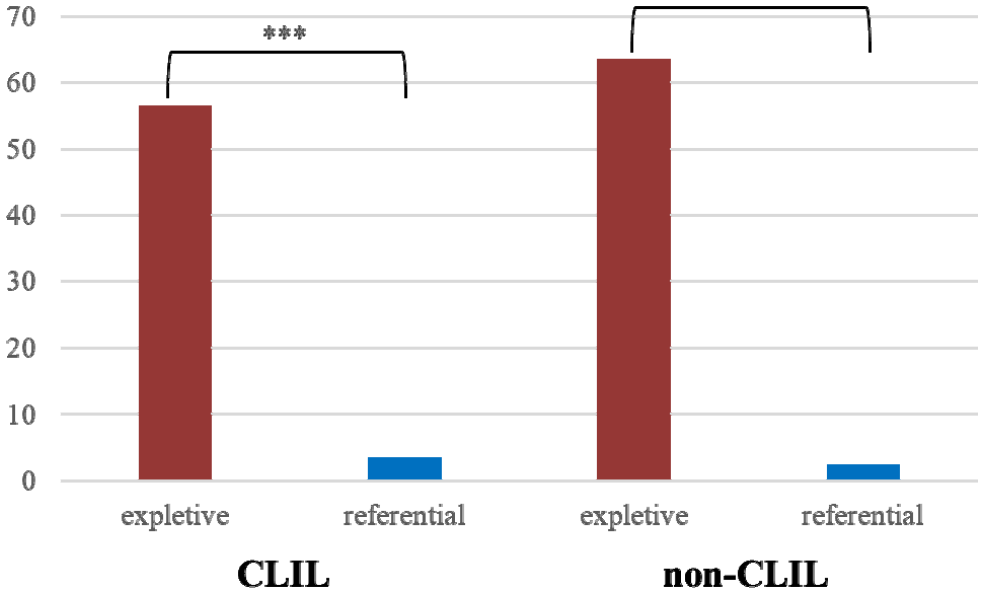


The data obtained in this study comes to confirm prior results and claims with data from older learners inasmuch as the erroneous production of inflection is a very marginal characteristic of the L2 learners' grammar (García Mayo & Villarreal Olaizola, 2011; Villarreal Olaizola, 2011; Villarreal Olaizola & García Mayo, 2009). Although there seems to be a subtle trend indicating that commission errors are more frequent with copula *be* compared to affixal inflection, the rates of commission in the three verbal contexts are too low to draw categorical conclusions in this regard.

5.2.4. Subject contexts

Figure 4 shows that CLIL and non-CLIL learners were found to omit referential subjects very infrequently (3.53% and 2.41%, respectively) compared with expletive subjects (CLIL: 56.52%, non-CLIL: 63.46%). The rate of omission of expletive subjects is significantly higher than that of referential subjects in both the CLIL (z=-3.913, p<.001***) and the non-CLIL group (z=-3.651, p<.001***).

Figure 4: Mean percent and Wilcoxon test results for subject omission



CLIL and non-CLIL learners thus behave alike in the provision of obligatory subjects in English. The fact that expletive subjects are significantly more problematic than referential subjects for both groups, especially for the non-CLIL learners, comes as no surprise, owing to the fact that expletive subjects have been claimed to entail

important complications for learners with pro-drop L1s such as Spanish (in line with studies in non-CLIL contexts such as White's, 1986, 1989, 2003b).

5.3. Discussion

This section elaborates on the data presented above to tackle first the conclusions from the between group comparisons and then the implications of the within-groups analysis.

Our first research question addressed the potential intergroup comparisons in terms of the provision of agreement morphology and overt subjects. Our data suggest that, overall, additional long-term CLIL exposure at a younger age does not have a significantly strong impact on the learners' L2 English grammar compared to EFL instruction, at least in the sample under analysis here. Still, the CLIL L2 learners who participated in this investigation do show a somehow more target-like performance than the learners enrolled in traditional EFL lessons only. This can be observed in the former's lower omission rate of expletive subjects, minor incidence of placeholder *is* and better command of auxiliary *be*. None of these trends found statistical support, but this fact does not preclude the observation of latent and insightful trends which help to shed light on the current discussion. Even though the use of placeholder *is* is agreed to signal an early stage of the acquisition of L2 English, our data reveal that this resource is half as common in the CLIL group as in the non-CLIL one. Closely related to this is the fact that CLIL learners show a lower omission rate and a negligible commission rate of auxiliary *be*. This result is relevant inasmuch as it has been previously attested in EFL contexts and with older learners that this auxiliary is acquired earlier than other auxiliary verbs and it plays a decisive role in resetting the unmarked null subject parameter in their L1 (i.e. Spanish) to the marked (i.e. non-pro-drop) value in the L2 English (Ruiz de Zarobe, 1998). It has also been observed that it is the CLIL group that performed slightly better as regards the provision of expletive subjects, though transfer from their null-subject L1 is still considerably frequent. The fact that Spanish lacks purely grammatical subjects like the English expletive pronouns *it* and *there* makes it quite difficult for these learners to understand their obligatoriness and thus to rapidly acquire a target-like performance in this respect. In sum, although it is true that the CLIL learners in this sample are far from having a target-like performance in terms of inflectional morphology and obligatory subjects, the previous observations point to a subtly more advanced morphosyntactic stage of these learners over their (age-matched) non-CLIL homologous.

The lack of highly significant differences between the provision of inflectional morphology and overt subjects between the CLIL and non-CLIL learners may stem

from two different factors. On the one hand, the age of the participants in this study. There is evidence in prior literature of the importance of age at testing, with older EFL learners outperforming younger CLIL and non-CLIL groups (Martínez Adrián & Gutiérrez Mangado, 2015a; Villarreal Olaizola, 2011), and also age of first exposure, with larger exposure being translated into faster rates of acquisition only when the child achieves a somehow mature cognitive development (Lázaro Ibarrola, 2012; cf. Lázaro Ibarrola, 2002 and Muñoz, 2006 on additional exposure at earlier stages). Lázaro Ibarrola (2002, 2012) observes that three years of additional CLIL exposure in a group of 15 years-olds who had been exposed to English since they were 5 explains the learners' higher provision of irregular past forms and the fact that they do not use placeholder *is* compared to EFL learners, while their production of affixal morphemes does not differ significantly from that of their EFL homologous. Our study complements these results by exploring learners with a similar onset age (5-6) but with an earlier age at testing (11-12) and a longer additional CLIL exposure (6 years). With it, we show that long-term exposure through CLIL (cf. Lázaro Ibarrola, 2002 and Muñoz, 2006 on EFL contexts) does not result in a significantly faster rate of acquisition compared to EFL instruction, but only in certain positive signs of a slightly advanced stage of acquisition. The age at testing may well have conditioned our results in that our learners seem not to have reached yet the stage at which the acquisition of morphosyntactic features accelerates, at around 12-13 (Lázaro Ibarrola, 2002, 2012). In this sense, it would be interesting to examine the morphosyntactic development of these learners longitudinally, when they surpass that age, whether they are still enrolled in CLIL programmes or not (as in Villarreal Olaizola, 2011, where the positive effect of three years of CLIL disappears one year after the CLIL Secondary Education programme is over).

On the other hand, the focus on meaning and communicative interaction that characterises CLIL programmes might be also at play in the lack of significant intergroup comparisons. In line with previous studies, our investigation underscores the fact that "mere exposure to quantity and quality input seems not to be enough to develop productive skills or accuracy rates to a target-like level" (Villarreal Olaizola, 2011: 205); otherwise, the extra 488 hours of exposure that the CLIL group has received in addition to their EFL lessons should have translated in a significantly better and more target-like performance than the group receiving EFL instruction only. As a matter of fact, our research calls for a focus on form in CLIL programmes and, particularly, for "grammar instruction [...] in context" (Slabakova, 2008, p. 407), as "[t]he explicit knowledge acquired in their EFL lessons [...] might be hard to retrieve in the context of communicative interaction" (Martínez Adrián & Gutiérrez Mangado, 2015a, p. 69; see also Ellis, 2001; García Mayo & Villarreal Olaizola, 2011; Martínez Adrián & Gutiérrez Mangado, 2015a; Muñoz, 2007; Pérez-Vidal, 2007; Pica, 2002;

Villarreal Olaizola, 2011). This is particularly relevant for the acquisition of affixal inflection, for which exposure alone has been claimed to be insufficient for a target-like performance (García Mayo & Villarreal Olaizola, 2011), owing to the fact that it is "often perceived as semantically redundant and having little communicative value" (Pawlak, 2008: 188). Although the third person morpheme has been found to be acquired rather late compared to other morphemes, there is evidence of the positive impact of planned form-focused instruction on its acquisition. Pawlak (2008), for instance, found that teenage Polish L2 English learners were more target-like in the provision of third person *–s* after treatment sessions with implicit corrective feedback in the form of recasts and output enhancement (i.e. clarification requests). More relevant is Basterrechea and García Mayo's (2014) study, where teenage CLIL learners were found to benefit from form-focused instruction through dictoglosses, with CLIL learners showing a higher – though non-significant – provision of affixal inflection than their EFL homologous, especially when the task was completed in pairs.

Our second research question focused on the intragroup differences as to the two main features under analysis: agreement morphology errors and null subjects. In this regard, one must consider that this investigation explores the learners' oral production, so the processing difficulties and pressure that this task may involve for the learner must not be underestimated. In the case of affixal inflection, for instance, this often results in the use of uninflected forms as the default option (Ionin, 2013; Prévost & White, 2000; cf. Villarreal Olaizola, 2011 on written data). Still, the general trends observed in this study conform to what previous studies on older learners have attested: a higher omission of affixal compared to suppletive inflection (as in García Mayo & Villarreal Olaizola, 2011; Villarreal Olaizola & García Mayo, 2009; Villarreal Olaizola, 2011). As discussed above, the explanation to this finding is rooted in minimalist theory, with copula and auxiliary *be* checking their agreement and tense features overtly and thus expressed morphologically, and lexical verbs checking theirs covertly, with or without overt morphological expression (Ionin & Wexler, 2002). This difference makes it more complicated for learners to achieve a target-like performance in this respect (García Mayo & Villarreal Olaizola, 2011), as the young learners who participated in this study demonstrated. Their performance has nonetheless conformed to the well-attested asymmetry between affixal and suppletive morphology, as is evident from their overall greater mastery of suppletive over affixal inflection and copula *be* over auxiliary *be*. Although the CLIL learners do show an apparent more target-like provision of suppletive inflection, their problems with affixal morphology and null subjects do not differ significantly from the non-CLIL group's morphosyntax. Hence, it seems that our data provide support for previous claims on the provision of these morphosyntactic features being independent of the type of instruction and increased exposure (García Mayo & Villarreal Olaizola, 2011; Martínez Adrián & Gutiérrez Mangado, 2015a; Villarreal Olaizola, 2011; Villarreal Olaizola & García Mayo, 2009).

The results also indicate that the grammar of our young L2 English learners seems to be unimpaired. Both the CLIL and non-CLIL learners in this study were found to have a good mastery of referential subjects but, at their early age, the input they have received may not have been enough to acquire the obligatory use of expletive subjects. Their use of placeholder *is* also evidences a quite underdeveloped L2 inflectional system (cf. the better performance of the slightly older L2 learners in Lázaro Ibarrola, 2012; Martínez Adrián & Gutiérrez Mangado, 2009), even in the case of the group that receives long-term additional exposure to the target language through CLIL instruction. However, while both CLIL and non-CLIL learners have certain problems to supply the inflected forms correctly, particularly the third person singular bound morpheme, they nonetheless seem well aware of the erroneous placement of affixal inflection in contexts other than the third person singular as well as of the wrong provision of inflected suppletive forms. Accordingly, they do not use inflectional morphology unsystematically; they use a high rate of bare uninflected forms as a default instead and barely produce commission errors, which points in the opposite direction of a possible representational impairment as regards agreement morphology (see Ionin & Wexler, 2002). This result is in keeping with previous investigations with older learners which claim that errors and variability in the provision of inflection just result from a mapping problem whereby learners are not able to supply the correct inflectional form even though both tense and agreement features are already available in their L1 (i.e. Spanish in this case) (García Mayo & Villarreal Olaizola, 2011; Villarreal Olaizola, 2011). An exception to this trend is copula *be* in our CLIL group. Its slightly higher commission error rate (significantly higher than that of affixal inflection) has, to the best of our knowledge, no prior antecedents in the literature, which calls for further research to try to determine the source of this unexpected finding. As a tentative hypothesis, we argue that this divergent trend may stem from the CLIL approach itself. It must be taken into account that, on the one hand, CLIL lessons have been claimed to focus more on meaning than on form in comparison with traditional EFL classes (Martínez Adrián, 2011) and, on the other, corrective feedback has been found to be less frequent in CLIL than in EFL settings (Milla & García Mayo, 2014), two aspects which may well have had an effect on the data attested in this investigation and, thus, should not be disregarded. The variety of forms that copula *be* takes might be regarded as an alternative explanation, but it seems unlikely in view of the lower rate of commission of auxiliary *be* in these same learners.

6. Conclusions

The main aim of this paper is to contribute to the existing debate on the potential benefits of CLIL programmes for the learner's L2 English morphosyntax with data from the Spanish monolingual community of Cantabria. To this end, we assessed

the potential impact of an early start and a long-term exposure to the target language through CLIL instruction on young (11- and 12-year-old) English learners' provision of agreement morphology and overt subjects, compared to age-matched learners with the same onset age but enrolled in traditional EFL programmes. Our investigation reveals that the provision of specific morphosyntactic characteristics such as inflection and obligatory subjects do not benefit so clearly from an earlier start and longer exposure to English through the increasingly popular CLIL programmes. Although we did find positive results for the CLIL group that point to a potential impact of CLIL instruction, such as a lower rate of placeholder *is* and null subjects, the effect of CLIL on these Spanish young learners' English morphosyntax is not highly significant compared to traditional EFL instruction.

In keeping with prior literature on the topic with older Spanish L2 English learners, our findings suggest that CLIL programmes may be more effective for the right provision of agreement morphology and overt subjects. Nonetheless, our research also calls for further focus on form and corrective feedback in CLIL programmes to achieve greater effectiveness and to try to minimise one of the well-attested weaknesses of this educational approach (Ball, 2016; Ball et al., 2015). These observations thus need further research to gauge the effect of implicit corrective feedback (Milla & García Mayo, 2014) and collaborative tasks (for Primary School learners, see García Mayo & Imaz Aguirre, 2019; for adult learners, García Mayo & Azkarai Garai, 2016 and Payant & Kim, 2019) on the learner's attention to form while engaged in meaning-focused communicative interaction. The triangulation with qualitative data coming from the observations of CLIL lessons can further provide insight into more effective ways of integrating the focus on formal aspects into content lessons. CLIL instructors should ideally receive training courses where the language component of their lessons is highlighted. They must be empowered so that they are able to identify the key language needed for their content units, make it more salient (Ball, 2016), and incorporate activities where students' knowledge of English grammar is called upon in order to process content information (see Ting, 2011). As clear from cross-linguistic influence being the source of some of the errors reported (e.g. subject omission), it is also suggested that CLIL instructors do not treat English in isolation from students' L1, a recommendation which holds true for EFL teachers as well.

The investigation has nonetheless some limitations. Firstly, the instrument used to collect the data complicates the comparison with homologous investigations using the story *Frog, where are you?* (e.g. García Mayo & Villarreal Olaizola, 2011; Villarreal Olaizola, 2011; Villarreal Olaizola & García Mayo, 2009). Secondly, it is necessary to compare these results with the learners' written production to gauge the potential effect of the processing difficulties and pressure inextricably linked to the oral task on

their provision of the morphosyntactic features surveyed (as in Villarreal Olaizola's 2011 tentative approximation).

This investigation would also benefit from an in-depth analysis of sequences which have not been included here but could provide insightful observations on both inflection and subject omission, such as utterances involving autocorrection and repair sequences, with which Martínez Adrián and Gutiérrez Mangado (2015a) observed an impact of the focus on meaning of CLIL instruction on the low number of successfully repaired sequences by teenage Spanish L3 English learners (also Lázaro & García Mayo, 2012). In view of the fairly high standard deviations in the data reported, it is necessary to carry out a more fine-grained exploration of individual trends. It would also be very interesting to measure the learners' target-like performance by assessing their frequency in use of affixal morphology and explore the relation between finiteness and subject omission, as both aspects will surely help to shed light on the observations and claims presented in this paper. Another aspect that would reward further research on the potential differences between our CLIL and non-CLIL learners is the examination of other morphosyntactic features and forms, as is the case of the past inflection (as in García Mayo & Villarreal Olaizola, 2011; Lázaro Ibarrola, 2012; Martínez Adrián & Gutiérrez Mangado, 2015a; Villarreal Olaizola, 2011; Villarreal Olaizola & García Mayo, 2009) and article omission and overuse (as in Gutiérrez-Mangado & Martínez-Adrián, 2018). As a final remark, designs where different degrees of intensity of the CLIL programme are compared (see Merino & Lasagabaster, 2018), particularly with a longitudinal research approach, should be welcome.

Acknowledgements

This research is part of the project *Bilingual teaching and learning in Cantabria: From primary to tertiary education*, funded by the University of Cantabria (ref. UC2016-GRE-10). We are also grateful to the Primary Schools, the students and the teachers who agreed to participate in this study, as well as to Pedro Alberto San Emeterio Bolado for the vignettes.

7. References

Aguilar, M. & Muñoz, C. (2014) The effect of proficiency on CLIL benefits in Engineering students in Spain. *International Journal of Applied Linguistics*, 24(1), 1-18.

Ball, P. (2016). Using language(s) to develop subject competences in CLIL-based practice. *Pulso. Revista de educación*, 39, 15-34.

Ball, P., Kelly, K. & Clegg, J. (2015) *Putting CLIL into Practice*. Oxford: Oxford University Press.

Basterrechea, M. & García Mayo, M. P. (2014) Dictogloss and the production of the English third person –*s* by CLIL and mainstream EFL learners: A comparative study. *International Journal of English Studies*, 14(2), 77-98.

Blanco-Suárez, Z., Gallardo-del-Puerto, F. & Gandón-Chapela, E. (2020) *The Primary Education Learners' English Corpus* (PELEC): Design and compilation. *RiCL, Research in Corpus Linguistics*, 8(1), 147-163.

Brown, R. (1973) *A First Language: The Early Stages*. Cambridge, MA: Harvard University Press.

Cadierno, T., Hansen, M., Lauridsen, J. T., Eskildsen, S. W, Fenyvesi, K., Jensen, S. H. & aus der Wieschen, M. V. (2020) Does younger mean better? Age of onset, learning rate and short-term L2 proficiency in young Danish learners of English. *Vigo International Journal of Applied Linguistics*, 17, 57-86.

Chomsky, N. (1981) *Lectures on Government and Binding*. Dordrecht: Foris.

Coyle, D. (2010) Foreword. In D. Lasagabaster & Y. Ruiz de Zarobe (eds) *CLIL in Spain: Implementation, Results and Teacher Training*. (pp. vii-viii) Newcastle upon Tyne: Cambridge Scholars.

Dalton-Puffer, C. (2008) Outcomes and processes in Content and Language Integrated Learning (CLIL): Current research from Europe. In W. Delanoy & L. Volkmann (eds) *Future Perspectives for English Language Teaching*. (pp. 139-157) Heidelberg: Carl Winter.

Ellis, R. (2001) Introduction: Investigating form-focused instruction. *Language Learning*, 51(s1), 1-46.

Eurydice. (2006) *Content and Language Integrated Learning at School in Europe*. Brussels: Eurydice European Unit.

Fernández Fontecha, A. (2009) Spanish CLIL: Research and official actions. In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (eds) *Content and Language Integrated Learning: Evidence from Research in Europe*. (pp. 3-21) Clevedon: Multilingual Matters.

Gallardo del Puerto, F., Gómez Lacabex, E & García Lecumberri, M. L. (2009) Testing the effectiveness of Content and Language Integrated Learning in foreign language contexts: The assessment of English pronunciation. In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (eds) *Content and Language Integrated Learning: Evidence from Research in Europe*. (pp. 63-80) Clevedon: Multilingual Matters.

___ & Martínez Adrián, M. (2013) ¿Es más efectivo el aprendizaje de la lengua extranjera en un contexto AICLE? Resultados de la investigación en España. *Padres y Maestros*, 34, 25-28.

García Mayo, M. P. (2003) Age, length of exposure and grammaticality judgments in the acquisition of English as a foreign language. In M. P. García Mayo & M. L. García Lecumberri (eds) *Age and the Acquisition of English as a Foreign Language*. (pp. 94-114) Clevedon: Multilingual Matters.

___ & Azkarai Garai, A. (2016) EFL task-based interaction: Does task modality impact on language-related episodes? In M. Sato & S. Ballinger (eds) *Peer Interaction and Second Language Learning: Research Agenda and Pedagogical Implications*. (pp. 241-266) Amsterdam & Philadelphia: John Benjamins.

___ & Imaz Agirre, A. (2019) Task modality and pair formation method: Their impact on patterns of interaction and attention to form among EFL primary school children. *System: An International Journal of Educational Technology and Applied Linguistics*, 80, 165-175.

___, Lázaro Ibarrola, A. & Liceras, J. M. (2005) Placeholders in the English interlanguage of bilingual (Basque/Spanish) children. *Language Learning*, 55(3), 445-489.

___ & Villarreal Olaizola, I. (2011) The development of suppletive and affixal tense and agreement morphemes in the L3 English of Basque-Spanish bilinguals. *Second Language Research*, 27(1), 129-149.

Goad H., White, L. & Steele, J. (2003) Missing inflection in L2 acquisition: Defective syntax or L1-constrained prosodic representations? *Canadian Journal of Linguistics*, 48, 243-263.

Guilfoyle, E. (1984) The acquisition of tense and the emergence of thematic subjects in child grammars of English. *The McGill Working Papers in Linguistics*, 2, 20-30.

Gutiérrez-Mangado, M. J. & Martínez-Adrián, M. (2018) CLIL at the linguistic interfaces. *International Journal of Immersion and Content-Based Education*, 6(1), 85-112.

Gutiérrez Martínez, A. & Ruiz de Zarobe, Y. (2017) Comparing the benefits of a metacognitive reading strategy instruction programme between CLIL and EFL Primary School students. *Estudios de Lingüística Inglesa Aplicada*, 17, 71-92.

Hawkins, R. & Casillas, G. (2008) Explaining frequency of verb morphology in L2 early speech. *Lingua*, 118, 595-612.

___ & Chan, Y-C. (1997) The partial availability of Universal Grammar in second language acquisition: The 'failed features' hypothesis. *Second Language Research*, 13, 187-226.

Haznedar, B. (2001) The acquisition of the IP system in child L2 English. *Studies in Second Language Acquisition*, 23, 1-39.

___ & Schwartz, B. (1997) Are there optional infinitives in child L2 acquisition? In E. Hughes, M. Hughes & A. Greenhill (eds) *Proceedings of the 21st Boston University Conference on Language Development*. (pp. 257-268) Somerville, MA: Cascadilla Press.

Hyams, N. (1989) The Null Subject Parameter in language acquisition. In O. Jaeggli & N. Hyams (eds) *The Null Subject Parameter*. (pp. 215-238) Dordrecht: Kluwer.

Ionin, T. (2013) Morphosyntax. In J. Herschensohn & M. Young-Scholten (eds) *The Cambridge Handbook of Second Language Acquisition*. (pp. 505-528) Cambridge: Cambridge University Press.

___ & Wexler, K. (2002) Why is 'is' easier than '*-s*'?: Acquisition of tense/agreement morphology by child second language learners of English. *Second Language Research*, 18, 95-136.

Jessner, U. (2014) On multilingual awareness or why the multilingual learner is a specific language learner. In M. Pawlak & L. Aronin (eds) *Essential Topics in Applied Linguistics and Multilingualism: Studies in Honor of David Singleton*. (pp. 175-184) Wien: Springer.

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*, 15(3), 23-27.

Juan-Garau, M. & Pérez Vidal, C. (2011) Trilingual primary education in the Balearic Islands. In I. Bangma, C. van der Meer & A. Riemersma (eds) *Trilingual Primary Education in Europe: Some Developments with Regard to the Provisions of Trilingual Primary Education in Minority Language Communities of the European Union*. (pp. 129-142) Leeuwarden: Fryske Akademy.

Judy, T. (2011) L1/L2 parametric directionality matters: More on the Null Subject Parameter in L2 acquisition. *EUROSLA Yearbook*, 11, 165-190.

___ & Rothman, J. (2010) From a superset to a subset grammar and the semantic compensation hypothesis: Subject pronouns and anaphora resolution in L2 English. In K. Franich, K. M. Iserman & L. L. Keil (eds) *BUCLD 34: Proceedings of the 34th Annual Boston University Conference on Language Development*. (pp. 197-208) Sommerville, MA: Cascadilla Press.

Lardiere, D. (2000) Mapping features to forms in second language acquisition. In J. Archibald (ed) *Second Language Acquisition and Linguistic Theory*. (pp. 102-129) Cambridge, MA: Blackwell.

___. (2008) Feature assembly in second language acquisition. In J. M. Liceras, H. Zobl & H. Goodluck (eds) *The Role of Formal Features in Second Language Acquisition*. (pp. 106-140) New York: Lawrence Erlbaum Associates.

___. (2009) Some thoughts on the contrastive analysis of features in second language acquisition. *Second Language Research*, 25(2), 173-227.

Lasagabaster, D. (2008) Foreign language competence and language integrated courses. *The Open Applied Linguistics Journal*, 1, 31-42.

Lázaro Ibarrola, A. & García Mayo, M. P. (2012) L1 use and morphosyntactic development in the oral production of EFL learners in a CLIL context. *International Review of Applied Linguistics*, 50, 135-160.

Lázaro Ibarrola, A. (2002) *La Adquisición de la Morfosintaxis del Inglés por Niños Bilingües Euskera/Castellano: Una Perspectiva Minimalista*. Unpublished PhD dissertation. Department of English and German, University of the Basque Country (Spain).

___. (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), 79-96.

MacWhinney, B. (2000) *The CHILDES Project: Tools for Analyzing Talk*. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Marsh, D. (1994) *Bilingual Education & Content and Language Integrated Learning.* Paris: International Association for Cross-cultural Communication, Language Teaching in the Member States of the European Union (Lingua) University of Sorbonne.

___. (ed) (2002) *CLIL/EMILE – The European Dimension: Actions, Trends and Foresight Potential*. Brussels: The European Commission.

Martínez Adrián, M. (2011) An overview of Content and Language Integrated Learning: Origins, features and research outcomes. *Huarte de San Juan. Filología y Didáctica de la Lengua*, 11, 93-101.

___ & Gutiérrez Mangado, M. 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) Clevedon: Multilingual Matters.

___ & Gutiérrez Mangado, M. J. (2015a) Is CLIL instruction beneficial in terms of general proficiency and specific areas of grammar? *Journal of Immersion and Content-Based Language Education*, 3(1), 51-76.

___ & Gutiérrez Mangado, M. J. (2015b) L1 use, lexical richness, accuracy and complexity in CLIL and NON-CLIL learners. *Atlantis, Journal of the Spanish Association for Anglo-American Studies*, 37(2), 175-200.

Merino, J. A. & Lasagabaster, D. (2018) The effect of Content and Language Integrated Learning programmes' intensity on English proficiency: A longitudinal study. *Journal of Applied Linguistics*, 28, 18-30.

Milla, R. & García Mayo, M. P. (2014) Corrective feedback episodes in oral interaction: A comparison of a CLIL and an EFL classroom. *International Journal of English Studies*, 14(1), 1-20.

Montrul, S. (2011) Multiple interfaces and incomplete acquisition. *Lingua*, 121, 591-604.

Muñoz, C. (2002) Relevance and potential of CLIL. In D. Marsh (ed) *CLIL/EMILE – The European Dimension: Actions, Trends and Foresight Potential*. (pp. 33-36) Brussels: The European Commission.

___. (2006) Accuracy orders, rate of learning and age in morphological acquisition. In C. Muñoz (ed) *Age and the Rate of Foreign Language Learning*. (pp. 107-125) Clevedon: Multilingual Matters.

___. (2007) CLIL: Some thoughts on its psycholinguistic principles. *RESLA, Revista Española de Lingüística Aplicada*, 20, 17-26.

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.

Ortega Durán, M. (2016) *Crosslinguistic Influence in L2 English Oral Production: The Effects of Cognitive Language Learning Abilities and Input*. Unpublished PhD dissertation. Department of Modern Language and Literatures and of English Studies, University of Barcelona (Spain).

Pawlak, M. (2008) The effect of corrective feedback on the acquisition of the English third-person *–s* ending. In D. Gabryś-Barker (ed) *Morphosyntactic Issues in Second Language Acquisition*. (pp. 187-202) Clevedon: Multilingual Matters.

Payant, C. & Kim, Y. (2019) Impact of task modality on collaborative dialogue among plurilingual learners: A classroom-based study. *International Journal of Bilingual Education and Bilingualism*, 22(5), 614-627.

Pérez-Cañado, M. L. (2012) CLIL research in Europe: Past, present, and future. *International Journal of Bilingual Education and Bilingualism*, 15(3), 315-341.

Pérez-Vidal, C. (2007) The need for Focus on Form (FoF) in Content and Language Integrated approaches: An exploratory study. *RESLA, Revista Española de Lingüística Aplicada*, 1, 39-45.

___ & Juan-Garau, M. (2010) To CLIL or not to CLIL? From bilingualism to multilingualism in Catalan/Spanish communities in Spain. In D. Lasagabaster & Y.

Ruiz de Zarobe (eds) *CLIL in Spain: Implementation, Results and Teacher Training*. (pp. 115-139) Newcastle upon Tyne: Cambridge Scholars.

___ & Juan-Garau, M. (2011) Trilingual Primary Education in Catalonia. In I. Bangma, C. van der Meer & A. Riemersma (eds) *Trilingual Primary Education in Europe: Some Developments with Regard to the Provisions of Trilingual Primary Education in Minority Language Communities of the European Union*. (pp. 68-92) Leeuwarden: Fryske Akademy.

___ & Roquet, H. (2015) The linguistic impact of a CLIL *Science* programme: An analysis measuring relative gains. *System*, 54, 80-90.

Phinney, M. (1987) The Pro-Drop Parameter in second language acquisition. In T. Roeper & E. Williams (eds) *Parameter Setting*. (pp. 221-238) Dordrecht: D. Reidel.

Pica, T. (2002) Subject matter content: How does it assist the interactional and linguistic needs of classroom language learners? *The Modern Language Journal*, 86(1), 1-19.

Pladevall Ballester, E. (2012) Child L2 English acquisition of subject properties in an immersion bilingual context. *Second Language Research*, 28(2), 217-241.

___. (2013) Adult instructed SLA of English subject properties. *Canadian Journal of Linguistics*, 58(3), 465-486.

Prévost, P. & White, L. (2000) Missing surface inflection or impairment in second language acquisition? Evidence from tense and agreement. *Second Language Research*, 16, 103-133.

Rizzi, L. (1993) Some notes on linguistic theory and language development: The case of root infinitives. *Language Acquisition*, 3(4), 371-393.

Ruiz de Zarobe, Y. (1997) Comportamiento de los pronombres expletivos en inglés: Aspectos contrastivos entre la primera y la segunda lengua. *Cuadernos de Investigación Filológica*, 23-24, 7-15.

___. (1998) Uniformidad morfológica y adquisición de sujetos en inglés lengua extranjera. *Langues et Linguistique*, 24, 171-186.

___. (2000) Concordancia copulativa, pronombres sujeto y adquisición de sistemas no-nativos. *Linguistica XL*, 2, 327-333.

___. (2007) CLIL in a bilingual community: Similarities and differences with the learning of English as a foreign language. *Vienna English Working Papers*, 16(3), 47-52.

___. (2008) CLIL and Foreign Language Learning: A longitudinal study in the Basque Country. *International CLIL Research Journal*, 1(1), 60-73.

___. (2011) Which language competencies benefit from CLIL? An insight into applied linguistics research. In Y. Ruiz de Zarobe, J. Sierra & F. Gallardo del Puerto

(eds) *Content and Foreign Language Integrated Learning: Contributions to Multilingualism in European Contexts*. (pp. 129-153) Bern: Peter Lang.

___ & Lasagabaster, D. (eds) (2010) *CLIL in Spain: Implementation, Results and Teacher Training*. Newcastle upon Tyne: Cambridge Scholars.

San Isidro, X. (2010) An insight into Galician CLIL: Provision and results. In D. Lasagabaster & Y. Ruiz de Zarobe (eds) *CLIL in Spain: Implementation, Results and Teacher Training*. (pp. 55-78) Newcastle upon Tyne: Cambridge Scholars.

___ & Lasagabaster, D. (2019a) The impact of CLIL on pluriliteracy development and content learning in a rural multilingual setting: A longitudinal study. *Language Teaching Research*, 23(5), 584-602.

___ & Lasagabaster, D. (2019b) Code-switching in a CLIL multilingual setting: A longitudinal qualitative study. *International Journal of Multilingualism*, 16(3), 336-356.

Slabakova, R. (2008) *Meaning in the Second Language*. Berlin: Mouton de Gruyter.

Ting, T. (2011) CLIL and Neuroscience: How are they related? In Y. Ruiz de Zarobe, J. Sierra & F. Gallardo del Puerto (eds) *Content and Foreign Language Integrated Learning: Contributions to Multilingualism in European Contexts*. (pp. 75-101) Bern: Peter Lang.

Tsimpli, I. M. & Dimitrakopoulou, M. (2007) The interpretability hypothesis: Evidence from *wh*-interrogatives in second language acquisition. *Second Language Research*, 23, 215-242.

Villarreal Olaizola, I. (2011) *Tense and Agreement in the Non-Native English of Basque-Spanish Bilinguals: Content and Language Integrated Learners vs. English as a School Subject Learners*. Unpublished PhD dissertation. Department of English and German, University of the Basque Country (Spain).

___ & García Mayo, M. P. (2009) Tense and agreement morphology in the interlanguage of Basque/Spanish bilinguals: CLIL versus non-CLIL. In Y. Ruiz de Zarobe & R. M. Jiménez Catalán (eds) *Content and Language Integrated Learning: Evidence from Research in Europe*. (pp. 157-175) Clevedon: Multilingual Matters.

White, L. (1986) Implications of parametric variation for adult second language acquisition: An investigation of the Pro-Drop Parameter. In V. Cook (ed) *Experimental Approaches to Second Language Acquisition*. (pp. 55-72) Oxford: Pergamon.

___. (1989) *Universal Grammar and Second Language Acquisition*. Amsterdam & Philadelphia: John Benjamins.

___. (2003a) Fossilization in steady state L2 grammars: Persistent problems with inflectional morphology. *Bilingualism: Language and Cognition*, 6(2), 129-141.

___. (2003b) *Second Language Acquisition and Universal Grammar.* Cambridge: Cambridge University Press.

Zobl, H. & Liceras, J. (1994) Functional categories and acquisition orders. *Language Learning*, 44, 159-180.

Appendix. The 8-vignette story used to collect the data

1.

2.

3.

4.

Human evaluation of three machine translation systems: from quality to attitudes by professional translators

Anna Fernández-Torné
Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental
Universitat Autònoma de Barcelona
anna.fernandez.torne@uab.cat

Anna Matamala
Departament de Traducció i d'Interpretació i d'Estudis de l'Àsia Oriental
Universitat Autònoma de Barcelona
anna.matamala@uab.cat

Abstract

This article aims to compare three machine translation systems with a focus on human evaluation. The systems under analysis are a domain-adapted statistical machine translation system, a domain-adapted neural machine translation system and a generic machine translation system. The comparison is carried out on translation from Spanish into German with industrial documentation of machine tool components and processes. The focus is on the human evaluation of the machine translation output, specifically on: fluency, adequacy and ranking at the segment level; fluency, adequacy, need for post-editing, ease of post-editing, and mental effort required in post-editing at the document level; productivity (post-editing speed and post-editing effort) and attitudes. Emphasis is placed on human factors in the evaluation process.

Keywords: machine translation, quality evaluation, human evaluation, automatic metrics, post-editing effort

Resumen

En este artículo se comparan tres sistemas de traducción automática poniendo especial atención en la evaluación humana. Los sistemas analizados son un sistema estadístico de traducción automática con adaptación al dominio, un sistema neuronal de traducción automática con adaptación al dominio y un sistema de traducción automática genérico. La comparación se lleva a cabo en una traducción del español al alemán de documentación industrial de componentes y procesos de máquina herramienta. El estudio se centra en la evaluación humana de la traducción automática, en concreto en los siguientes aspectos: fluidez, adecuación y ranquin a nivel de segmento; fluidez, adecuación, necesidad de posedición, facilidad de posedición

y esfuerzo mental requerido en la posedición a nivel de documento; productividad (velocidad de posedición y esfuerzo de posedición) y actitudes. Se hace énfasis en los factores humanos del proceso de evaluación.

Palabras clave: traducción automática, evaluación de la calidad, evaluación humana, métricas automáticas, esfuerzo de posedición

1. Introduction

Machine translation research has seen two interesting developments in recent years: firstly, the rise of neural machine translation (Cho et al., 2014; Castilho et al., 2017) and, secondly, the willingness to go beyond automated metrics such as BLEU (Papineni et al., 2002) or TER (Snover et al., 2006) and seek feedback from human participants (Callison-Burch et al., 2012). Our research combines the two approaches and aims to compare three different machine translation systems with a focus on human evaluation in a domain-specific scenario where resources are scarce. Making proper domain adaptation is not only a goal but also a challenge which has been researched extensively within statistical machine translation (SMT) (Foster & Kuhn, 2007, Axelrod et al., 2011; Bisazza et al., 2011; Gascó et al., 2012; Sennrich, 2012; Eetemadi et al., 2015), but to a lesser extent in neural machine translation (NMT) (Luong & Manning, 2015; Freitag & Al-Onaizan, 2016; Crego et al., 2016). Large generic machine translation systems are freely available online and are used even in cases where domain-adapted systems may be more suitable, but evaluations comparing large generic MT systems and domain-adapted systems, with an emphasis on human evaluation, are missing. The systems under analysis in our research are a domain-adapted statistical machine translation (SMT) system, a domain-adapted neural machine translation (NMT) system and a generic machine translation (GT) system. They have been compared in three domains and language pairs: reports and press releases from non-profit international organisations (from English into Spanish) (INTORG), industrial documentation of machine tool components and processes (MTOOL) (from Spanish into German), and the installation and maintenance documentation for elevators (ELEV) (from Spanish into French). Automated metrics have been computed and the global results have already been presented (Etchegoyhen et al., 2018), but the aim of this article is to focus on the results of just one language pair and domain, namely MTOOL (from Spanish into German), in order to provide a more thorough discussion, with a focus on human factors that were not previously discussed.

Section 2 describes the corpora and models used. Section 3 describes methodological aspects, i.e. the measures used in the human evaluation, the tool

selected to perform the experiment, the participants' selection procedure and profile, and the development of the test. Section 4 discusses the results of the MTOOL evaluation. The article concludes with some thoughts on future research avenues concerning human factors in machine translation research.

2. Corpora and MT models in MTOOL

This research was developed as part of the AdapTA project, in which data were obtained from project partners. For the MTOOL (from Spanish into German) corpus, the training data provided by the partner specialised in the domain were particularly scarce. Thus, only 25,256 parallel segments were gathered, 1,984 segments were used as development sets, and three sets of 50 sentence pairs were selected for testing purposes. The selection was not random but took into account three factors to guarantee that sentences were representative and could be used in tasks replicating a real professional scenario: the presence of specific domain vocabulary, the average sentence length (as in any technical field, sentences are mostly short in this domain too), and the presence of coherent segments at the contextual level. The content was highly specialised and dealt with industrial documentation of machine tool components and processes. This scenario replicates a typical situation for which there is a high demand from a professional point of view and limited training resources. To complement the scarce data provided in the form of translation memories for MTOOL, out-of-domain data were compiled, with a total of 1,784,385 additional segments obtained from freely available corpora (see Etchegoyhen et al., 2018 for further technical details). The main function of these generic datasets was to serve as a basis for the NMT models.

As explained by Etchegoyhen et al. (2018), SMT systems were phrase-based models built with Moses (Koehn et al. 2007), with phrases of maximum length 5 and n-gram language models of order 5 built with KenLM (Heafield, 2011). For NMT, the attention-based encoder-decoder approach (Bahdanau et al., 2015) was followed, using the OpenNMT toolkit (Klein et al., 2017). The translations from the online generic system were obtained from Google Translate in June 2017 in which, to the best of our knowledge, translations from Spanish into German were produced using their phrase-based SMT engine. Domain adaptation in MTOOL was carried out using the system that performed best during tests: for SMT, the phrases from the entire generic dataset were combined through fill-up and, for NMT, it was carried out through fine-tuning (Luong & Manning, 2015), by training the generic networks on the in-domain data.

3. Methodological aspects

A field quasi-experiment, i.e., one "taking place in real life [...] in which the criterion of randomization (of participants in a sample, for instance) cannot be met" (Van Peer, Hakemulder and Zyngier, 2012: 90) was planned. Thus, a more realistic but less controlled environment was favoured, prioritising its ecological validity since "a laboratory often fails to replicate the everyday conditions under which cultural phenomena occur" (ibid: 89). The aim was to gather both qualitative and quantitative data. The experiment was approved by UAB's Ethical Committee on Animal and Human Research (CEEAH) and, following the committee's advice, the experiment included one part in which the participants were paid their requested fees as professional translators to carry out a series of tasks and another part (replying to questionnaires) which was voluntary.

3.1. Selecting the measures for human evaluation

The human evaluation took into account three factors: quality at the segment and document levels, productivity, and translators' attitudes. At the segment level three indicators were gathered to assess quality: fluency, adequacy and ranking. Fluency is understood to be the extent to which a translated segment flows naturally in the target language without grammar and spelling mistakes and is considered to be genuine language by native speakers (Koehn and Monz, 2006). It was measured on a 1 to 4 scale, with 1 indicating that the text was incomprehensible and 4 indicating that the text was flawless, following TAUS guidelines. Adequacy, measured on another 4-point scale, with 1 indicating none of the meaning is represented in the translation and 4 indicating everything is represented in the translation, is considered to be the amount of information from the original segment that is present in the translated segment (Koponen, 2010). As regards ranking, it consists of placing in order different translated versions from the same original segment from best (1) to worst quality (3).

At the document level, the subjective perception of participants regarding five quality aspects was also gathered, namely fluency, adequacy, need for post-editing, ease of post-editing, and mental effort involved in the post-editing. Those five aspects were rated on a 10-point scale and were presented to the participants as follows, with not specific definition added:

- How fluent the raw machine translated text was, with 1 indicating that the text was not fluent and 10 indicating that it was very fluent.

- How much of the information in the source text was present in the raw machine translated text, with 1 indicating none of the information in

the source text was represented in the translation and 10 indicating all information in the source text was represented in the translation.

- How much post-editing the text required, with 1 indicating the translation required very few editing and 10 indicating the translation required a lot of editing.

- How easy the post-editing was, with 1 indicating the post-editing was found very difficult and 10 indicating the post-editing was found very easy.

- How much mental effort the post-editing required, with 1 indicating the post-editing required very low mental effort and 10 indicating the post-editing required a lot of mental effort.

They were also given the opportunity to add comments after each statement. Concerning productivity, the objective measures chosen were post-editing speed and post-editing effort. Post-editing speed refers to the "average number of words processed by the post-editor in a given timespan" (TAUS, n.d.), measured in words per hour. Post-editing effort is defined as "the average percentage of word changes applied by the post-editor on the MT output provided" (TAUS, n.d.). The effort is measured on a 0 to 10 scale in which 0 means that no changes needed to be made on the MT output and 10 implies that all the text or most of it was changed. It is based on the edit distance (Levenshtein's algorithm) normalised by segment length (i.e., divided by the number of characters of whatever segment is longer: either the automatically translated one or the post-edited one).

As regards attitudes, our aim was to assess the attitude of the participants prior to the test and its evolution during the experiment through a questionnaire administered before and after the three tasks. The questionnaire included a 1-to-10 scale in which participants gave their opinion on the following aspects:

- general machine translation quality (without post-editing), with 1 indicating the raw MT is of very poor quality and 10 indicating it is a very good raw MT.

- usefulness of machine translation for translators, with 1 indicating that MT is useless and 10 indicating it is very useful.

- inclination to use machine translation as a starting point, with 1 indicating a very low inclination and 10 indicating a very high inclination to use MT.

- interest in post-editing, with 1 indicating a very low interest and 10 indicating a very high interest in the use of MT.

- boredom of post-editing tasks, with 1 indicating that post-editing tasks are not boring at all and 10 indicating they are very boring.

- cognitive effort involved in post-editing tasks, with 1 indicating a very low cognitive effort and 10 indicating a very high cognitive effort involved in post-editing tasks.

- and quality of post-edited machine translated texts, with 1 indicating the post-edited MT texts are of very poor quality and 10 indicating they are very good post-edited machine translated texts.

3.2. Selecting the tool

After analysing various tools such as CASMACAT, MATECAT, PET, Translog II, Appraise, Costa MT, MT-Equal, TransCenter and CATaLog Online, TAUS DQF was chosen due to its extensive use both in academia and in the industry (Görög, 2014; Valli, 2015). It has a user-friendly interface that makes the process easier both for the researcher and the translator. TAUS DQF also generates reports of the results automatically and randomises the presentation of segments in the ranking task. However, the current version of the tool has two drawbacks, i.e. it does not allow the post-editor to have a global view of the text and it does not allow the post-editor to go back to previous segments (Moran, Saam @ Lewis, 2014).

3.3. Selecting the participants

A priori non-probabilistic purposive sampling and snowball sampling techniques were used for the recruiting of respondents (Bryman, 2012) according to the following criteria: they should be professional translators working in the language pair being researched and native speakers of the target language. They were identified through distribution lists and email contacts. In MTOOL (Spanish into German), 22 professionals participated in the experiment, but due to technical issues only socio-demographic data from 21 were recorded. Seventeen were women (81%) and the age range was between 32 and 67. Most participants (95%) had university education and they all had at least 2 years' experience in translation. 19 translators (91%) had experience in the revision of third-party texts and 11 (52%) had also worked as professional post-editors.

3.4. Test development

The test lasted 4 hours approximately and, to avoid participants' fatigue, it was divided into two sessions that participants undertook at their own convenience in

a 3-week period. First of all, participants were informed via email of the tasks to be carried out in the first session. After signing the informed consent sheet, they were asked to participate in a voluntary section in which a general questionnaire collected socio-demographic data, followed by a post-editing pre-questionnaire (see Section 3.1).

Next, they were asked to post-edit three texts (containing 50 segments each with a total of 843, 985 and 953 words) for which they were paid a fee. The order of the presentation of texts and the MT system used to translate such texts was randomised and each individual participant given instructions in a specific order. After post-editing each text, participants were requested to assess fluency, adequacy, need for post-editing, ease of post-editing, and mental effort involved in the post-editing on a 1 to 10 scale (see Section 3.1).They were also given the opportunity to add comments after each statement.Finally, they were requested to reply again to the same questions as in the pre-questionnaire on post-editing, to see how much their attitude had changed after the post-editing task.

Once they finished the first part, participants received a second e-mail with the instructions for the second session, in which they had to carry out different tasks on the same 150 segments: a fluency evaluation task, an adequacy assessment task, and a ranking task.

3.5. Automated metrics

Automated metrics were computed for the three systems. Table 1 summarises the values and shows the positive impact of domain adaptation both in SMT and NMT in contrast with a generic system, which is especially relevant taking into account the limited amount of training data when compared to generic systems. It also shows how fine-tuned NMT systems seem to perform better than the other ones. For the BLEU metric statistical significance at $p < 0.05$ was found between all pairs of systems, i.e., between NMT and GT, between NMT and SMT, and between GT and SMT.

Table 1: Objective automated metrics

	SMT	NMT	GT
BLEU	19.830	27.715	12.265
METEOR	35.260	41.471	25.668
TER	69.378	62.203	85.055

4. Discussion of results

Results from the human evaluation will be presented separately for each element assessed. A statistical analysis was carried out using IBM SPSS v.20, with a significance level of 0.05.

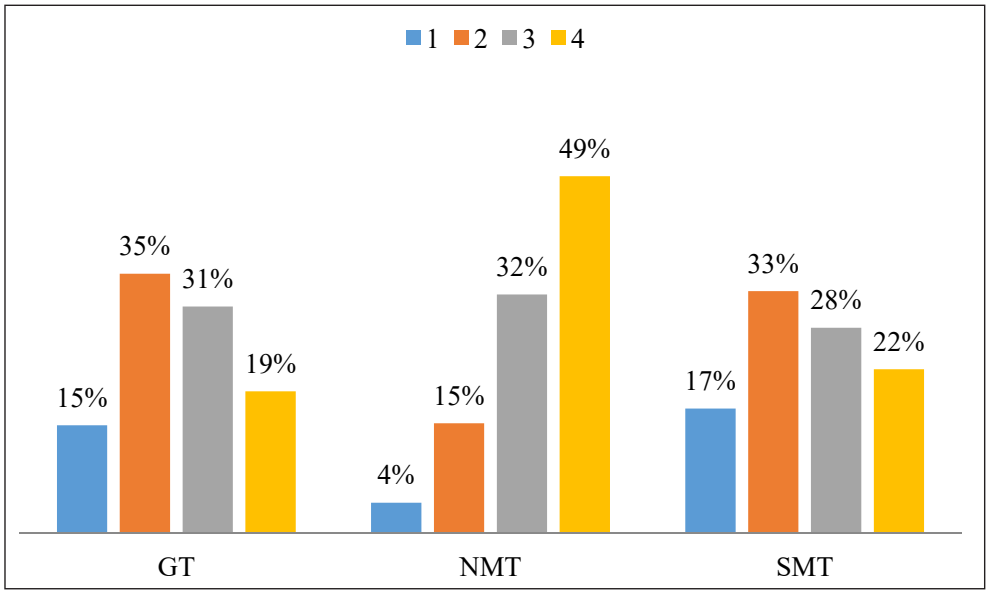
Chi-square tests (Saldanha & O'Brien, 2013) were employed to compare the distribution of qualitative data, i.e., the adequacy, fluency and ranking assessments at the segment level. Discrete quantitative data, such as quality assessments at the textual level, were analysed using a Mann-Whitney U test (Mann and Whitney, 1947) to compare groups, while for continuous quantitative data, such as PE speed and effort, the Bonferroni-corrected Mann-Whitney U test was used for multiple comparisons (Dunn, 1964). Attitude assessments were considered as paired discrete numerical variables and a Wilcoxon signed-rank test (Wilcoxon, 1945) was then applied to determine whether there was a statistically significant change in their opinions. Moreover, Spearman's correlation (Schober, Boer & Schwarte, 2018) was used to see if socio-demographic data and participants' professional experience had any influence on the different assessments.

An inter-rater reliability analysis was also performed using the quality assessment variables at the segment level through the intra-class correlation coefficient (ICC) estimates. Thus, the ranking obtained an ICC of 0.924, reaching 0.956 in the case of the adequacy and 0.979 in the case of the fluency, which are excellent levels of reliability.

4.1. Quality at the segment level

In terms of adequacy, assessed on a 1 to 4 scale, the GT and SMT systems show a similar distribution, with most segments being assessed as a 2 and a 3 (35% and 31% in GT; 33% and 28% in SMT, respectively), whilst in the NMT system participants rate most segments with higher marks, 3 (32%) and 4 (49%), as shown in Graph 1.

Graph 1: Adequacy metrics



The median for all three systems is 3, but the mode for the NMT system is 4, much higher than for the GT and SMT systems (mode = 2). The same differences are found when mean rates are compared, as the NMT system shows a mean of 3.25 (*SD* = 0.86) and the GT and SMT systems show similar lower mean values (M = 2.55, *SD* = 0.96 and M = 2.56, *SD* = 1.02 respectively), as Table 2 shows:

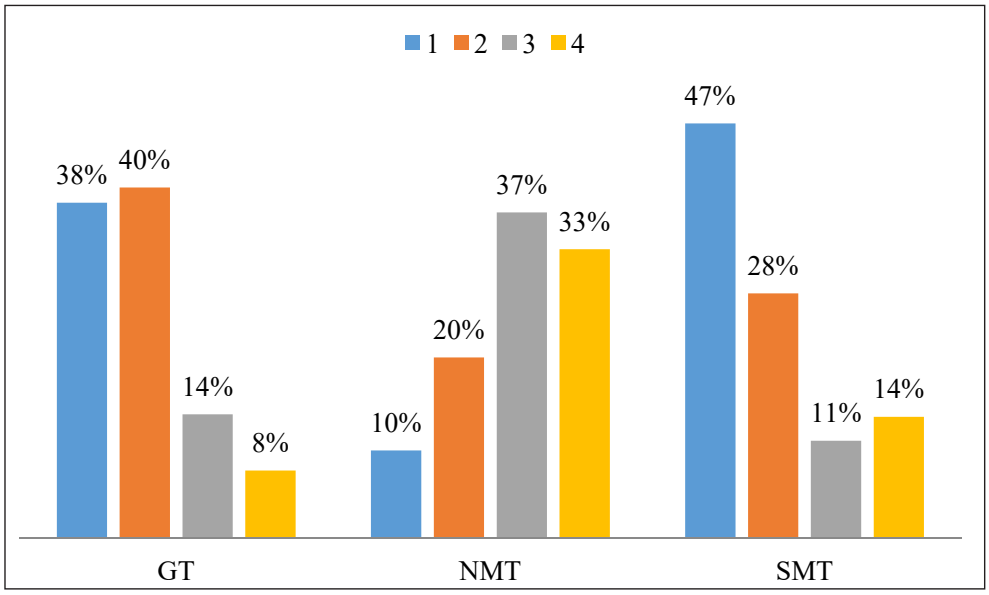
Table 2: Descriptive statistics for adequacy

	NMT	GT	SMT
Mean	3.25	2.55	2.56
Median	3	3	3
Mode	4	2	2

Differences between GT and SMT are not statistically significant, but they are between GT and NMT ($X^2(3) = 296.28$, $p < 0.001$; $Z = 510.5$, $p < 0.05$) and between NMT and SMT ($X^2(3) = 271.48$, $p < 0.001$; $Z = 1,896$, $p < 0.05$), hence proving that the NMT system is the best rated in terms of adequacy.

In terms of fluency, in the GT and SMT systems more than 75% of the segments are rated with low values, as shown in Graph 2. On the contrary, the NMT system only has 30% of segments in this low range, showing an entirely different pattern.

Graph 2: Fluency metrics



All descriptive values, included in Table 3, show better values for the NMT system, followed by the GT and SMT systems. The mean value for the NMT system (M = 2.92, SD = 0.96) is higher than that of the GT and SMT systems, although the difference between the GT system (M = 1.906, SD = 0.91) and the SMT system (M = 1.909, SD = 1.06) is almost non-existent.

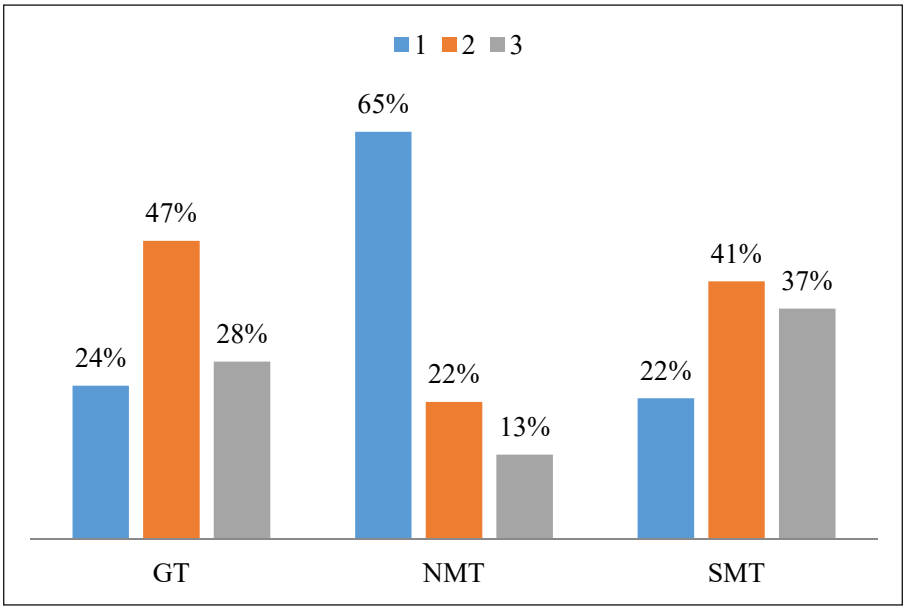
Table 3: Descriptive statistics for fluency

	NMT	GT	SMT
Mean	2.92	1.91	1.91
Median	3	2	2
Mode	3	2	1

Chi-square tests show that there are statistically significant changes in fluency between the three systems: SMT vs. NMT ($X^2(3) = 528.32$, $p < 0.001$), SMT vs. GT

($X^2(3) = 55.38$, $p < 0.001$), NMT vs. GT ($X^2(3) = 541.15$, $p < 0.001$).As far as the ranking task is concerned, the NMT system is selected as the best-rated translation engine in 39.7% of cases, whilst GT is selected in 31% of segments and the SMT system in 29.3%. When looking at the segments selected in the first place, one can see that 24% is linked to GT, 22% to the SMT system and 65% to the NMT system, as shown in Graph 3.

Graph 3: Ranking results



As regards mean values, the NMT system obtains lower values, which in this case show a better assessment as rank 1 represents the best quality and rank 3, the worst: NMT ($M = 1.49$, $SD = 0.72$), GT ($M = 2.04$, $SD = 0.72$) and SMT ($M = 2.14$, $SD = 0.76$).

Table 4: Descriptive statistics for ranking

	NMT	GT	SMT
Mean	1.49	2.04	2.14
Median	1	2	2
Mode	1	2	2

Chi-square tests show that there are statistically significant changes between the three systems regarding their ranking, as shown in Table 5.

Table 5: Ranking results

	Chi-square	Wilcoxon
GT vs. SMT	$X^2(2) = 1{,}583.98$, $p < 0.001$	$Z = 8{,}850$, $p < 0.05$
GT vs. NMT	$X^2(2) = 353.75$, $p < 0.001$	$Z = 18{,}972$, $p < 0.05$
NMT vs. SMT	$X^2(2) = 714.83$, $p < 0.001$	$Z = 3{,}181$, $p < 0.05$

4.2. Quality at the document level

Only 12 replies were recorded for the GT and NMT systems and 13 replies were recorded for the SMT system. Few qualitative comments were added by participants, whose contribution to this task was considered voluntary due to the constraints imposed by the ethical committee.

As regards fluency, mean values show that the NMT system obtains the best ratings ($M = 5$, $SD = 1.65$) in contrast to GT ($M = 3.33$, $SD = 1.15$) and SMT ($M = 3.31$, $SD = 2.17$). In this instance, only the differences between GT and NMT are statistically significant ($U = 33.00$, $p < 0.05$) and between NMT and SMT ($U = 34.50$, $p < 0.05$). Participants provided some contradictory comments concerning the NMT such as "[a]stonishingly, several segments were translated perfectly (10), but then with other segments [*sic*] content was missing or the sentence was not understandable" (P1) or "[n]ot at all fluent, although a few sentences were okay" (P3). Nevertheless, these comments were more positive than those received for the other systems, where participants indicated that "[e]ven simple German sentence structure was not correct, partly the text contained completely untranslated terms" (P3) or "[t]he source quality is not good" (P6). In any case, none of the mean values is high, 5 being the best of all three.

In terms of adequacy, the NMT system again obtains a higher mean value ($M = 6.92$, $SD = 1.44$) compared to the GT ($M = 4.83$, $SD = 2.21$) or the SMT system ($M = 5.54$, $SD = 2.54$). However, differences are only statistically significant when comparing GT and NMT ($U = 30.50$, $p < 0.05$). Participants indicated that in the NMT system "[t]he programm [*sic*] has problems with longer, convoluted sentences and skips parts" (P16) and in the SMT system "[o]ften words in the source language remained, parts of sentences weren't translated at all" (P1).

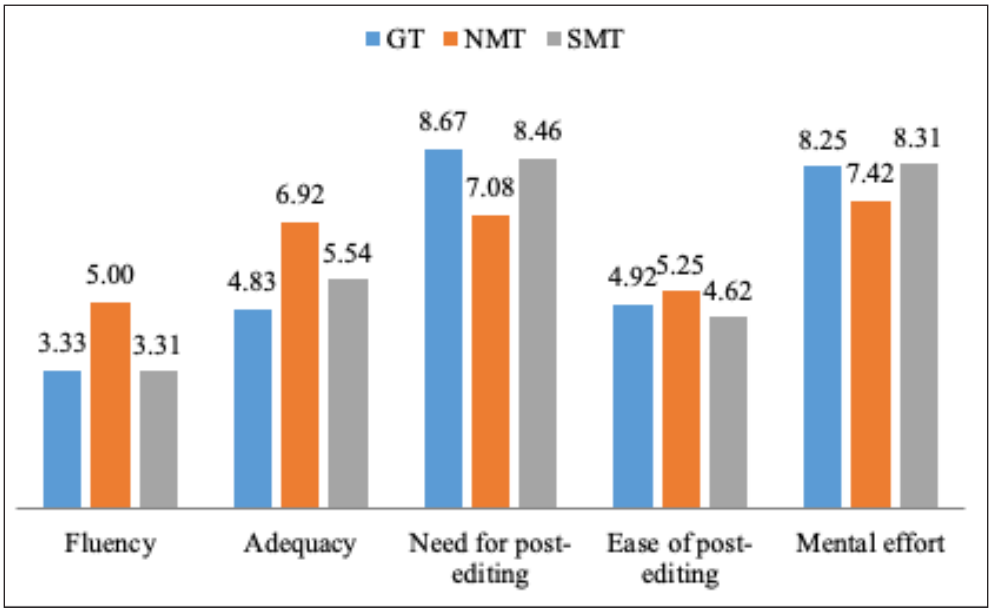
With respect to the need for post-editing, the NMT system (M = 7.08, *SD* = 1.98) is considered to require less post-editing compared to GT (M = 8.67, *SD* = 0.78) and SMT (M = 8.46, *SD* = 1.45). These descriptive results are partially confirmed, since the only statistically significant differences are found between the GT and NMT systems ($U = 33.00$, $p < 0.05$). One participant considered that in the SMT system "[n]early all the segments had to be post-edited" (P20) and another one indicated that in GT "[i]ndividual segments were very well translated, but all in all there was a lot of work" (P1). Regarding the NMT system, comments reflect opposing views: "[i]t required a lot of post editing. The MT text was of low quality and not reliable" (P3) versus "some segments didn't require any editing, others were complicated" (P1).

In relation to the ease of post-editing, mean values for all three systems are quite similar and are not statistically significant: NMT (M = 5.25, *SD* = 2.67), GT (M = 4.92, *SD* = 2.97), and SMT (M = 4.62, *SD* = 3.07). Qualitative data show that the difficulty was the low accuracy of the terminology due to the high degree of specialisation of the texts, which was highly dependent on the knowledge of each participant of the domain. The experimental conditions linked to the chosen tool also had an impact, as participants could not go back to a previous segment and correct it, as mentioned above. In this regard, one participant working with GT indicated that "[i]t was not easy. I had some terminology issues. In a real translation, I would have gone back at the end to change some important terms that I got wrong to make sure they are translated correctly and consistently throughout the file. The result as it is now is awful and would definitely require further editing" (P3). Another participant made the following comment concerning the NMT system: "[t]erminology and context presented the biggest challenge, given that it was impossible to access previously edited segments" (P6). Also, when dealing with the SMT output, similar comments are found: "[i]t was often difficult to understand the whole sentence at once" (P20). It is interesting to notice that one participant indicated that post-editing the NMT system was "[w]ay easier than G3. But the translation still needed some work" (P16). G3 was a GT output.

With regard to mental effort, participants considered that post-editing the NMT output (M = 7.41, *SD* = 2.19) required less effort than post-editing GT (M = 8.25, *SD* = 2.18) and SMT outputs (M = 8.31, *SD* = 1.89), although the differences are not statistically significant. Qualitative data show that the effort was mostly related to the degree of specialisation of the text, as indicated by participant 12 in the following comment in which s/he states that the challenge is "mostly to find out about specialised vocabulary".

Graph 4 presents a summary of mean values for the measures discussed until this point.

Graph 4: Mean values for quality at document level



4.3. Productivity

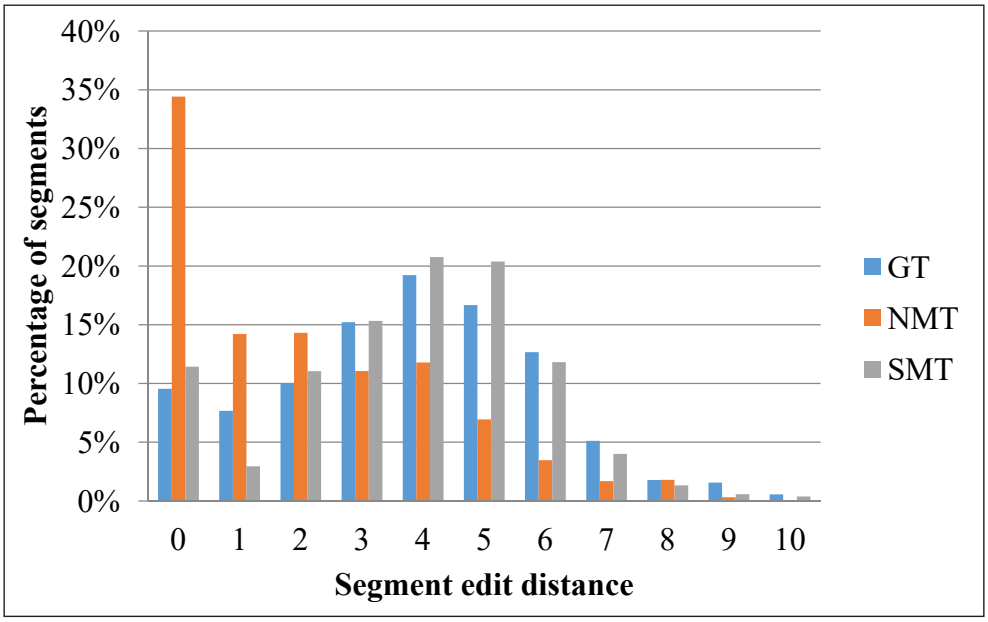
Data from two participants could not be used, as they had not carried out the task following the instructions provided. In terms of post-editing speed, NMT texts were translated at a speed of 1,207 words per hour (*SD* = 630.81), compared to 1,018 words for GT output (*SD* = 500.40) and 996 words per hour for SMT output (*SD* = 483.26). However, a Mann-Whitney U test does not find any statistical differences among the three systems in any case: SMT vs. NMT (U = 160.00 p > 0.05), SMT vs. GT (U = 174.00, p > 0.05), NMT vs. GT (U = 152.00, p > 0.05).

When looking for correlations with the participants' experience as translators, revisers or post-editors –computed through a Spearman's correlation in which the numeric value of the years of experience has been used–, there seems to be no correlation between PE speed and their experience. There is also no correlation between PE speed and PE effort. However, as regards quality assessments at the document level, there seems to be a positive correlation between PE speed and fluency (r_s = -0.59, p < 0.05) and ease of PE (r_s = 0.39, p > 0.05) assessments at the document level in the case of the NMT system, and paradoxically a negative correlation between PE speed and ease of PE (r_s = -0.42, p > 0.05) at the document level in the case of the SMT system.

In terms of post-editing effort, a big difference is observed for 0 edit distance: 34.42% of the segments of the NMT system are included in this range, whilst the

percentage is much lower for the GT (9.56%) and the SMT systems (11.43%). Moreover, the NMT system performs better than GT and SMT in edit distances 1 and 2 (14.21% and 14.32% respectively). Thus, more than 60% of its segments fall in the lowest edit distances, as compared to 27% for GT and 25% for the SMT system. All these values seem to indicate that the NMT is the system that requires less post-editing effort in our experiment. Graph 5 summarises the results.

Graph 5: Edit distances in segments



Mean values on post-editing effort prove the observations in the edit distance distribution: NMT presents a mean of 20.49 (*SD* = 20.82), which is much lower than those of GT (*M* = 37.56, *SD* = 21.51) and SMT (*M* = 37.14, *SD* = 20.21). The differences in terms of PE effort are statistically significant when comparing all systems, as described next: SMT vs. NMT ($U = 266.43$, $p < 0.001$), SMT vs. GT ($U = 479.25$, $p < 0.001$), NMT vs. GT ($U = -212.81$, $p < 0.001$).

Again, when correlating PE effort with the participants' professional experience, there only seems to be a positive correlation between PE effort and their experience as revisers ($r_s = 0.13$, $p < 0.001$). Despite the fact that the correlation is statistically significant, the strength of association is very low.

When correlating PE effort with quality assessments at the document level, statistically significant correlations with a moderate-to-low strength of association are

found for all assessments in the case of the NMT systems. Thus, for higher PE effort values, there are lower values in fluency (r_s = 0.10, p < 0.05), adequacy (r_s = -0.34, p < 0.001) and ease of PE (r_s = -0.21, p < 0.001), but higher values in the need for PE (r_s = 0.25, p < 0.001) and mental effort (r_s = 0.18, p < 0.001). In the case of the SMT and GT systems, a negative correlation exists between PE effort and adequacy (r_s = -0.34, p < 0.001 and r_s = -0.18, p < 0.001 respectively).

4.4. Attitudes

Participants' attitude towards several aspects of MT and PE are not particularly positive, contrary to Cadwell et al.'s (2016) findings in relation to institutional translators from the European Commission's Directorate-General for Translation (DGT). As shown in Table 6, in relation to MT, even though they do not seem to be inclined to use it and think MT texts are of low quality, they paradoxically perceive it to be reasonably useful. As far as PE is concerned, they regard it as a relatively boring task which requires a high cognitive effort, although their interest in PE can be considered fair and their perception of post-edited text quality is actually quite high.

Table 6: Previous attitudes on MT and PE

	Mean	Median	SD
MT quality	2.82	3	1.08
MT usefulness	5	5	2.19
MT use inclination	3.18	2	2.56
Interest in PE	4.55	5	2.42
Boredom associated with PE	4.91	5	2.74
PE cognitive effort	7.55	8	2.07
Quality of PE texts	6.73	7	1.68

A closer analysis of previous attitudes by age (Table 7) shows some interesting results: translators seem to have more positive attitudes towards MT quality and usefulness and to be more inclined to use MT as they grow older. They also seem to be less interested in PE and consider PE more boring as their age increases. However, in terms of their attitude towards PE cognitive effort and the quality of PE texts, there seems to be no age-related pattern.

Table 7: Attitudes according to age

	32-36 years old (4)			37-42 years old (6)			Over 43 years old (11)		
	Mean	Median	SD	Mean	Median	SD	Mean	Median	SD
MT quality	2.00	2	1.41	2.50	2.5	0.58	3.40	3	1.14
MT usefulness	4.00	4	1.41	4.00	4.5	1.41	6.20	7	2.59
MT use inclination	1.00	1	0.00	2.00	2	0.82	5.00	5	2.83
Interest in PE	6.50	6.5	2.12	4.25	4.5	0.96	4.00	2	3.24
Boredom associated with PE	3.00	3	2.83	5.25	5	2.06	5.40	3	3.36
PE cognitive effort	7.50	7.5	3.54	7.25	7	2.22	7.80	8	1.92
Quality of PE texts	6.00	6	1.41	7.00	7.5	1.41	6.80	6	2.17

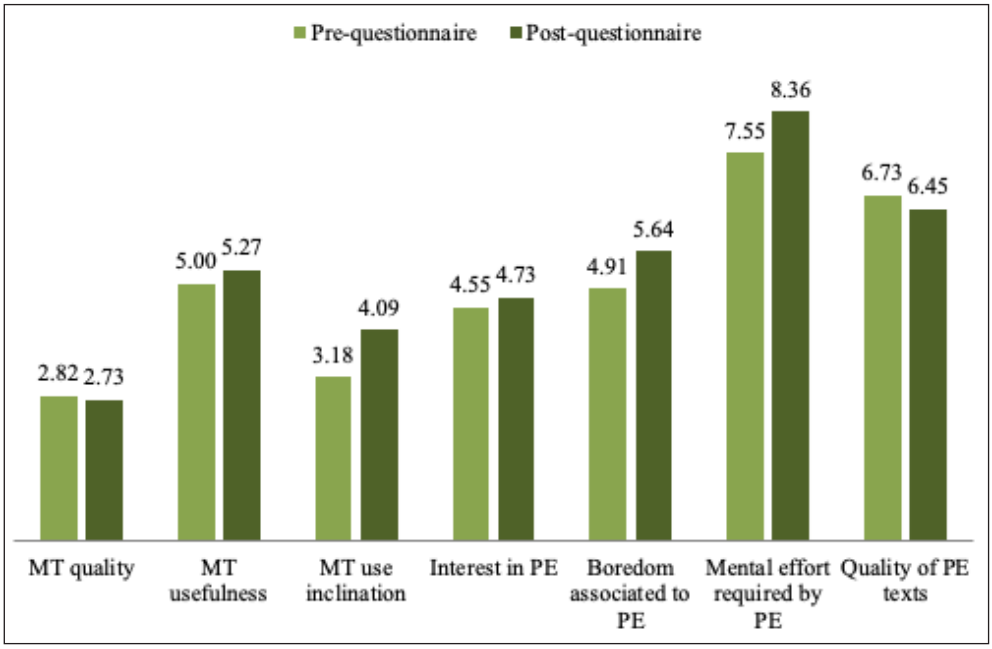
Previous attitudes seem to be related to translators' experience. Although correlations found between different aspects are not statistically significant, the strength of association is moderate. Thus, there is a positive correlation between the attitude towards MT quality and their experience as translators (r_s = 0.46, p > 0.05), as revisers (r_s = 0.34, p > 0.05) and as post-editors (r_s = 0.41, p > 0.05). However, only the experience as revisers presents a moderately positive correlation with the attitude towards MT usefulness (r_s = 0.46, p > 0.05) and towards the inclination to use MT (r_s = 0.39, p > 0.05). Their interest in PE is linked to their experience as post-editors (r_s = 0.35, p > 0.05), while boredom associated with PE has more to do with the experience as revisers (r_s = 0.36, p > 0.05). Nevertheless, it must be noted that the more experience they have as post-editors, the less boring they seem to consider PE (r_s = -0.76, p < 0.01) and the harder they tend to consider it in terms of PE cognitive effort (r_s = 0.37, p > 0.05). Lastly, their attitude toward the quality of post-edited texts is positively correlated with their experience as translators (r_s = 0.50, p > 0.05).

When correlating previous attitudes with the quality assessments at the document level, there are some statistically significant correlations. Thus, it is found that a positive attitude towards the quality of PE texts is correlated with a more positive assessment of the adequacy of texts (r_s = 0.37, p < 0.05). Also, the higher the PE cognitive effort

is thought to be, the lower the ease of PE (r_s = -0.41, p < 0.05) and the higher the perceived PE mental effort (r_s = 0.39, p < 0.05) assessments are. There is also a positive correlation between the mental effort assessment and both the attitudes towards MT usefulness (r_s = 0.361, p < 0.05) and towards the quality of post-edited texts (r_s = 0.46, p < 0.01).

As regards the change of attitudes, participants who did not fill in the questionnaire both before and after the task were not taken into account. A total of 11 replies were collected, which are summarised in Graph 6, where the changes in mean values from the pre-questionnaire to the post-questionnaire are presented.

Graph 6: Changes in attitude: mean values



In most aspects assessed there is a negative change in participants' attitudes. A positive change can only be found when assessing the usefulness of machine translation for translators (from M = 5.00 *SD* = 2.19 to M = 5.27, *SD* = 2.00), the inclination to use machine translated texts as texts as a starting point (from M = 3.18, *SD* = 2.56 to M = 4.09, *SD* = 2.30) and the interest in post-editing (from M = 4.55, *SD* = 2.42 to M = 4.73, *SD* = 2.41). It is worth highlighting that most aspects are assessed with very low rates: the quality of machine translation, which is rated with a 2.82 (*SD* = 1.08), drops to 2.72 (*SD* = 1.35). Post-editing is considered to be more boring (from M = 4.91, *SD*

= 2.74 to *M* = 5.64, *SD* = 3.41) and requiring a higher degree of mental effort (from *M* = 7.55, *SD* = 2.07 to *M* = 8.36, *SD* = 1.69) after carrying out the tasks. Similarly, the quality that professional translators assign to post-edited texts decreases slightly from 6.73 (*SD* = 1.68) to 6.45 (*SD* = 1.63). Statistically significant differences are only found for the mental effort ($Z = -1.84$, $p < 0.05$).

Changes in attitude towards any of the aspects do not appear to be related to any age group in particular. They also do not seem to be related to the years of experience translators have in the translation field or with the fact they have experience revising third-party texts. Previous experience in PE seems to be the only aspect having a statistically significant impact on the positive change in attitude as far as the inclination to use MT is concerned ($Z = -2.03$, $p < 0.05$).

5. Conclusions and summary of results

Results of the comparison of three machine translation systems in the machine-tool domain for the Spanish-German language pair show that the NMT system is ranked highest on all assessed aspects, while GT ranks second in 6 out of the 10 assessed aspects and SMT ranks second in just 4 out of the 10. Table 8 shows all the elements assessed. It indicates that the MT system performs better on a 1 to 3 ranking column and then whether statistically significant differences were found when comparing systems. For attitudes, the symbols used indicate whether a positive or a negative change was found before and after the test, and whether it was statistically significant.

Table 8: Summary of results

MTOOL	**Ranking**			**Significant differences**		
Segment-level Quality	**GT**	**NMT**	**SMT**	**GT vs. NMT**	**GT vs. SMT**	**NMT vs. SMT**
Adequacy	3	1	2	✔	X	✔
Fluency	2	1	3	✔	X	✔
Ranking	2	1	3	✔	✔	✔
Text-level Quality						
Fluency	2	1	3	✔	X	✔
Adequacy	3	1	2	✔	X	X
Need for PE	3	1	2	✔	X	X
Ease of PE	2	1	3	X	X	X
PE mental effort	2	1	3	X	X	X
Productivity						
PE speed	2	1	3	X	X	X
PE effort	3	1	2	✔	✔	✔
Attitude	**Change**					
MT quality	-			X		
MT usefulness	+			X		
MT use inclination	+			X		
Interest in PE	+			X		
PE boredom	+			X		
PE cognitive effort	+			✔		
Quality of PE texts	-			X		

When analysing both automated metrics and human results, it must be pointed out that there is an almost perfect match between the results obtained automatically

and those obtained from the human assessments. NMT is the system which is unanimously awarded the best results. GT and SMT, however, vie for the last position: while GT ranks second in 6 out of the 10 human assessments, it ranks third in the three automated metrics.

Table 9 shows the existing correlations between BLEU and different human assessments. Although the results might not be statistically significant, the strength of the association is what is relevant here (between -1 and +1, indicating negative or positive associations respectively):

Table 9: Correlations between BLEU and human assessments

Adequacy and BLEU	r_s = -0.09, $p > 0.05$
Fluency and BLEU	r_s = 0.48, $p < 0.001$
Ranking and BLEU	r_s = -0.30, $p < 0.001$
PE speed and BLEU	r_s = 0.05, $p > 0.05$
PE effort and BLEU	r_s = -0.50, $p < 0.001$

Leaving aside adequacy and PE speed, where the correlations with BLEU are very low, significant positive correlations are found between fluency and BLEU, and negative correlations are found between the other aspects, so that the greater the fluency, the lower the ranking (hence, a better result) and the lower the PE effort, the higher the BLEU metric, which highlights the existing correlations between PE effort and BLEU.

In view of these findings, we may conclude that the NMT system works much better than both the GT and the SMT systems in a highly technical, specialised domain such as that of machine-tools despite the low amount of both in-domain (25,256) and out-of-domain (1,784,385) training data, which is in line with the findings of other researchers such as Castilho et al. (2017b) in the educational domain, of Wu et al. (2016) for Wikipedia segments, of Bentivogli et al. (2016) for transcribed speeches and of Klubička et al. (2017) in the news field.

As regards subjective opinions and attitudes, some interesting results have been obtained that leave the door open for further research: participants indicate that the quality is highly variable depending on the segments, ranging from perfect translations to unacceptable ones, which shows the potential for automatic quality assessment prior to post-editing to reduce effort and negative assessments. However, their general

attitude towards machine translation and post-editing is not a positive one and does not improve much after the experience. Two aspects that may have had an effect are the experimental design, which did not allow them to go back into the text and make corrections, and the fact that they were dealing with texts of varied quality. Our research has also shown some correlations (or lack of correlations) by age and years of experience in different fields (i.e. translation, revision and post-editing), a field worth exploring in future research in which the assessment of machine translation will hopefully not just rely on automated metrics but also on human factors. In this regard, it is worth the potential in future research of other methodological tools such as focus groups or interviews with different user profiles to obtain more qualitative data on professional users attitudes which can be triangulated with quantitative measures. It also remains to be seen how future training in the area of post-editing will impact on professional attitudes towards this task.

Acknowledgements

This work was partially funded by the Spanish Ministry of Economy and Competitiveness via the AdapTA project (RTC-2015-3627-7). We would like to thank MondragonLingua Translation & Communication as the coordinator of the project and the translators who participated in the experiments. Anna Matamala is a member of Transmedia Catalonia, a research group funded by the Catalan government under SGR call (2017SGR113).

6. References

Axelrod, A., He, X., & Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '11)*, 355-362.

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. *Proceedings of the International Conference on Learning Representations.* CoRR abs/1409.0473.

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 29, 65-72.

Bentivogli, L., Bisazza, A., Cettolo, M. & Federico, M. (2016). Neural versus Phrase-Based Machine Translation Quality: a Case Study. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 257–267.

Bisazza, A., Ruiz, N., & Federico, M. (2011). Fill-up versus interpolation methods for phrase-based SMT adaptation. *International Workshop on Spoken Language Translation (IWSLT 2011)*, 136-143.

Bryman, A. (2012). *Social Research Methods*. Oxford: Oxford University Press.

Cadwell, P., Castilho, S., O'Brien, S., & Mitchell, L. (2016). Human factors in machine translation and post-editing among institutional translators. *Translation Spaces*, 5, 222-243.

Callison-Burch, C., Koehn, P., Monz, C., Post, M., Soricut, R., & Specia, L. (2012). Findings of the 2012 Workshop on Statistical Machine Translation. *Proceedings of the 7th Workshop on Statistical Machine Translation*, 10-51.

Castilho, S., Moorkens, J., Gaspari, F., Calixto, I., Tinsley, J., & Way, A. (2017a). Is neural machine translation the new state of the art? *The Prague Bulletin of Mathematical Linguistics*, 108(1), 109-120.

Castilho, S., Moorkens, J., Gaspari, F., Sennrich, R., Sosoni, V., Georgakopoulou, P., Lohar, P., Way, A., Barone, A.V.M., & Gialama, M. (2017b). A comparative quality evaluation of PBSMT and NMT using professional translators. *Proceedings of MT Summit XVI*, 1, 116-131.

Cho, K., van Merrienboer, B., Bahdanau, D., & Bengio, Y. (2014). On the properties of neural machine translation: Encoder–Decoder approaches. *Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, 103-111.

Crego, J., Kim, J., Klein, G., Rebollo, A., Yang, K., Senellart, J., Akhanov, E., Brunelle, P., Coquard, A., Deng, Y., Enoue, S., Geiss, C., Johanson, J., Khalsa, A., Khiari, R., Ko, B., Kobus, C., Lorieux, J., Martins, L., Nguyen, D., Priori, A., Riccardi, T., Segal, N., Servan, C., Tiquet, C., Wang, B., Yang, J., Zhang, D., Zhou, J., & Zoldan, P. (2016). *Systran's pure neural machine translation systems*. CoRR abs/1610.05540.pdf.

Dunn, O. J. (1964). Multiple comparisons using rank sums. *Technometrics*, 6, 241–252.

Eetemadi, S., Lewis, W., Toutanova, K., & Radha, H. (2015). Survey of data-selection methods in statistical machine translation. *Machine Translation*, 29, 189-223.

Etchegoyhen, T., Fernández-Torné, A., Azpeitia, A., Martínez García, E., & Matamala, A. (2018). Evaluating Domain Adaptation in Machine Translation Across Scenarios. *Proceedings of the Eleventh International Conference on Language Resources Evaluation (LREC 2018)*, 6-15.

Foster, G., & Kuhn, R. (2007). Mixture-model adaptation for SMT. *Proceedings of the Second Workshop on Statistical Machine Translation (StatMT '07)*, 128-135.

Freitag, M., & Al-Onaizan, Y. (2016). *Fast Domain Adaptation for Neural Machine Translation.* CoRR abs/1612.06897.pdf.

Gascó, G., Rocha, M. A., Sanchis-Trilles, G., Andrés-Ferrer, J., & Casacuberta, F. (2012). Does more data always yield better translations? *Proceedings of the 13th European Chapter of the Association for Computational Linguistics*, 152-161.

Görög, A. (2014). Quantifying and benchmarking quality: the TAUS Dynamic Quality Framework. *Tradumàtica*, 12, 443-454.

Heafield, K. (2011). Kenlm: Faster and smaller language model queries. *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT '11)*, 187-197.

Klein, G., Kim, Y., Deng, Y., Senellart, J., & Rush, A.M. (2017). Opennmt: Open-source toolkit for neural machine translation. *Proceedings of the 20th Annual Conference of the European Association for Machine Translation*, 67-72.

Klubička, F., Toral, A., & Sánchez-Cartagena, V.M. (2017). Fine-grained human evaluation of neural versus phrase-based machine translation. *The Prague Bulletin of Mathematical Linguistics*, 108, 121-132.

Koehn, P., & Monz, C. (2006). Manual and automatic evaluation of machine translation between European languages. *Proceedings of the Workshop on Statistical Machine Translation*, 102-121.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R. et al. (2007). Moses: Open source toolkit for statistical machine translation. *Proceedings of the 45th Annual Meeting of the ACL*, 177-180.

Koponen, M. (2010). Assessing machine translation quality with error analysis. *Electronic proceedings of the KäTu symposium on translation and interpreting studies*, 4, 1-12.

Luong, M.T., & Manning, C.D. (2015). Stanford neural machine translation systems for spoken language domains. *Proceedings of the International Workshop on Spoken Language Translation*, 76-79.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, 18, 50–60.

Moran, J., Saam, C., & Lewis, D. (2014). Towards desktop-based CAT tool instrumentation. *Proceedings of the Third Workshop on Post-Editing Technology and Practice (WPTP-3)*, 99-112.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.J. (2002). BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, 311-318.

Schober, P., Boer, C., & Schwarte, L. (2018). Correlation Coefficients: Appropriate Use and Interpretation. *Anesthesia & Analgesia*, 126 (5), 1763-1768.

Sennrich, R. (2012). Perplexity minimization for translation model domain adaptation in statistical machine translation. *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 539-549

Snover, M., Dorr, B., Schwartz, R., Micciulla, L., & Makhoul, J. (2006). A study of translation edit rate with targeted human annotation. *Proceedings of Association for Machine Translation in the Americas*, 223-231.

TAUS. (nd.). *Taus dynamic quality framework: Getting started*. Technical report.

Valli, P. (2015). The TAUS Quality Dashboard. *Proceedings of the 37th Conference Translating and the Computer*, 127–136.

Van Peer, W., Hakemulder, F., & Zyngier, S. (2012). *Scientific methods for the humanities*. Amsterdam: John Benjamins Publishing Company.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1, 80–83.

Wu, Y., Schuster, M., Chen, Z., Le, Q.V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, L., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., & Dean, J. (2016). *Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. CoRR abs/1609.08144.

"When being specific is not enough": Discrepancies between L2 learners' perception of definiteness and its linguistic definition —

Sugene Kim
Nagoya University of Commerce & Business, Japan
sugene_kim@nucba.ac.jp

Abstract

This paper explores the sources of difficulties that second language (L2) learners encounter when using English articles. Eighty-four Korean college students completed a forced-choice elicitation task before and after receiving instruction on article use and provided written accounts of article choices. The analysis of the task performance and written accounts indicated the participants' noticeable tendency to prioritize specificity over definiteness, resulting in the overuse of *the* with specific indefinites. Not infrequently, the participants estimated a "nonspecificity hierarchy" for nonspecific definites, often leading to the infelicitous use of *a(n)*. The overuse of *the* with modified noun phrases suggests that L2 learners attempt to construe semantic context (i.e., ±definite) on the basis of the syntactic structure. Furthermore, the participants' correct use of *a(n)* for singular count indefinites sometimes stemmed from assuming the number of a target noun to be single rather than considering its multiple existence and, thus, its indefinite nature. These findings underline the necessity of teaching the specificity feature to indicate to learners that (1) English articles are prototypical realizations of encoding definiteness, which requires the mutual identifiability of a unique referent, and (2) specificity, which presupposes identifiability assumed by the writer/speaker alone, is not marked by articles in English.

Keywords: English articles, specificity, definiteness, article errors, L2 learners

Resumen

Este trabajo explora las dificultades que los estudiantes de una segunda lengua (L2) encuentran en el uso de los artículos en inglés. Un total de ochenta y cuatro

estudiantes universitarios coreanos completaron una tarea escrita de elección forzada antes y después de recibir clases sobre el uso del artículo, y proporcionaron informes con las razones de su elección de las distintas formas de los artículos. El análisis de los resultados de la tarea y de los informes escritos mostró una tendencia notable de los participantes a priorizar la especificidad sobre la definitud, de lo que resultó el uso excesivo de "the" con indefinidos específicos. No pocas veces tuvieron en cuenta una "jerarquía de no especificidad" para definidos inespecíficos, que con frecuencia determina el uso infeliz de "a(n)." El uso excesivo de "the" con sintagmas de sustantivos modificados sugirió que los estudiantes de L2 intentan interpretar el contexto semántico (es decir, ±definido) sobre la base de la estructura sintáctica. Además, el uso correcto de los participantes de "a(n)" para indefinidos de recuento singular se produjo en ocasiones por suponer que el número de un sustantivo objetivo es único en lugar de considerar su existencia múltiple y, por lo tanto, su naturaleza indefinida. Estos resultados ponen de manifiesto la necesidad de que en las clases se llame la atención sobre la característica de especificidad para indicar a los alumnos que (1) los artículos son realizaciones prototípicas de la codificación de definitud, que requiere la identificación mutua de un referente único, y que (2) la especificidad, cuya identificación es asumida sólo por el escritor/hablante, no es marcada por los artículos en inglés.

Palabras clave: artículos en inglés, especificidad, definitude, errores de artículo, estudiantes de L2

1. Introduction

Extensive empirical evidence indicates that correct article use is difficult for second language (L2) learners to master, especially for those whose mother tongue (L1) does not have functional equivalents (e.g. Korean, Japanese, Russian, and Turkish) (Butler, 2002; Ionin, 2006; Ionin et al., 2004, 2008; Kim & Lakshmanan, 2009; Ko et al., 2010; Master, 1997). Research on L2 writing, for instance, has noted that inaccurate article use is one of the most frequent errors (Bitchener et al., 2005) and that even highly advanced student writers often fail to exhibit native-like article use (Lennon, 1991; Spada & Tomita, 2010). L2 learners' infelicitous article use is undoubtedly attributed to the rules governing the system, which are notoriously unwieldy (Dulay et al., 1982) and lack one-to-one correspondence between an article and its semantics—i.e., whether the noun phrase is (in)definite, (non)specific, or generic in reference (Kachru, 2010).

In written discourse, articles create an understanding between the writer and reader, and their misuse easily causes ambiguity because different articles lead to different interpretations. As Halliday & Hasan (1976:74) noted, "Whenever the

information is contained in the text, the presence of an article creates a link between the sentence in which it occurs and that containing the referential information." Given that article use cannot be avoided irrespective of the genre and that English article errors continue to plague the compositions of L2 writers, the importance of identifying features that influence article misuse and providing proper pedagogical solutions cannot be overstated.

In the almost three decades since Master (1990:461) suggested the necessity of providing "a coherent grammar for teaching the articles as a system," such attempts have been few and far between. Of course, the relative dearth of research might be attributable to the trend in which most language education studies have moved beyond methods that emphasize discrete grammar teaching, instead prioritizing lesson content and contextualization for students (Richards & Rodgers, 2014). However, learners need to be equipped with language awareness—or "explicit knowledge about language and conscious perception [of] . . . language use" (Garrett & James, 2000:330)—so that they can use such knowledge independently while engaging in the encoding process. Regardless of one's position on the usefulness of focus on form or forms, it is indisputable that both teachers and learners look for help in pedagogical grammar or reference books that address the article system in a manner that clarifies how meaning is mapped onto forms (Young, 1996).

2. Pedagogical frameworks for teaching the English article system

Drawing upon McEldowney's (1977) suggestion to produce a simplified framework for teaching English articles, Master (1990) introduced a binary system in which four features of article use—definiteness, specificity, countability, and number—are reduced to a meaning contrast between "classification" (marked by the indefinite or zero article) and "identification" (marked by the definite article). The system obscures the distinction between generic and specific uses of a noun, which basically requires the same encoding, and collapses them into a single feature—identification. Although the notion of specificity is useful in establishing a discourse referent, it was set aside "as a red herring" in explaining the article system (Langacker, 1991:104), presumably because English does not mark specificity but selects articles on the basis of definiteness.

Master (1997) suggested another pedagogical framework for students at the beginning, intermediate, and advanced levels of proficiency. Following Little (1994), he proposed that in the teaching of beginners, sustained attention should not be directed to article usage rules, except for teaching words that commonly take articles. When students advance to the intermediate level, cognitive methods such as Master's (1990) binary schema can be implemented, providing sufficient time for learners to practice

a single distinction at a time. When students attain an advanced level, teaching rules governing article usage is not as appropriate as allowing them to learn the articles as lexical items in context. He emphasized the importance of encouraging learners to keep a record of their own errors, enabling errors to become "an essential part of the learning process" (Lewis, 1993:6).

Master (2002) proposed yet another pedagogy based upon a canonical information structure in which new information is mentioned last (or to the right of the main verb) and marked with the indefinite or zero article, whereas given information is mentioned first (or to the left of the main verb) and marked with the definite article (Yule, 1998). After confirming that given information adheres to a canonical structure most of the time and new information does approximately half of the time, he experimented with the applicability of an information structure as a framework to teach English articles. The results showed that the experimental group, taught with the suggested pedagogy, made noticeable improvement compared to the other two groups of students who received either traditional instruction on article use or no instruction.

The effectiveness of dictionary use for making target-like article choices was tested in Miller's (2006) study with advanced L2 learners. A comparison of scores on the pre- and posttests, which consisted of gapped and non-gapped exercises, indicated that the overall increase in correct article use was larger for learners in the experimental group, who were requested to use the dictionary for the posttest. However, the experimental group outperformed the control group only in the gapped exercise, suggesting that dictionary use facilitates participants' ability to identify the correct article but not to determine whether and where an article is needed. More recently, Kim (2018) evaluated the effectiveness of a lexicographic approach to teaching the English article system using an English–Korean bilingualized dictionary—*Naver Dictionary*. She showed that dictionary consultation helps L2 learners determine nominal countability and associated article use, although the way the dictionary provides countability information is not always explicit and/or sufficiently user-friendly (e.g. inadequate labeling or the absence of nominal countability information).

Compared with the number of research articles investigating English learners' understanding of the article system (e.g., Chan, 2016), actual use of articles (e.g., Mizuno, 1999), underlying reasons for article errors (e.g., Ionin et al., 2004, 2008; Chan, 2017), or sources of difficulty (e.g., Sarker & Baek, 2017), fewer experimental studies have been conducted to delineate a pedagogical framework for teaching English articles. Amid reports attesting that L2 learners seem to use articles without a clear understanding of them (e.g., Butler, 2002), theoretical research in the domain has not influenced teaching practice on the whole (Lopez & Sabir, 2019). The present study was conceived in an attempt to link research and pedagogy by first examining whether there

is insufficiency in the way the article system is taught and then addressing the problem area(s) identified. Specifically, this study addressed the following research questions:

(1) Does focused instruction on English article use improve L2 learners' ability to use articles correctly?

(2) Are there particular types of difficulties that L2 learners encounter even after receiving focused instruction on English article use?

Drawing upon the answers to these questions, this paper suggests a systematic approach to teaching the article system in a manner that does not overwhelm learners with an immense volume of information. A quasi-experiment was conducted using a within-subjects design in which each participant is tested under the control condition first and then under the treatment condition (Price, 2012).

3. Method

3.1. Participants and research setting

The participants were a homogeneous group of 84 Korean college students at a major research university in Seoul, Republic of Korea. They were first-year female students (aged 18 to 19) taking a mandatory freshman English course. The course was designed to improve all four language skills with a primary focus on academic literacy development, encompassing the general features of college-level English reading and writing. The class met twice a week for 75 minutes each time over a 15-week semester. Judging from the scores of the placement test—the TOEFL ITP (Institutional Testing Program) test—which is administered by the school and scored by ETS (Educational Testing Service) to determine the level of English courses they should take, the participants could be collectively described as upper-intermediate to pre-advanced learners of English, corresponding to the CEFR (Common European Framework of Reference for Languages) C1 level. The language background survey results indicated that the participants had studied English for an average of 10.5 years before entering college and that none had lived in English-dominant countries.

3.2. Instruments

A 32-item forced-choice elicitation task was designed to be used as both a diagnostic test (pretest) and a posttest. Targeting the use of English articles, the task contained sentences from various sources, such as online newspaper and magazine articles, Ionin et al. (2004), and Yoo (2004). Effort was exerted to ensure that items of

varying specificity–definiteness value combinations and noun types, such as concrete and abstract nouns used in countable and uncountable forms, were included in random order because the literature has consistently identified making countability judgment errors (e.g., Tsang, 2017) and equating specificity with definiteness (e.g., Chan, 2016; Ionin et al., 2004) as major obstacles to achieving target-like article use among L2 learners. According to Brown's dichotomy (as cited in Celce-Murcia & Larsen-Freeman, 1999), specificity crucially differs from definiteness in that specificity refers to the shared knowledge from the writer/speaker's viewpoint within the knowledge base of the writer/speaker and reader/hearer, irrespective of the latter's knowledge status. Since revisions were made to the original sentences by shortening sentences, changing the sentence structure, or simplifying vocabulary, three English native-speaking professor colleagues—all of whom had a PhD in applied linguistics or English literature—were asked to evaluate the naturalness of the revised sentences and confirm the correctness of article use in these sentences. Of the 40 initially prepared test items, eight were removed because there were discrepancies among the professors regarding article use for the target noun in the given context. The finalized elicitation task is presented in Appendix 1, and the correct answers are marked in bold.

3.3. Procedure

This study employed a one-group pretest–posttest design. To estimate the current understanding of English article use, the participants were pretested in Week 1 and were required to choose the correct article for each item without using a dictionary. In Week 15, they took a posttest with the same elicitation task, with an approximately 3-month interval between the two tests to minimize possible practice effects (Bachman, 1990). One week prior to the posttest, the participants received focused instruction on article use for two consecutive sessions. In all other weeks, the curriculum included no direct grammar instruction.

During the first session, the 14-page chapter about the rules concerning English article use from *Top 20: Great Grammar for Great Writing* (Folse et al., 2008)—hereafter abbreviated as *Top 20*—was used as the instructional material. As is customary for most English grammar books written for international students, *Top 20* explains the rules based on nominal countability and definiteness, such as the use of *a(n)* to introduce or classify a singular count noun (see example 1 below), the use of either *a(n)* or *the* for a general truth regarding a singular count noun (see examples 2 and 3), and the use of *the* with specific noun references (see example 4).

(1) Jambalaya is **a** rice dish that is native to south Louisiana.

(2) **A** piano has 96 keys. (= Piano**s** have 96 keys.)

(3) **The** tiger is native to India. (= Tiger**s** are native to India.)

(4) **The** title of this course sounds interesting.

Regarding the use of no article, it does not differentiate between the zero article (Ø1) occurring with noncount and plural nouns (e.g., *water* and *cats*) and the null article (Ø2) occurring with certain singular count and proper nouns (e.g., *lunch* and *Chicago*) (Chesterman, 1991) because singular nouns preceded by Ø2 are currently interpreted as either noncount nouns or set phrases (e.g., *on edge* or *to save face*) (Master, 1997). Therefore, the distinction between Ø1 and Ø2 was not made during the instruction; instead, both were collectively referred to as the zero article (Ø), indicating that no salient article is used.

Table 1: Binary schema for English article use

Purpose (definiteness)	Countability		
	Count noun		Noncount noun
	singular	plural	
classification (–definite)	*a(n)*	Ø	Ø
definition (–definite)	*a(n), the*		
identification (+definite)	*the*	*the*	*the*

The second session took place in a computer lab. During the session, the students were introduced to the binary schema (see Table 1) onto which the English article usage rules covered in *Top 20* were simply tabulated. The rules for proper nouns and nouns in idiomatic or conventional expressions were excluded because the use of articles for these nouns is affected by factors beyond whether the noun is (un)countable or whether it takes a singular or plural form. Following Master (1990), the purposes of article use were classified as "classification" or "identification" for the indefinite or definite use of a target noun. Unsurprisingly, the schema turned out to be largely identical to that of Celce-Murcia and Larsen-Freeman (1999). The only difference was that the purpose "definition" was added to refer to a noun used for referencing an entire class as a whole, satisfying the description inherent in the noun (Lyons, 1999). (Whether and how meanings differ according to different article use in generic contexts is outside the scope of this study. See Chesterman (1991) for a detailed discussion.) Employing the binary schema, the participants worked through the exercise questions in *Top 20* as guided in-class practice. As a reference for countability, the participants used *Naver Dictionary*—the most popular bilingualized online dictionary among Korean students. Countability is indicated in *Naver Dictionary* using the codes [C] for count nouns, [U]

for uncount nouns, and [U, C] or [C, U] for nouns that can be used in both count and noncount contexts.

After the second session was completed, the participants took the posttest the following week. The posttest was administered in the computer lab so that the participants could consult the online dictionary for countability information as needed. As with the pretest, the participants chose the correct answer(s) for each item on the posttest. This time, they were additionally asked to provide full written explanations for their article choices in either English or their L1, Korean, to shed light on what specifically causes Korean learners of English to misuse articles. Before administering the posttest, the instructor demonstrated how to give a written account of article selection and provided guidelines on what to report after completing each question.

3.4. Data Analysis

To examine whether using the binary schema can effectively teach L2 learners the English article system (Research Question 1), the participants' pre- and posttests were scored by checking whether the answer given was correct, incorrect, or partially correct. When the respondents chose either *a(n)* or *the*—not both—for singular count nouns used for definition purposes, the item was scored as partially correct with a half point awarded. The posttest scores were compared with the pretest scores by a paired-samples *t*-test with a significance level set at .05. In addition, mean correct answer rates for each of the specificity–definiteness value combinations were calculated to examine whether specificity affects detection of the semantic context (i.e., ±definite).

To investigate the difficulties the participants encountered in their article use (Research Question 2), the number of responses for each option—*a(n)*, *the*, and Ø—was counted for all items. The written accounts of article choices were classified according to Butler's (2002) classification scheme, which first classifies reasons for article use as specific or nonspecific depending on whether learners "were able to identify rules of grammar or other reasons for selecting the articles they chose" (p. 458). (Nonspecific reasons such as plausible choice, elimination, and no clue will not be discussed in detail in this paper). Cases in which the participants left no written comments were categorized separately as "blank answers." Following Chan (2017), both reasons for non-target-like article use and incorrect hypotheses for target-like article use were examined. Then, the written accounts pertaining to "non-target-like article use" or "erroneous theories inadvertently leading to target-like article use" were analyzed to identify particular types of difficulties that the participants had encountered in attempting to use English articles correctly. They were coded thematically using "paradigmatic analytic procedures to produce taxonomies and categories out of the

common elements across the database" (Polkinghorne 1995:5). Comments offered in Korean were translated verbatim into English.

4. Results and discussion

4.1. Performance on the pre- and posttest

To answer the first research question of "whether focused instruction on English article use improves L2 learners' ability to use articles correctly," the pre- and posttest means were compared using the paired-samples *t*-test. The results are summarized in Table 2 (the mean pre- and posttest scores of each item are provided in Appendix 1).

Table 2: Performance comparison according to purpose and specificity

Purpose (definiteness)	Specificity	Item number	Test	*M*	*SD*	*P*
classification (–definite)	–specific	1, 2, 9, 18, 20, 24, 25, 28	Pretest	58.2%	29.76	.040*
			Posttest	63.7%	29.99	
	+specific	3, 10, 12, 15, 26, 27, 31	Pretest	40.5%	19.60	.232
			Posttest	42.0%	19.17	
		subtotal	Pretest	49.9%	27.01	.001*
			Posttest	53.6%	27.72	
definition (–definite)	–specific	4, 16, 21	Pretest	65.3%	14.98	.039*
			Posttest	99.6%	.56	
identification (+definite)	–specific	6, 23, 29, 32	Pretest	65.8%	14.07	.289
			Posttest	71.7%	18.38	
	+specific	5, 7, 8, 11, 13, 14, 17, 19, 22, 30	Pretest	89.6%	10.66	.018*
			Posttest	99.4%	1.22	
		subtotal	Pretest	82.8%	15.94	.007*
			Posttest	91.5%	15.94	
		total	Pretest	65.8%	26.82	.000*
			Posttest	74.5%	29.36	

* $p < .05$.

The overall means increased from 65.8% on the pretest to 74.5% on the posttest. The *p*-value was far less than the preselected alpha ($p < .001$), confirming that formal instruction exerts a positive effect in helping L2 learners acquire the English article system (Master, 1997). For the nouns used for definition purposes, such as Items 16 and 21, the pre- and posttest means were 65.3% and 99.6%, respectively; the mean difference was statistically meaningful ($p = .039$).

Item 16: ***A/The*** *paper clip is handy when holding several sheets of paper together.*

Item 21: *Typically, Ø dandelions bloom in both the spring and the fall.*

The fact that no respondents correctly chose both *a* and *the* in Item 16 on the pretest clearly indicates their lack of knowledge of the relevant grammar rule (Chan, 2016). However, the posttest mean increased by more than 34% and reached almost 100%, suggesting that being introduced to the descriptive rule was sufficient to enable the participants to apply it correctly.

Table 3: Performance comparison of the nouns used for identification purposes

Purpose (definiteness)	Reference type	Item number	Test	*M*	*SD*	*P*
identification (+definite)	anaphoric	5, 7, 11	Pretest	100.0%	.00	1.000
			Posttest	100.0%	.00	
	associative	8, 13, 14, 23, 29, 32	Pretest	68.5%	12.61	.062
			Posttest	81.0%	19.91	
	cataphoric	6, 17, 19, 22, 30	Pretest	89.8%	6.88	.044*
			Posttest	99.0%	1.39	

* $p < .05$.

Table 3 describes the results for the nouns used for identification purposes broken down by reference type—anaphoric, associative anaphoric (hereafter shortened to "associative"), or cataphoric reference. An anaphoric reference occurs when a word/phrase in a text refers back to other ideas in the text for its meaning (Lyons, 1999), as exemplified in (5) below. An associative reference means that first mentions of new referents within a discourse can be identified via another already present referent (Allan, 2009), as shown in (6). A cataphoric reference, or backwards anaphora, occurs when a word/phrase refers to ideas later in the text (Chesterman, 1991), as in (7).

(5) *An elegant, dark-haired woman entered the compartment, and I immediately recognized* ***the woman****.*

(6) *They've just got in from New York.* ***The plane*** *was five hours late.*

(7) *I remember* ***the beginning*** *of the war very well.*

As shown in Table 3, all participants correctly chose *the* for the anaphoric references on both tests. The pretest means for the associative and cataphoric references were relatively lower—68.5% and 89.8%, respectively. The posttest means increased to 81.0% for the former (p = .062) and 99.0% for the latter (p = .044), and the mean difference was significant only for the latter.

Considering the several different ways of classifying nouns, Table 4 compares the participants' performance on the concrete and abstract nouns used in their singular forms for classification purposes. Nouns used for identification or definition purposes are excluded because their native-like article usage—*the* for the former and both *a* and *the* for the latter—often makes it impossible to conjecture about the respondents' determination of countability.

Table 4: Distribution of article choices for concrete and abstract nouns

Noun type	Item number	Test	M	SD	P	Average number of responses (%)		
						a(n)	*the*	Ø
Concrete	3, 9, 10, 12, 18, 20, 25, 27, 28, 31	pretest	49.8%	28.39	.089	42 (49.8%)	40 (47.9%)	2 (2.4%)
		posttest	51.8%	29.79		44 (51.8%)	41 (48.2%)	0 (0.0%)
Abstract	2, 15, 24, 26	pretest	39.3%	10.99	.057	33 (39.3%)	33 (39.3%)	18 (21.4%)
		posttest	46.4%	7.97		39 (46.4%)	44 (52.4%)	1 (1.2%)

Note. Cells for the correct answers are shaded.

The pre- and posttest means for the concrete nouns were 49.8% and 51.8%, respectively, and their difference was not significant (p = .089). Most participants who made an article error on the items concerning concrete nouns, such as Item 18, incorrectly chose *the* (47.9% on the pretest and 48.2% on the posttest), not Ø (2.4% on the pretest and 0.0% on the posttest), on both the pre- and posttests. This

result suggests that the single greatest obstacle they faced was identifying the context suggested by the text.

Item 18: *I'm having some difficulties with my visa application. I think I need to find* ***a*** *lawyer with lots of experience. I think that's the right thing to do.*

The pre- and posttest means for the abstract nouns were 39.3% and 46.4%, respectively, and the mean difference was not statistically meaningful (p = .057). For the items concerning abstract nouns, such as Item 26, the pretest correct answer rate (39.3%) was approximately 10% lower than the rate for concrete nouns (49.8%) possibly because the participants had to overcome the obstacle of definiteness distinction coupled with the countability judgment. As shown in Table 4, approximately two-thirds of the participants made pretest errors on either the former (39.3%), leading to the misuse of *the*, or the latter (21.4%), resulting in non-target-like Ø.

Item 26: [The first line of a magazine article]

The night before he died, Michael Jackson ran through ***a*** *six-hour dress rehearsal of his concert.*

After the participants received the instruction, the posttest mean of abstract nouns increased by approximately 7%, and the problem areas were more or less narrowed down to context detection, adding support to Kim's (2018) proposal for a lexicographic approach to teaching articles. During the instruction, deliberate attention was devoted to informing the participants that (1) abstract nouns can be used in countable forms without substantially changing the meaning (Celce-Murcia & Larsen-Freeman, 1999) and (2) countability is a "variable, context-sensitive feature that should be checked" in a dictionary (Kim, 2018:217). Regardless of whether they performed correctly or incorrectly on the task, mention of dictionary consultation was made mostly for these abstract nouns, as in "The dictionary marks the countability status of *rehearsal* as [C, U] or [C]. Countable and first mentioned, thus [the correct answer is] *a*" or "According to the dictionary, *rehearsal* is countable. Nonetheless, *the* is correct because of the modifier" for Item 26. Presumably because of the newly acquired awareness of the noncount-to-count shift that many abstract nouns undergo (Greenbaum & Nelson, 2009) accompanied by dictionary consultation for countability status, the average number of respondents who made errors in judging the countability of abstract nouns decreased considerably from 18 to 1.

In sum, the *t*-test confirmed that teaching the English article system using the binary schema facilitates L2 learners' overall ability to use English articles correctly. However, the participants' performance on the nouns used for classification or

identification purposes differed substantially depending on whether the plus/minus designation of specificity coincided with that of definiteness.

4.2. Reasons behind non-target-like article choices

To answer the second research question of "whether there are particular types of difficulties that L2 learners encounter even after receiving focused instruction on English article use," the participants' written accounts of article choices on the posttest were analyzed according to the classification schemes of Butler (2002) and Chan (2017). The analysis showed that for target-like article use, 83.2% of the reasons were specific, 8.1% were nonspecific, and 8.7% were blank answers. Of the specific reasons, inappropriate hypotheses accounted for 26.3% of the target-like article use. For non-target-like article use, specific and nonspecific reasons constituted 72.4% and 22.2% of the reasons underlying article misuse, respectively, and blank answers constituted 5.4%. Of the specific reasons, 88.3% concerned problems with referentiality (i.e., ±definite), 0.5% involved the misdetection of countability, and 11.2% reflected nongeneralizable or idiosyncratic hypotheses. The thematic analysis of the written accounts pertaining to non-target-like article use or erroneous theories inadvertently leading to target-like article use identified three inappropriate hypotheses that the participants frequently applied in attempting to use English articles correctly. In the following subsections, these inappropriate hypotheses are discussed in detail.

4.2.1. Prioritization of specificity over definiteness

Consistent with previous research findings (e.g., Chan, 2017; Kim & Lakshmanan, 2009), a vast majority of the participants tended to fluctuate between specificity and definiteness—a regular pattern discerned in English article use among article-less L1 groups. Their inability to distinguish the two semantic features and the false prioritization of specificity over definiteness induced misuse of *the* in indefinite contexts (Ionin et al., 2004) or misuse of *a(n)* in indefinite contexts, suggesting that specificity affects L2 learners' article use in both definite and indefinite environments. As shown in Table 2 in the previous section, the participants' performance scores were notably lower when these two semantic features were in conflict, and both the pre- and posttest means for the specific indefinites were by far the lowest of all context types.

Table 5: Distribution of article choices for specific indefinites and nonspecific definites

Noun type	Item number	Test	Average number of responses (%)		
			a(n)	*the*	Ø
specific indefinite	3, 10, 12, 15, 26, 27, 31	pretest	34 (40.5%)	45 (53.6%)	5 (5.8%)
		posttest	35 (42.0%)	49 (58.0%)	0 (0.0%)
nonspecific definite	6, 23, 29, 32	pretest	21 (25.3%)	55 (65.8%)	8 (8.9%)
		posttest	16 (19.0%)	60 (71.7%)	8 (9.2%)

Note. Cells for the correct answers are shaded.

As illustrated in Table 5, more than half of the participants incorrectly chose *the* for specific indefinites on both the pre- and posttests. The participants' lack of understanding of what denotes definiteness was clearly attributable to a significantly strong overuse of the definite article. The rationale behind such selections was almost the same across the specific indefinite items—the target noun is "specific." The participants' performance on these items was almost impervious to the explicit instruction; most participants remained resistant to considering contexts as indefinite on the posttest. One of the participants who was in the top decile of her class applied the specificity feature almost exclusively to determine article use, commenting that "[the correct article is] *the* because the target noun is specific" or "*a* because it's nonspecific" for most of the task items.

Likewise, approximately 20% of the participants falsely chose *a(n)* for nonspecific definites on the posttest for a similar reason that "the target noun is *nonspecific*" in the given context. Other adjectives frequently used to explain their choices included *unknown, undecided,* and *not previously mentioned*.

An interesting observation emerged regarding the definiteness distinction: Not infrequently, L2 learners assess the "hierarchy of nonspecificity" and determine the context accordingly. For example, although the uniqueness—or identifiability—presupposition for the definite (Heim, 1991) is satisfied for both Item 32 and Item 23, the posttest mean of the former was 63.1%, while that of the latter was more than 13% higher at 76.2%.

Item 32: [Announcement on the International Manga Awards home page]
Submissions for the 28th International Manga Awards are now closed. All entries will be reviewed by our judging panel, and ***the*** *winners of each category will be announced at the awards ceremony next month.*

Item 23: *Several days ago, Mr. James Peterson, a famous politician, was murdered. Police are trying to find* ***the*** *murderer of Mr. Peterson.*

For Item 32, more than one-third of the respondents chose Ø (misinterpreting the context as indefinite) in place of *the*, stating that "*winners of each category* are undecided since the review process has not even started yet" or something similar—i.e., the target noun is almost *completely* nonspecific. For Item 23, approximately 23% incorrectly chose *a*. Their reasons were largely identical: "Although we don't know who the murderer is because (s)he is still at large, the murderer is out there anyway"—i.e., the target noun is *partially* nonspecific. Of course, whether this "nonspecificity hierarchy" theory can account for some L2 article errors requires further confirmation. However, the number of respondents who incorrectly interpreted the context as indefinite on the posttest for Item 23 (n = 20) was markedly lower than that for Item 32 (n = 31).

A highly plausible explanation for L2 learners' fluctuation between specificity and definiteness might relate to the inadequate—or even misleading—descriptions used in most English teaching materials, in which the term "specific" is employed to explain (in)definiteness. *OxfordDictionaries.com*, for instance, defines the indefinite article as "a determiner that implies that the thing referred to is *non-specific*" (emphasis added); for the definite article, *Top 20* explicitly directs learners to use it "to refer to a *specific* thing or person" (emphasis added). The fact that specificity and definiteness are two distinct semantic features and that "the specificity distinction cross-cuts the definiteness distinction" (Ionin et al., 2008:557) is almost never introduced. This practice might well have led L2 learners to believe that "a specific reference requires *the*" (Butler, 2002:464). Most of these learners are ignorant of using nouns in the [+specific, –definite] condition, which inevitably results in the overuse of *the* with specific indefinites. The participants' performance on Item 31 exemplifies how persistent this misconception can be: The posttest means were a scant 1.2%, ranking the lowest of all items.

Item 31: **A** *man we both know proposed to me last night, but I'm too embarrassed to tell you who it was.*

In line with Chan (2016:74), the term *specific* was "predominantly used by the respondents to explain definiteness." This finding indicates that most L2 learners experience difficulty in distinguishing between specificity and definiteness and, consequently, between definiteness and indefiniteness (Ionin et al., 2004).

4.2.2. Construal of semantic context on the basis of syntactic structure

While article use essentially depends on semantic context, L2 learners commonly distinguished articles on the basis of the coexistence of modifying information

(syntactic structure) (Butler, 2002; Chan, 2017). For Item 3, for instance, 44% of the respondents incorrectly chose *the* on the posttest.

Item 3: [A memo declining an invitation to dinner]
I'm sorry, but I will be out of town for the weekend. I am visiting ***a*** *classmate from my English class. Her name is Samantha Brown, and she lives in Boston.*

Analysis of their reasons suggested that most of these students held the misconception that modification functions to "uniquely identify a specific subset and indicate definite reference" (Chan, 2017:25). They made a similar comment that "*classmate* is used in its specific sense because it is modified," which is correct, "and so it should take *the*," which is incorrect.

Most English teaching materials elucidate that the definite article is used to refer to a specific thing or person, including those made specific by so-called "structural information" (Hawkins, 1978)—e.g., prepositional phrases or relative clauses—that helps locate the referent. *Top 20* shows examples such as ***The window*** *in the kitchen has been closed all day* and ***The pilots*** *who work for that airline will go on strike at midnight*, in which the boldfaced definite noun phrases are modified by a prepositional phrase or a relative clause (marked with an underline). Most L2 learners seem preoccupied with the phrase "made specific by prepositional phrases or adjective clauses" and falsely equate "being modified" with "being definite."

Compared with specific indefinites, the nouns used as nonspecific definites did not seem to pose as serious a problem; both their pre- and posttest means were approximately 25% higher than the means for the specific indefinites. For Item 32, for example, 63% of the participants chose the correct answer on the posttest, and most participants cited "*winners* is modified" as their reason for selecting *the*.

Item 32: *All entries will be reviewed by our judging panel, and* ***the*** *winners of each category will be announced at the awards ceremony next month.*

The propensity of associating modification with definite contexts must have had a favorable influence in this case because there always exists descriptive content modifying the definites used for cataphoric references. Given that cataphoric use accounts for 40% of all instances of *the* (Yoo, 2009), it appears necessary for teachers and writers of materials to explicitly indicate that the cataphoric use of *the* does not cover those occasions in which modification occurs for classification purposes.

4.2.3. Numerical approach to understanding indefiniteness

Approximately one-third of the participants chose an article at least once on the posttest on the basis of the number of the target noun. Regardless of the context suggested by the text, they chose *a(n)* for most singular count nouns, stating that "the number of the target noun is *one*." Adopting this numerical approach led to the correct answer for singular nouns used for classification purposes, although their understanding of indefiniteness was completely opposite of what it should be. That is, the indefinite article is used when the reader/hearer does not know exactly which one is referred to and thus refers to *multiple* possibilities, whereas the definite article denotes the existence of a *single* entity.

For instance, the posttest mean of Item 31 was the lowest (1.2%), and only one student chose the correct answer. The student, who achieved the highest score on the posttest (91% out of 100%), briefly outlined her reason as follows: "Because *one* man proposed."

> Item 31:**A** *man we both know proposed to me last night, but I'm too embarrassed to tell you who it was.*

Similarly, for Item 25, which had the second lowest posttest mean (4.8%), all four students who selected the correct article were found to have adopted the single-entity strategy. They all similarly commented that "we are talking about *one* city, not many, thus *a*."

> Item 25: *Among many cities in Asia, Hong Kong is also frequently described as* ***a*** *place where East meets West.*

Expectedly, applying the numerical approach for singular nouns used for identification purposes led to erroneous article choices. For Item 29, for instance, almost half of the participants incorrectly selected *a* on the posttest. The application of the numerical approach was accountable for 19% of such misuse (and the prioritization of specificity over definiteness was accountable for 81%).

> Item 29: *Chris went to a newly opened café near his office for a relaxing cup of coffee. To his disappointment, there were screaming kids running around. He wanted to talk to* ***the*** *manager, although he didn't know who he or she was. . .*

Of the 84 participants, only two appeared to understand the concept of indefiniteness correctly. These participants cited the reason "[there can exist] unspecified *many*" at least once when choosing *a(n)* for a singular count noun used for classification purposes—such as Items 3, 9, 12, 18, 25, 27, 28, and 31. Unfortunately,

accurate conceptualization of the semantic feature of indefiniteness did not correlate with posttest performance on these items.

5. Conclusions and implications

The comparison of the participants' mean pre- and posttest scores using the paired-samples *t*-test confirmed that focused instruction on English article use improves L2 learners' ability to use articles correctly in a statistically meaningful way (Research Question 1). However, most participants had trouble grasping the context suggested by the text even after receiving focused instruction on article use. The analyses of the participants' written accounts of article selection and the pattern of options chosen for each item identified three inappropriate hypotheses that the participants frequently applied in attempting to use English articles correctly (Research Question 2)—prioritization of specificity over definiteness, construal of semantic context on the basis of syntactic structure, and numerical approach to understanding indefiniteness. The findings of this study suggest the possible shortcomings of teaching English articles as a binary system based on nominal countability and definiteness. The participants' performance on the posttest showed that they exhibited traceable, nonrandom error patterns of incorrectly associating the definite article with specific contexts (Ko et al., 2009). Such observations were supported by the analysis of reasons, which revealed that a great portion of article errors were constrained by specificity, which most L2 learners erroneously equated with definiteness. Further, most participants (mis)used *the* with modified noun phrases regardless of the definiteness of the target noun (Ionin et al., 2004) on the grounds that they are made specific by modifying information.

The problems identified in this study have bearing on how we can help L2 learners construe the semantics of English articles. First, it is considered necessary to bring the specificity feature into focus because the binary schema does not seem sufficient to decipher the abstruse meanings of English articles. Supporting Ionin et al.'s (2004) Fluctuation Hypothesis, it was obvious that [+specific] triggers overuse of *the* independent of definiteness. To provide input regarding which feature—specificity or definiteness—is appropriate in making an article choice, specificity needs to be incorporated into instruction on the English article system to demonstrate how [+specific] is not necessarily related to [+definite] and vice versa (Master, 1990). Given that most L2 learners already employ the specificity feature in the distinction of definiteness, teachers may consider explaining the definite–indefinite dichotomy in relation to specificity to resolve discrepancies between learners' perception of definiteness and its linguistic definition. For specific infinites, such as a first-mentioned referent in an introductory sentence (Thomas, 1989), learners can be guided to understand that specificity, which assumes "unilateral" identifiability by the writer/

speaker alone, does not satisfy the presupposition of being definite. In this regard, pedagogical strategies might elucidate that (1) the determination of identifiability, which is associated with the definite article (Yule, 1998), depends on both the writer/speaker's and the reader/hearer's "mutual" identifiability of a unique referent within the discourse context and that (2) whether the writer/speaker has one salient referent in mind is completely irrelevant to article use. Additionally, the (in)definiteness of a nonspecific noun needs to be accounted for in terms of whether there exists only one or a multiple of the noun.

Although the findings of the present study have important pedagogical implications, several limitations should be acknowledged. One of the major limitations concerns the representativeness of the sample. The small number of homogeneous participants did not provide sufficient data for broader generalization of the findings. Future research might be pursued on a larger scale with participants with a wide range of proficiency levels. Further, a longitudinal study is suggested to analyze L2 learners' article use in the actual context of use rather than by a one-shot test with a small number of test items to examine the lasting benefit accrued from conceptualizing the article system as a formulated rule. It would be equally meaningful to administer a retention test after the posttest to measure whether the knowledge acquired is internalized. Another limitation may be found in the qualitative data collection method. Due to time constraints, the participants' accounts of article choices were elicited in the form of written documentation, not in interviews, which would have facilitated richer description of their article use. Last, in the absence of a control group, one cannot completely rule out the possibility that the participants implicitly acquired the article rules during the three-month period between the pre- and posttests.

Despite these limitations, this study suggested the primary causes of English article misuse among upper-intermediate to pre-advanced Korean learners of English. To help L2 learners better grasp the article system, it is essential to provide them with explicit, well-defined instructions appropriate for their proficiency levels. Given that learner errors stem from a range of sources, it might be an elusive goal to devise a pedagogy that meets the needs of diverse learners. However, it is feasible to help them overcome specific types of difficulties they commonly encounter by answering a question "generated by [a target learner's] interlanguage" (Swan, 1994:51). To this end, teaching the specificity feature is deemed essential to impart the simple facts that English articles are realizations of encoding definiteness (Halliday & Hasan, 1976) and that there is no morphological indicator for marking specificity in the English article system.

6. References

Allan, K. (Ed.). (2009). *Concise Encyclopedia of Semantics*. Oxford: Elsevier.

Bachman, L. F. (1990). *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Bitchener, J., Young, S., & Cameron, D. (2005). The effect of different types of corrective feedback on ESL student writing. *Journal of Second Language Writing, 14*(3), 191–205.

Butler, Y. G. (2002). Second language learners' theories on the use of English articles: An analysis of the metalinguistic knowledge used by Japanese students in acquiring the English article system. *Studies in Second Language Acquisition, 24*(3), 451–480.

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's guide.* Boston, MA: Heinle & Heinle.

Chan, A. Y. W. (2016). How much do Cantonese ESL learners know about the English article system? *System, 56*, 66–77.

Chan, A. Y. W. (2017). Why do Hong Kong Cantonese ESL learners choose a certain English article for use? *The Asian Journal of Applied Linguistics, 4*(1), 16–29.

Chesterman, A. (1991). *On definiteness: A study with special reference to English and Finnish.* Cambridge: Cambridge University Press.

Dulay, H., Burt, M., & Krashen, S. (1982). *Language two*. New York: Oxford University Press.

Folse, K. S., Solomon, E. V., & Smith-Palinkas, B. (2008). *Top 20: Great grammar for great writing* (2nd ed.). Boston, MA: Thomson Heinle.

Garrett, P., & James, C. (2000). Language awareness. In M. Byram (Ed.), *Routledge encyclopedia of language teaching and learning* (pp. 330–332). London: Routledge.

Greenbaum, S., & Nelson, G. (2009). *An introduction to English grammar* (3rd ed.). Harlow, UK: Pearson Longman.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman Group.

Hawkins, J. A. (1978). *Definiteness and indefiniteness: A study in reference and grammaticality prediction*. London: Croom Helm.

Heim, I. (1991). Artikel und definitheit [Articles and definiteness]. In A. v. Stechow, & D. Wunderlich (Eds.), *Semantics: An international handbook of contemporary research* (pp. 487–535). Berlin: de Gruyter.

Ionin, T. (2006). *This* is definitely specific: Specificity and definiteness in article systems. *Natural Language Semantics, 14*(2), 175–234.

Ionin, T., Ko, H., & Wexler, K. (2004). Article semantics in L2 acquisition: The role of specificity. *Language Acquisition, 12*(1), 3–69.

Ionin, T., Zubizarreta, M. L., & Maldonado, S. B. (2008). Sources of linguistic knowledge in the second language acquisition of English articles. *Lingua, 118*(4), 554–576.

Kachru, Y. (2010). Pedagogical grammars for second language learning. In M. Berns (Ed.), *Concise encyclopedia of Applied Linguistics* (pp. 172–178). Oxford: Elsevier.

Kim, S. (2018). A lexicographic approach to teaching the English article system: Help or hindrance? *Lexikos, 28*, 196–220.

Kim, L. K., & Lakshmanan, U. (2009). The processing role of the Article Choice Parameter: Evidence from L2 learners of English. In M. P. García Mayo, & R. Hawkins (Eds.), *Second language acquisition of articles: Empirical findings and theoretical implications* (pp. 87–114). Amsterdam: John Benjamins.

Ko, H., Ionin, T., & Wexler, K. (2009). L2 acquisition of English articles by Korean speakers. In C. Lee, G. B. Simpson, Y. Kim, & P. Li (Eds.), *The handbook of East Asian Psycholinguistics: Korean* (Vol. 3) (pp. 286–304). Cambridge: Cambridge University Press.

Ko, H., Ionin, T., & Wexler, K. (2010). The role of presuppositionality in the second language acquisition of English articles. *Linguistic Inquiry, 41*(2), 213–254.

Langacker, R. W. (1991). *Foundations of cognitive grammar: Descriptive application* (Vol. 2). Stanford, CA: Stanford University Press.

Lennon, P. (1991). Error and the very advanced learner. *International Review of Applied Linguistics in Language Teaching, 29*(1), 31–44.

Lewis, M. (1993). *The lexical approach: The state of the ELT and a way forward*. Hove, UK: Language Teaching.

Little, D. (1994). Words and their properties: Arguments for a lexical approach to pedagogical grammar. In T. Odlin (Ed.), *Perspectives on pedagogical grammar* (pp. 99–122). Cambridge: Cambridge University Press.

Lopez, E., & Sabir, M. (2019). Article pedagogy: Encouraging links between linguistic theory and teaching practice. *RELC Journal, 50*(1), 188–201.

Lyons, C. (1999). *Definiteness*. Cambridge: Cambridge University Press.

Master, P. (1990). Teaching the English articles as a binary system. *TESOL Quarterly, 24*(3), 461–478.

Master, P. (1997). The English article system: Acquisition, function, and pedagogy. *System, 25*(2), 215–232.

Master, P. (2002). Information structure and English article pedagogy. *System, 30*(3), 331–348.

McEldowney, P. L. (1977). A teaching grammar of the English article system. *International Review of Applied Linguistics in Language Teaching, 15*(2), 95–112.

Miller, J. (2006). An investigation into the effect of English learners' dictionaries on international students' acquisition of the English article system. *International Education Journal, 7*(4), 435–445.

Mizuno, M. (1999). Interlanguage analysis of the English article system: Some cognitive constraints facing the Japanese adult learners. *International Review of Applied Linguistics in Language Teaching, 37*(2), 127–152.

Naver Bilingual (English–Korean/Korean–English) Dictionary. http://endic.naver.com/

Polkinghorne, D. E. (1995). Narrative configuration in qualitative analysis. *International Journal of Qualitative Studies in Education, 8*(1), 5–23.

Price, P. C. (2012). *Research methods in psychology: Core concepts and skills*. Boston, MA: Flat World Knowledge.

Richards, J. C., & Rodgers, T. S. (2014). *Approaches and methods in language teaching* (3rd ed.). Cambridge: Cambridge University Press.

Sarker, B. K., & Baek, S. (2017). Revisiting fluctuations in L2 article choice in L1-Korean L2-English learners. *Journal of Psycholinguistic Research, 46*(2), 367–393.

Spada, N., & Tomita, Y. (2010). Interactions between type of instruction and type of language feature: A meta-analysis. *Language Learning, 60*(2), 263–308.

Swan, M. (1994). Design criteria for pedagogic language rules. In M. Bygate, A. Tonkyn, & E. Williams (Eds.), *Grammar and the language teacher* (pp. 45–55). New York: Prentice Hall.

Thomas, M. (1989). The acquisition of English articles by first- and second-language learners. *Applied Psycholinguistics, 10*(3), 335–355.

Tsang, A. (2017). Judgement of countability and plural marking in English by native and non-native English speakers. *Language Awareness, 26*(4), 343–359.

Yoo, I. W. (2004). *The English articles and nouns* [PDF document]. https://ocw.mit.edu/courses/global-studies-and-languages/21g-213-high-intermediate-academic-communication-spring-2004/readings/MIT21G_213S04_articles.pdf

Yoo, I. W. (2009). The English definite article: What ESL/EFL grammars say and what corpus findings show. *Journal of English for Academic Purposes, 8*(4), 267–278.

Young, R. (1996). Form–function relations in articles in English interlanguage. In R. Bayley, & D. R. Preston (Eds.), *Second language acquisition and linguistic variation* (pp. 135–175). Amsterdam: John Benjamins.

Yule, G. (1998). *Explaining English grammar*. Oxford: Oxford University Press.

Appendix 1

Forced-choice elicitation task and pre- and posttest means

Circle the correct answer for each question. If there is more than one correct answer, you can select multiple choices.

No.	Item	pretest M	posttest M
1–2	[The first line of a magazine article]	94.0%	100.0%
	A / The / **Ø** creative ideas need **a** / the / Ø special climate to grow.	28.6%	45.2%
3	[A memo declining an invitation to dinner]	59.5%	56.0%
	I'm sorry, but I will be out of town for the weekend. I am visiting **a** / the / **Ø** classmate from my English class. Her name is Samantha Brown, and she lives in Boston.		
4	A / The / **Ø** lions are almighty creatures.	79.8%	100.0%
5	Julian ordered a cup of coffee and a dessert, but he didn't touch a / **the** / Ø dessert.	100.0%	100.0%
6–8	At a gallery, I saw a beautiful landscape painting. I really wanted to meet an / **the** / Ø painter of a / **the** / Ø painting, but a / **the** / Ø gallery owner said he didn't know who painted it.	81.0%	98.8%
		100.0%	100.0%
		84.5%	100.0%
9–10	[A text message to a friend who lives nearby]	58.3%	66.7%
	My cat suddenly started to drag his back legs. Do you know **a** / the / Ø good veterinarian in our neighborhood who specializes in treating cats? As you know, I am keeping a pet for the first time and I don't know what to do!	28.6%	32.1%
	[A reply from her friend]		
	I know **a** / the / Ø veterinarian but I am not sure whether he specializes in cats. I will ask around and get back to you soon!		

11	Robert was discussing an interesting book in his class. I went to discuss a / **the** / Ø book with him afterwards.	100.0%	100.0%
12	Susanna works in a "Lost and Found" in an airport. Early this morning, a man approached her and said he was trying to find **a** / the / Ø red-haired girl who must have flown in on Flight 703. Since Susanna was unable to help him, she took him to the nearest check-in desk.	50.0%	50.0%
13–15	We have just arrived from New York. A / **The** / Ø plane was five hours late. After waiting nearly an hour for a bus at the airport, we decided to get a taxi. While driving, a / **the** / Ø driver told us that there was **a** / the / Ø bus strike in downtown Chicago.	70.2% 73.8% 57.1%	97.6% 100.0% 59.5%
16	**A** / **The** / Ø paper clip is handy when holding several sheets of paper together.	44.6%	98.8%
17	A / **The** / Ø happiness that I felt when Charlene became pregnant was beyond description.	86.9%	96.4%
18	I'm having some difficulties with my visa application. I think I need to find **a** / the / Ø lawyer with lots of experience. I think that's the right thing to do.	76.2%	79.8%
19–20	A / **The** / Ø fact that you've known them for years cannot be **an** / the / Ø excuse to not ask them first.	100.0% 60.7%	100.0% 70.2%
21	Typically, a / the / **Ø** dandelions bloom in both the spring and the fall.	71.4%	100.0%
22	A / **The** / Ø tea that I received for my birthday is high-quality.	95.2%	100.0%
23	Several days ago, Mr. James Peterson, a famous politician, was murdered. Police are trying to find a / **the** / Ø murderer of Mr. Peterson.	71.4%	76.2%
24	It is important to draw **a** / the / Ø distinction between what you want and what you need.	39.3%	42.9%

25	Among many cities in Asia, Hong Kong is also frequently described as **a** / the / Ø place where East meets West.	8.3%	4.8%
26	[The first line of a magazine article] The night before he died, Michael Jackson ran through **a** / the / Ø six-hour dress rehearsal of his concert.	32.1%	38.1%
27–28	[In-flight announcement] Ladies and gentlemen, **a** / the / Ø passenger requires medical attention. If there is **a** / the / Ø doctor on board, please identify yourself to one of the cabin crew immediately.	54.8% 100.0%	57.1% 100.0%
29	Chris went to a newly opened café near his office for a relaxing cup of coffee. To his disappointment, there were screaming kids running around. He wanted to talk to a / **the** / Ø manager, although he didn't know who he or she was. He looked around but there was none looking like one, so he had to hurriedly finish his coffee and leave the place.	42.9%	48.8%
30	An / **The** / Ø anger he felt after the accident nearly ended his career.	85.7%	100.0%
31	**A** / The / Ø man we both know proposed to me last night, but I'm too embarrassed to tell you who it was.	1.2%	1.2%
32	[Announcement on the International Manga Awards home page] Submissions for the 28th International Manga Awards are now closed. All entries will be reviewed by our judging panel, and a / **the** / Ø winners of each category will be announced at the awards ceremony next month.	67.9%	63.1%

Teachers' oral corrective feedback and learners' uptake in high school CLIL and EFL classrooms

Ruth Milla
Universidad del País Vasco (UPV/EHU)
Departamento de Didáctica de la Lengua y la Literatura
ruth.milla@ehu.eus

María del Pilar García Mayo
Universidad del País Vasco (UPV/EHU)
Departamento de Filología Inglesa y Alemana y de Traducción e Interpretación
maria.pilar@garciamayo@ehu.eus

Abstract

Oral corrective feedback (OCF) has been reported to be affected by several factors such as learners' age, level of proficiency or the OCF types provided by the teacher. However, little research has been carried out on the variable learning context, even though OCF and uptake vary in rates and types in second language (SL) and foreign language (FL) settings. Moreover, OCF has been clearly under-researched in classrooms that follow a content and language integrated learning (CLIL) approach. As CLIL programs are being widely implemented mainly in European settings and differences in context characteristics suggest variations in OCF and learners' uptake, the present study aimed to compare the recorded classroom interaction data (22 hours 43 minutes) from an intact class of learners (N=26) in their last year of secondary education (age 17-18), attending the lessons of an English as a FL (EFL) teacher and a Business Studies (CLIL) teacher. Findings show significant differences as to the proportion and OCF types used, as well as different learners' behavior regarding the rates of uptake and repair and the uptake after the use of recasts. Pedagogical implications are offered as to how to maximize the potential benefits of OCF in FL classrooms.

Keywords: Oral corrective feedback, corrective feedback episodes, EFL, CLIL, learning context

Resumen

Estudios previos han hallado que la retroalimentación correctiva oral (RCO) puede variar dependiendo de factores tales como la edad y el nivel de conocimiento de lengua de los aprendices o los tipos de RCO que facilita el profesor. Sin embargo, existe

escasa investigación sobre la variable del contexto de aprendizaje, aunque la RCO y la respuesta de los aprendices difiere en cantidad y tipo en contextos de segunda lengua y contextos de lengua extranjera (LE). Además, la RCO apenas ha sido investigada en aulas que siguen un enfoque de aprendizaje integrado de contenido y lengua extranjera (AICLE). Debido a que los programas de AICLE se están implementando ampliamente en contextos europeos y que las diferencias entre los contextos sugieren posibles variaciones en cuanto a la RCO y la respuesta de los aprendices, este trabajo tiene como objetivo comparar la interacción oral grabada (22 horas 43 minutos) en un aula intacta (N=26) de segundo curso de bachillerato (edad=17-18) en las clases de inglés como LE con otra de estudios empresariales (AICLE). Los resultados muestran diferencias significativas en cuanto a los tipos de RCO utilizados y la respuesta de los aprendices ante las reformulaciones. Se presentan implicaciones pedagógicas relativas a cómo obtener el máximo beneficio de la RCO en aulas de LE.

Palabras clave: Retroalimentación oral, episodios de retroalimentación oral, ILE, AICLE, contexto de aprendizaje**.**

1. Introduction

Oral corrective feedback (OCF) has been defined as "a reactive type of form-focused instruction which is considered to be effective in promoting noticing and thus conducive to learning" (Yang & Lyster, 2010: 237). OCF is a pedagogical technique that has been claimed to be beneficial for the process of second language acquisition (SLA) (Russell & Spada, 2006; Sheen, 2011). This oral corrective technique is part of what has been termed as corrective feedback episodes (CFE). We have chosen the term CFE as it is the most frequently used in the literature (Lyster, Saito & Sato, 2013; Mackey, Gass & McDonough, 2000) after Lyster & Ranta's (1997) seminal study. Typically, CFEs consist of three moves: Error, OCF and Uptake. Example (1), which belongs to the database of the present study, as all the examples in this paper, illustrates them:

(1) Learner: (...) there *haven't been any victims.
Teacher: there weren't any victims. You are talking about the past, right? There weren't any victims. There weren't any. What other word do you have for victim?
Learner: but was today!
Teacher: yes, but the tense that you have is past: "were involved". It is not: "there has been an accident" and then you can use the present. No, the past. Another word for victims?

A CFE consists of three moves: the first is the error that the learner produces within oral interaction, in example 1 an error having to do with the use of tenses. The second turn represents the teacher's OCF move, which can take different forms. In a seminal study on OCF, Lyster & Ranta (1997), identified six types of OCF: recasts and explicit correction on the one hand, which have been grouped into a larger category referred to as reformulations since they offer the target form, and, on the other, repetition, clarification requests, elicitation and metalinguistic cues, which have been termed prompts (Lyster 2002, *et passim*) and try to elicit learners' self-repair. Of all these, recasts are the most frequently used types and they have been further divided into two different types: conversational and didactic (Sheen, 2006). Conversational recasts are more implicit and less direct, as teachers use longer sentences to reformulate the error and, therefore, they tend to be a less salient type of OCF. Didactic recasts, on the contrary, are more explicit and direct, and teachers use shorter reformulations of the erroneous utterance and isolate the repaired form to make it more visible. Recasts of this latter type become more salient and more easily perceived by the learners and they typically appear in form-oriented lessons. In example 1, the teacher reformulates the erroneous verb tense and provides the target form, using an explicit correction move. Finally, the third move corresponds to the learner's reaction to the OCF provided, which is referred to as uptake. The uptake move does not appear in all CFEs and, when it does, it can be of several types: the learner may fail to repair the error (the 'needs repair' category in Lyster & Ranta, 1997), as in (1) above; the error is repaired by the learner ('self-repair') or by another learner ('peer-repair'). In (1) one can find yet another move, which consists of further OCF in the form of a metalinguistic explanation by the teacher, who then continues with the topic of the lesson.

OCF is a topic that has been explored widely in the field of SLA (Nassaji & Kartchava, forthcoming). Different factors have been considered in previous studies, such as learners' individual differences (ID) or the type of OCF that could be more appropriate to address different types of errors (Kartchava & Ammar, 2014). Regarding the former, learners' age (Oliver & Grote, 2010; Panova & Lyster, 2002), learners' beliefs (Kartchava, 2016; Yang, 2016) and proficiency level (Ammar & Spada, 2006; Li, 2014) have been reported to have an impact on learners' response to oral corrections. As for the latter, research has reported the predominance of recasts, with the exception of OCF in high school classrooms, where prompts have been claimed to be more frequent (Brown, 2016). In terms of their effectiveness, both recasts (Goo & Mackey, 2013; Long, 2015) and prompts have been found to lead to successful uptake and repair, depending on variables such as error type. Thus, prompts have been found to be more effective for grammar errors while recasts appear to lead to higher rates of uptake when used for pronunciation or lexical errors (Bryfonski & Ma 2020, Gurzynski-Weiss, 2010; Saito, 2013). In addition, a balanced and tailored provision of

types has been recommended (Li, 2014; Li, 2018; Lyster & Mori, 2006; Saito & Lyster, 2012).

The variable instructional context, i.e., the type of learning setting, has been acknowledged to affect learners' reaction to OCF in general and their uptake of the different OCF types (Mackey & Goo, 2007; Sheen, 2004) but it has been very scarcely researched. Although previous studies have explored oral CFEs in SL (Kartchava & Ammar, 2014a; Loewen, 2004; Lyster, 2004) and foreign language (FL) settings (Goo, 2012; Havranek & Cesnik, 2001; Yilmaz, 2012) only a handful of studies have established comparisons between the two (Milla & García Mayo, 2014; Llinares & Lyster, 2014; Lochtman, 2007; Lyster & Mori, 2006; Sheen, 2004), with findings showing relevant differences between the two contexts.

The aim of the present study is to investigate OCF in two different settings, a traditional English as a Foreign Language (EFL) classroom and a Content and Language Integrated (CLIL) classroom, in order to assess the extent to which instructional context has an impact on CFEs and learner uptake. We will adopt the definition of CLIL provided by Dalton-Puffer (2011:183) as an educational approach where "[...] curricular content is taught through the medium of a FL [foreign language] typically to learners participating in some form of mainstream education at the primary, secondary or tertiary level". In what follows we first provide a summary of the studies that have compared OCF in different language contexts, to then move onto the actual details of the current study, its major findings and implications.

2. Literature review

To the best of our knowledge, Sheen (2004) was the first study that compared CFEs in different learning contexts. She examined four different classroom settings: French immersion (FI) and English as a Second Language (ESL) in Canada, ESL in New Zealand and EFL in Korea and focused on recasts (reformulations of the erroneous utterance by providing the target form) due to the small proportion of the rest of OCF types in her database. Her findings showed differences as to OCF provision and uptake: the use of recasts, although high in all settings, was significantly higher in New Zealand and Korean classrooms than in FI and ESL in Canada. Significant differences were also found between the amount of recasts in Korean EFL and New Zealand ESL settings.

In addition to the differences in OCF provision, the rate of uptake after recasts and subsequent repair was found to be higher in the Korean EFL and New Zealand ESL contexts than in Canadian ESL and immersion classrooms. Sheen attributes

this difference in uptake to the learners' orientation to form rather than meaning in Korean EFL and New Zealand ESL, which consequently led to greater noticing of this OCF type. The fact that an ESL setting is oriented to form and not to meaning contrasts with other SL contexts, typically oriented to meaning. This difference might be related to the fact that this was an intensive course with a native speaker teacher and the learners were somewhat older. Sheen (2004) calls for more research on different topics, such as the impact of contextual factors on OCF patterns and learner uptake.

Lyster & Mori (2006) also considered how instructional setting could be a relevant factor in OCF provision. They observed and recorded intact lessons in two different learning settings at the elementary-school level. There were 18.3 audio-recorded hours of FI lessons for English speaking learners in Canada with French as a Second Language (FSL), taken from Lyster & Ranta's (1997) seminal study, and 14.8 hours of video recordings of Japanese immersion (JI) for English speakers in the USA, that is, Japanese as a foreign language (JFL), taken from the data reported on in Mori (2002).

Lyster & Mori (2006) analysed the lessons with Part A of the communicative orientation of language teaching (COLT) coding scheme used by Spada & Fröhlich (1995). They found that FSL lessons had a more experiential orientation and the focus of the lesson was generally on meaning and rarely on form, while JFL lessons had a more analytic orientation and the focus of the lesson was predominantly on form. The authors then compared CFEs in each of the two settings. Their analysis revealed that the proportion of errors corrected by the teachers was similar (67% and 61%) and that OCF types were similarly used across the two contexts, with recasts being the predominant type (54% and 65% of all OCF moves), prompts in a smaller proportion (38% and 26%) and explicit correction relatively infrequent (7% and 9%).

However, findings related to uptake revealed differences: rates of uptake and repair were higher in JFL than in FSL. Uptake of the different types varied with the largest amount coming from prompts in the FSL classrooms (62%) and from recasts in the Japanese language learning setting (61%). Similarly, the proportion of repair was reversed, the highest amount of repair after prompts being found in FSL (53%) and after recasts in JFL (68%). Uptake and repair following explicit correction moves was similarly small in both settings, less than 10%.

Based on these findings, the authors proposed the Counterbalance Hypothesis (CH), which states that:

> *[...] instructional activities and interactional feedback that act as a counterbalance to a classroom's predominant communicative orientation are likely to prove more effective than instructional activities and interactional feedback that are congruent with its predominant communicative orientation. (Lyster & Mori, 2006: 269)*

Thus, the teachers in the meaning-focused lessons of the FSL context obtained more learners' uptake with the use of form-focused teaching techniques, such as prompts. On the other hand, in the JFL classrooms, which were found to be more oriented to form, more meaning-focused or implicit types resulted in larger rates of uptake and repair due to learners' awareness of OCF in these settings. The authors explain that the COLT coding scheme helps to recognize the orientation of a given classroom to form or to meaning. The analysis of the activities can help researchers to recognize the learners' orientation, which, in turn, seems to predict their ability to perceive and use the different OCF types. Therefore, the authors advocate for a balanced provision of OCF, using different types in order for the learners to be able to notice them. They also call for more 'fine-grained' classroom research where the TL is the same in all the settings and where classrooms with FL instruction are compared with immersion settings.

Lochtman (2007) is another example of a comparative study. The data were gathered in German as a Foreign Language (GFL) classrooms (Lochtman, 2002) and in FSL classrooms (Lyster & Ranta, 1997). The comparison revealed differences in OCF provision: teachers in FL settings tended to offer prompts while SL teachers preferred to use recasts. As for uptake, similar results were found, with higher rates in response to prompts in both settings but in GFL recasts also obtained remarkable rates of repair. Therefore, Lochtman's (2007) results were in line with those of Lyster & Mori's (2006) study regarding the finding that instructional context influences the three moves of CFEs.

There is a language learning setting that has been underresearched as far as OCF is concerned, namely, the CLIL setting. CLIL occurs typically in FL contexts and, although it derives from immersion programs in Canada, it includes not only the communicative use of the language, but also skills, contents and competences, with a holistic vision of the language. CLIL teachers are not native speakers of the TL or language teachers, but subject teachers, who normally plan their content lessons alongside the traditional FL lessons taught by language experts. Nevertheless, as CLIL is an umbrella term and has been implemented profusely in primary, secondary and university levels throughout Europe (Pérez Cañado, 2012), there are multiple types of programs that can be found under this approach, as we will see in what follows. All in all, CLIL lessons differ from FL lessons since the focus has moved from language form towards content. In this sense, OCF can be hypothesized to take different forms and lead to different rates of learners' uptake.

A recent study comparing FSL in Canada, JFL in the USA and CLIL classrooms in Spain was carried out by Llinares & Lyster (2014). The researchers used data from Lyster & Mori's (2006) FSL and JFL classrooms and included a CLIL context, which,

as mentioned above, involves an integration of form and meaning –language and subject matter– and more hours of exposure to the TL. In this study, CLIL learners, primary school children, had Spanish as their L1 and were enrolled in a bilingual program, with English as the TL. Based on Lyster & Mori's (2006) comparison of two different immersion classrooms, Llinares & Lyster (2014) performed a three-way analysis of CFEs examining the frequency and distribution of OCF types as well as repair and uptake of those types and tried to identify the factors that contribute to similarities or differences across the instructional settings.

Llinares & Lyster (2014) reported that OCF types occurred in a similar proportion in the three settings: recasts were the most frequently used type, followed by prompts and the least used type was explicit correction. As for uptake, the pattern was reversed: higher uptake after recasts was found in CLIL and JFL while FSL learners showed more uptake after prompts. Recasts were much more effective – in terms of repair– in CLIL classrooms, with the opposite happening in FSL classrooms. In JFL, similarly high rates of repair were found for recasts, prompts and explicit correction. Finally, as for recast type, the researchers use the distinction between conversational and didactic recasts (Sheen, 2006), mentioned above. In Llinares & Lyster (2014) study, CLIL and JFL teachers used a greater amount of didactic recasts while FSL teachers preferred conversational recasts, a feature that the authors present as a possible explanation for the differences in uptake and repair: the explicitness of didactic recasts may favour the learners' awareness of the correction and, in turn, increase the effectiveness of this OCF move.

Llinares & Lyster (2014) examined classroom differences and reported that interaction in CLIL and JFL shared more characteristics than JFL and FSL, which were both termed as immersion contexts by Lyster & Mori (2006). Llinares & Lyster (2014) explain that this finding has to do with the fact that, as there are different types of CLIL programs (Lasagabaster & Sierra, 2010), immersion programs differ from one another as well (Tedick & Cammarata, 2012). Thus, in each of the contexts, teachers' beliefs and previous experience shape OCF patterns and the type of instruction seems to influence learners' noticing of OCF as well. The authors consider it interesting to explore CFEs in secondary level classrooms, where CLIL teachers' background is different, since they are subject matter specialists and have no specific training as language teachers. Llinares & Lyster (2014) call for further research on the effect of the instructional context variable on OCF patterns.

In a recent meta-analysis, Brown (2016) considered the teachers' background in relation to their provision of CF. The author explains that teachers with more L2 training tended to provide more prompts than recasts and pay more attention to lexical than to phonological errors. Therefore, it seems interesting to compare the

behavior regarding the OCF of teachers with previous linguistic training and those who are content subject teachers with a certified knowledge of the target language, such as CLIL teachers in secondary education.

The lack of research on OCF in CLIL classrooms (Dalton-Puffer & Nikula, 2014) was what prompted Milla & García Mayo (2014) to carry out another comparative study, where the corrective behavior of a CLIL and an EFL teacher as well as the uptake and repair patterns of a group of 30 intermediate level learners in different lessons were analyzed. The learners were 17-18 years old and belonged to an intact class in the second year of post-compulsory secondary education in a trilingual program (Spanish, language X and English), in which about 30% of teaching time was devoted to each of the languages. Following a classroom observation procedure, the authors audio-recorded a total of 377 minutes of three CLIL lessons (Business Studies in English) and four EFL lessons. Besides the recording, the first author observed the lessons, which were analyzed using the COLT scheme as in Lyster & Mori (2006), revealing that CLIL lessons were clearly oriented to meaning while EFL lessons were more form-oriented. This finding contrasts with Llinares & Lyster's (2014) CLIL classrooms, where attention to form happened in a content/meaning-oriented setting. The reason may lie in the fact that in Milla & García Mayo (2014) the CLIL teacher was a subject specialist with no specific training in language teaching, as is typical in secondary education in Spain. Conversely, primary school CLIL teachers are generally English language teachers that also teach subject matters in English. Therefore, it would be expected that CLIL secondary school teachers are less oriented to form and their lessons more focused on meaning, in a similar way to immersion classrooms.

The type of feedback provided by the teachers was classified into the six types identified by Lyster & Ranta (1997): recasts, repetitions, clarification requests, elicitations, metalinguistic information and explicit correction. Milla & García Mayo (2014) also examined which type of feedback promoted immediate uptake. The analysis of the CFEs in the two contexts revealed significant differences as to the amount of errors corrected by the EFL (72%) and CLIL (53%) teachers. The authors also found that the CLIL teacher used recasts almost exclusively while the EFL teacher used the whole spectrum of types, favoring explicit correction, elicitation, repetition, and metalinguistic feedback. Consider examples (2) and (3), which illustrate how the two teachers address a grammar error, by means of a recast in a CLIL classroom and by means of metalinguistic information and elicitation in an EFL classroom, respectively:

(2) Learner: the value it has when the company *start...
Teacher: ok, when the company starts ... and do you remember that in order to calculate we have a simple formula? OK? It is...? (addressing another learner). Do you remember?

(3) Learner: something you did *give an enormous sense of achievement.
Teacher: the verb is OK, David, but not the tense. "Something you did", it's past so you cannot say give.
Learner: xxx I don't know.
Teacher: if the sentence is in the past, you will need a verb in the past, so?
Learner: gave.

In example 2, the CLIL teacher reformulates the grammar error, but he asks a question and continues with the lesson. Therefore, the learners might miss the correction, as they do not have an opportunity to react to the recast. According to the Noticing Hypothesis (Schmidt, 1990), noticing is essential for L2 learning. If learners miss the corrective intention of recasts, it is likely that they do not have an effect on their language learning process. In example 3, we find a very explicit indication of where the error occurs and a second move by the teacher with an elicitation.

The findings in Milla & García Mayo (2014) were in line with Lyster & Mori's (2006), where the more form-oriented teachers (JFL) preferred prompts or didactic recasts and the more meaning-oriented teachers (FSL) used conversational recasts in a remarkably higher proportion with respect to other types. Milla & García Mayo (2014) showed that, although the two teachers used different types of OCF, both used recasts very often and, therefore, no significant differences in the use of OCF between the two classrooms were reported, except for repetition and explicit correction, which were not used by the CLIL teacher. This lack of significance was attributed to the small amount of data that were obtained out of the recorded lessons. The authors were interested in unravelling the details of the CFEs occurring in the two contexts and carried out a qualitative analysis of OCF moves in order to explore the differences identified but not confirmed by the statistical analysis. The qualitative analysis revealed that the teachers not only provided different types of OCF but also used the types in a different way. Thus, the CLIL teacher made topic continuation moves after the correction, not allowing learners to react to OCF moves. On the contrary, the EFL teacher displayed a wider variety of OCF types and combined them, using what Lyster & Ranta (1997) termed as 'multiple feedback' (i.e. using several OCF types for the same error in the same move). Example 4, with data from our database, illustrates this phenomenon. The EFL teacher uses an explicit correction move and a metalinguistic cue, coded as explicit correction, following the conventions in Lyster & Ranta's (1997) seminal study. This type of OCF move turns out to be really salient and therefore leads to uptake in most cases.

(4) Learner: *suits [swi:ts] and...
Teacher: not sweets [explicit correction]. Sweet is something that you eat and is full of sugar [metalinguistic cue]. Suits [su:ts] [explicit correction] yes?

As for the results of the analysis of the uptake move, OCF led to higher rates of uptake in EFL lessons (82%) than in CLIL (52%). Regarding the learners' immediate response to the OCF types, it was found that there was a higher proportion of uptake after recasts and clarification requests in EFL and after elicitation and recasts in CLIL. Although these differences were not significant, the researchers reported that there was a tendency for learners to respond more positively to implicit types in a form-oriented lesson and to explicit correction in the meaning-focused lessons of CLIL. These findings are in line with Lyster & Mori (2006) and thus explained by the CH. However, the authors acknowledged the limitations of their data and called for further research on OCF in the two settings, EFL and the under-researched CLIL context.

3. The present study

The main aim of this study is to explore CFEs in two different learning settings, namely CLIL and EFL, and assess the extent to which the type of instructional context affected the teachers' corrective behavior in terms of the amount and type of OCF chosen. The learners' uptake was also examined in both contexts as well as the effect that each of the OCF types had on their production. As mentioned above, in previous research Milla & García Mayo (2014) reported few significant quantitative differences between the two contexts, which was probably due to the small dataset analyzed. However, interesting qualitative differences were observed. The current study has enlarged the database and has also analyzed two other EFL and CLIL teachers to discard the potential impact of individual teachers' idiosyncrasies on the results.

The following research questions were considered:

1. Is there a difference in the amount and types of OCF provided in CLIL and EFL classrooms?

2. Does type of error influence quantity and quality of OCF in each of the classrooms?

3. Does OCF lead to more uptake in CLIL or EFL classrooms? Do learners react differently to OCF types in the two classrooms?

3.1. Participants

Both the teachers and learners participating in this study belonged to a public high school. For the main study, two teachers volunteered: the CLIL teacher was a male with a background in Economics and 20 years of teaching experience, the last

seven years of which he had used English as the language of instruction. He had a C1 level in the language according to the Common European Framework of Reference for Languages (Council of Europe, 2001). The EFL teacher was a female with 26 years of experience, with a degree in English Studies and several professional courses completed throughout her academic life.

The two participant teachers taught the same group of learners, 26 adolescents (17-18 year old, 14 female, 12 male students) in their second year of post-compulsory education with a proficiency level in English between B1 and B2 of the Council of Europe (2001), as attested by the Oxford Placement test administered to them. As suggested in previous research (Basturkmen, 2012), there was only one group of learners so that the teachers' behavior could be better compared.

Regarding the context, the participants belonged to a trilingual (Spanish, language X, English) programme, with approximately 30% of the time conducted in each of the languages. The EFL teacher, who met with the participants three hours per week, followed a communicative methodology but she also used a grammar book and put great emphasis on formal aspects of the language. Reading comprehension and writing were also very important as the participants had to take the university entrance exam at the end of their academic year. The methodology used by the CLIL teacher, who met with the participants four hours per week, was based on reading and discussing some notes and articles he himself provided, as well as carrying out practical exercises. The focus was mainly on content, also to prepare the students for the university entrance exam. More information about the lessons will be provided below.

Although the main corpus in the present study comes from the 2nd year post-compulsory education teachers and learners, in order to discard a potential teacher effect, another pair of teachers were also recorded and the interaction data in their lessons was analysed. By observing another pair of teachers, we could assess whether the differences between the CLIL and EFL teachers in 2nd year also existed in 1st year as well and were not due to the 2nd year teachers' idiosyncrasies. Several CLIL and EFL lessons were recorded in a 1st year classroom in the trilingual programme at the same school. In the 1st year group we also observed lessons from two subjects, EFL and CLIL. The EFL lessons were taught by a female teacher with 24 years of experience. The CLIL lessons (*Science for the Contemporary World*, a compulsory subject) were taught by another female teacher with 14 years of teaching experience. The participants were 25 16-year-old 1st year post-compulsory education students (22 female and 3 male) with a B1 English proficiency as attested by the Oxford Placement Test (18 learners obtained B1, 6 B2, 1 C1).

3.2. Procedure

In order to perform a comparative analysis of the CFEs occurring in EFL and CLIL contexts, we employed a classroom-observation methodology, a standard procedure in this type of studies (Llinares & Lyster, 2014; Lochtman, 2007; Lyster & Mori, 2006). Fifteen CLIL and twelve EFL lessons were observed and audio-recorded (a total of 22 hours 43 minutes). As in previous studies (Lyster & Mori, 2006), we used the COLT scheme (Spada & Fröhlich, 1995) to analyze the predominant orientation of the lessons, either to meaning or language form.

Besides this, 5 CLIL and 6 EFL lessons were recorded in the 1st year (8 hours 5 minutes in total).

3.3. Data coding and analysis

The recorded data were transcribed following CHILDES (McWhinney, 2000) conventions and each of the moves in the CFEs was analyzed. First, the number and type of errors were examined in each of the classrooms, EFL and CLIL. Secondly, OCF and the use of the different OCF types were analyzed. The OCF types were the ones in Lyster & Ranta's (1997) taxonomy, namely, recasts, clarification requests, repetitions, metalinguistic clues, elicitation and explicit correction. The teachers' use of the larger categories, reformulations (recasts and explicit correction) and prompts (clarification requests, repetitions, elicitation, metalinguistic cues, explicit correction) was also analysed. Finally, uptake of OCF and of the different types was compared in each of the classrooms. Data were transcribed and codified by the two authors independently and inter-rater reliability reached 98% agreement. Table 1 below displays the codes used for this analysis.

Table 1: Transcription codes employed in the codification of the CFEs

CODE	MEANING	CATEGORY
*L1	UNSOLICITED L1 USE	ERROR TYPES
*G	GRAMMAR ERROR	
*P	PRONUNCIATION ERROR	
*L	LEXICAL ERROR	
NC	NON-CORRECTED ERROR	CF TYPES
RC	RECAST	
CL	CLARIFICATION REQUEST	
RP	REPETITION	
EL	ELICITATION	
ML	METALINGUISTIC CUES	
EC	EXPLICIT CORRECTION	
RF	REFORMULATION	
PM	PROMPT	
UN	NO UPTAKE	UPTAKE TYPES
NR	NEEDS REPAIR	
SR	SELF-REPAIR	
Pe	PEER REPAIR	

Moreover, following Lyster & Ranta's (1997) conventions, multiple feedback moves occurring in the same CFE were codified as single feedback moves. Thus, the following equivalences were used: Recast or Explicit correction together with Metalinguistic cues or Elicitation were codified as Explicit correction. Then, Metalinguistic cues together with Elicitation were codified as Elicitation. The analysis of multiple feedback moves was therefore carried out from a qualitative descriptive, as will be seen below.

Data were analyzed quantitatively by means of the R program (R Development Core Team, 2008) and non-parametric tests (Fisher and Chi-square). Data were also analyzed qualitatively, with descriptions of CFEs that illustrate specific aspects that were not shown in the statistical analyses.

3.4. Results

In what follows, the results of the COLT analysis in the 2nd year EFL and CLIL classrooms will be presented and the differences between the two classroom contexts

highlighted. Regarding participant organization of the activities, in EFL lessons a teacher-fronted methodology was found more than half of the time (59%), but group or pair activities and individual work of the learners also occurred. On the contrary, in CLIL lessons teacher-fronted activities were prevalent (94% of the total time) and only a very small amount of time was allowed for individual activities.

Regarding content focus, EFL lessons were divided into language and thematic content, with little time devoted to thematic content. As expected, CLIL lessons were focused on thematic content and no time at all was devoted to language content alone or in combination with thematic content. In the EFL lessons, the teacher and the learners worked together more than half of the time (about 59%) and the learners worked with their peers and with texts the rest of the session, whereas in the CLIL classroom the lessons were controlled by the teacher most of the time. This means that in the EFL sessions there were more opportunities for learner interaction than in the CLIL lesson, which leads to learners having fewer opportunities for free production and repair of the errors, as we will see below. Finally, as for modality, EFL activities were more centred on oral skills, while in CLIL the focus was on both oral and written skills. As has been noted in previous research, learners' performance may vary dependent on context-related variables of the setting where they operate (Brown, 2016; Nassaji, 2020). The differences between the two classroom settings in our study are clear and, consequently, differences in CFEs are also expected.

After examining the lesson orientation in each of the contexts, CFEs were analyzed, starting with their first move, that of error. Learners showed a different behavior depending on the context: there was a larger number of errors in CLIL (562) than in EFL (171) and the type of errors also differed, with the unsolicited use of the L1 being a very frequent error type[1] in CLIL and very scarce in EFL, as Table 2 below shows.

Table 2: Number and percentages of error types in EFL and CLIL

ERROR TYPE	EFL	CLIL
L1 USE	28 (16.4%)	392 (69.8%)
GRAMMAR	43 (25.1%)	45 (8%)
PRONUNCIATION	73 (42.7%)	104 (18.5%)
LEXICAL	27 (15.8%)	21 (3.7%)
TOTAL	171	562

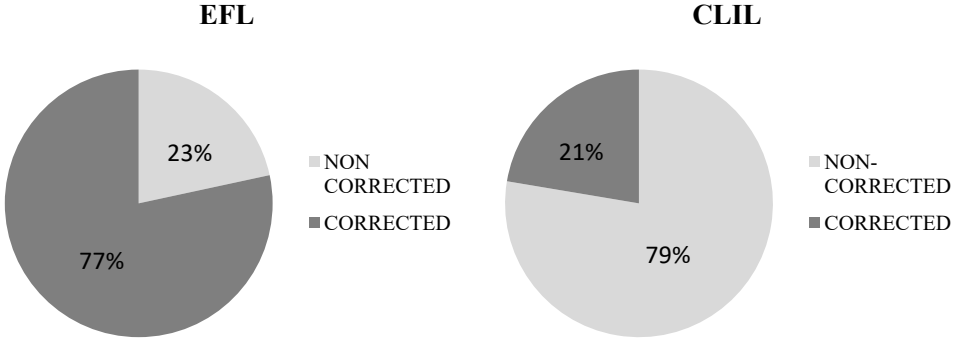
1 We are aware that a moderate use of the L1 is no longer considered problematic in the second/foreign language classroom (Antón & DiCamilla, 1998). However, as will be explained, the teachers in our study did consider L1 use an error and acted accordingly. That is the reason why unsolicited L1 use has been analyzed.

Let us provide now the answers to the three research questions in the present study. The first question considered whether there could be a difference in the amount and types of OCF provided in the CLIL and the EFL classrooms. In order to answer it, data corresponding to the OCF provided by the 2nd year CLIL and EFL teachers were analyzed. Based on the findings from previous studies (Milla & García Mayo, 2014; Llinares & Lyster, 2014; Lochtman, 2007; Lyster & Mori, 2006), we expected to find significant differences regarding the amount and the types of OCF preferred by the EFL and the CLIL teachers. Specifically, more explicit types and prompts were expected in the EFL lessons, although recasts were expected to be the most frequently selected OCF type in both settings (Brown, 2016).

As explained above, in order to discard a potential teacher effect that could be influencing the results, data were collected from a second group of learners in their 1st year of post-compulsory education at the same school with another pair of teachers, EFL and CLIL (Science in this case). We discarded the teacher effect since we obtained statistically similar results in the behavior of the EFL teachers in 1st and 2nd year as well as in the 1st and 2nd year CLIL teachers, both concerning proportion of errors corrected, analyzed by Chi-square tests (EFL p=0.5833834; CLIL p=0.6121218) and OCF types, analyzed by Fisher tests (EFL p=0.5508; CLIL p=0.05938). We are aware that the similarities could be due to the fact that they belong to the same school but the data from the two teachers in 1st year indicate that in the present study CLIL and EFL teachers behaved in a different manner, and these differences were not due to the idiosyncrasy of the individual teachers.

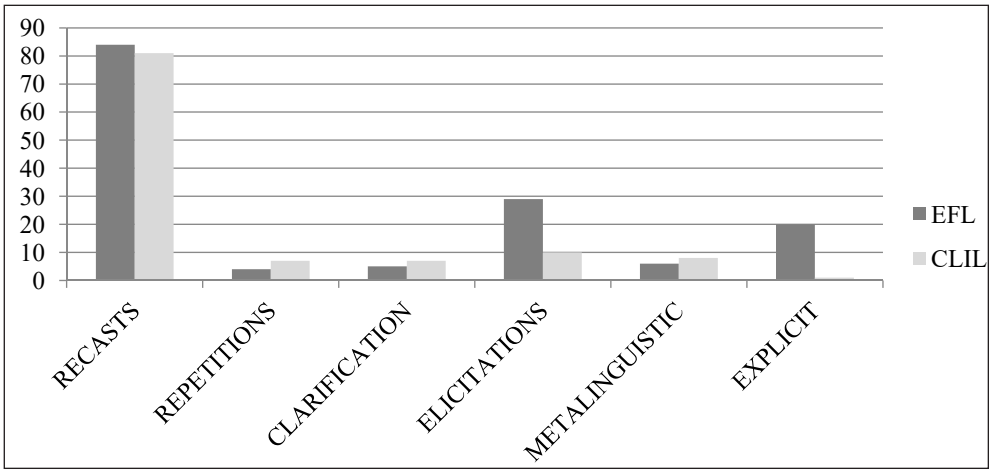
Regarding the results of 2nd year classrooms, the proportion of errors corrected was analyzed. Graph 1 shows striking differences (confirmed by the Chi-square test, p=0) between the two teachers: the EFL teacher corrected more than 77% of errors (131 out of 171), while only 21% of them were corrected by the CLIL teacher (121 out of 562 errors).

Graph 1: Percentage of errors corrected and not corrected in EFL and CLIL



Significant differences were also found in the use of the OCF types (Fisher test, *p*=0). While both teachers preferred recasts over the rest of the OCF types, as commonly found in other contexts (Sheen, 2004), our EFL teacher also resorted relatively frequently to prompts such as elicitation and, within reformulations, both recasts and explicit correction were used. Graph 2 illustrates these findings. As for the use of the larger categories, reformulations and prompts, no differences were found when comparing the two classrooms (Chi-square test, *p*=.875), since the proportion of the use of these categories was similar with reformulations being used much more frequently (107 cases in EFL, 68%, and 84 in CLIL, 70%) than prompts (50 cases in EFL, 32%, and 36 in CLIL, 30%). However, if OCF types are analyzed in detail, Graph 2 shows how the EFL teacher used both types of reformulations (both didactic recasts and explicit correction), while the CLIL teacher only resorted to recasts, which were of an implicit nature in his case.

Graph 2: Total number of OCF Types in EFL and CLIL Classrooms



Considering recent findings on the potential benefits of the use of the L1 in SL and FL contexts and of translanguaging (García, 2019), analyses were carried out tallying lexical, grammar and pronunciation errors and not those related to L1 use. Significant differences were found as the teachers' proportion of correction (*p*= 0) as well as the use of OCF types (*p*=.007) were different in the two contexts.

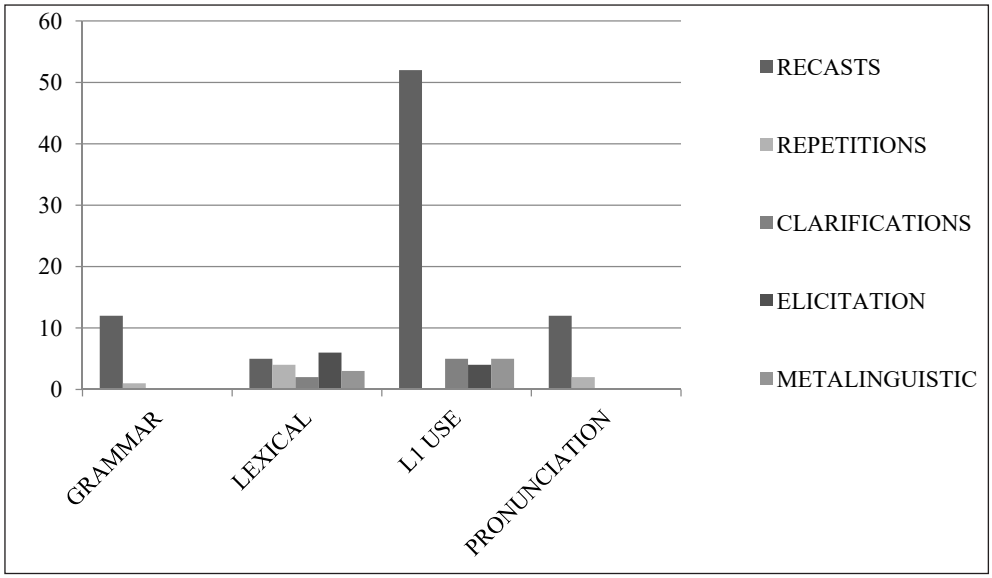
Our second research question considered whether the type of error could influence the quantity and quality of OCF provided in each of the classrooms. Error type was hypothesized to affect the amount and type of OCF preferred by the teachers. Regarding the proportion of each error type corrected in the two classrooms,

significant differences were found. Thus, the EFL teacher corrected 68% of grammar errors, whereas the CLIL teacher addressed 31%. L1 use instances were addressed in 67% of the cases in EFL but 18% in CLIL. As for lexical errors, 70% and 81% of them received OCF in the EFL and the CLIL classrooms, respectively. The most striking contrast was found in pronunciation errors, which were corrected in 91% of the cases in EFL and 22% in CLIL.

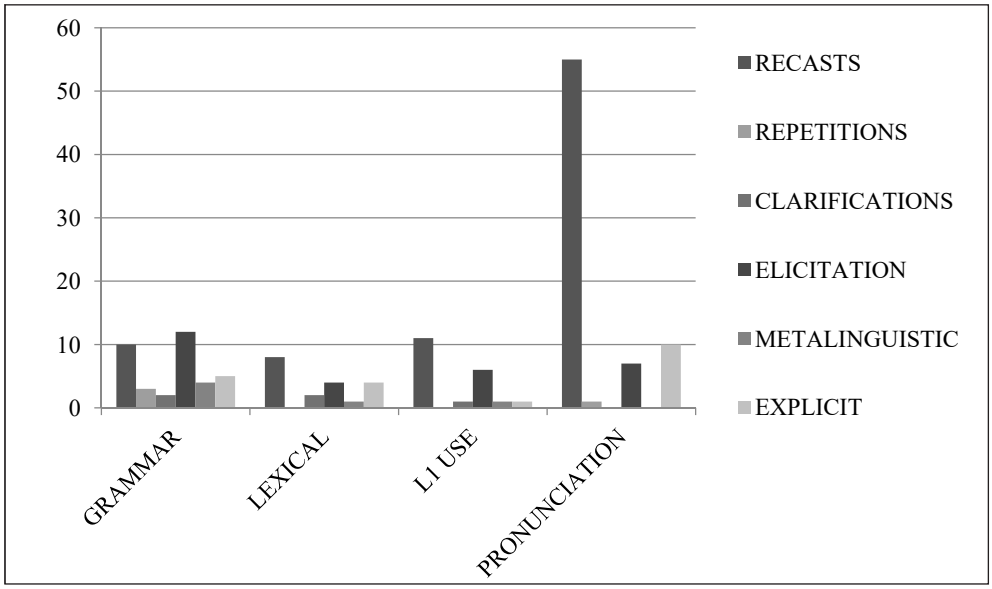
As for the OCF types used for each of the types of error, we predicted that recasts would be used to address phonological or lexical errors while prompts such as elicitation or metalinguistic cues would be chosen for grammar errors, as in previous research (Lyster & Ranta, 1997; Lyster & Mori, 2006). On the basis of the main focus of the lessons, we predicted that the EFL teacher would address grammar errors more often and by means of prompts, while the CLIL teacher was expected to show more concern for vocabulary errors and use more implicit types of OCF such as recasts. Our findings showed that the type of error had an impact on the OCF types used and significant differences were found both for each of the teachers (Chi-square test: CLIL p= 0; EFL p=0) as well as between the teachers, as explained in what follows.

Regarding grammar errors, Graphs 3 and 4 show that the CLIL teacher almost exclusively selected recasts to address this error type, while the EFL teacher used the whole spectrum of OCF types (Fisher test, p=.002). As for lexical errors, more variation can be observed in the OCF types preferred by the CLIL teacher, which is coherent with his meaning-oriented lessons. These types of meaning errors were more important for him and thus he addressed them more carefully. This concern for meaning over form is shown in the data and it was also expressed by the teacher himself in informal conversations and in a questionnaire he completed for a follow-up study (Milla & García Mayo, in press). No significant differences can be reported in the use of OCF types for this type of error (p=.073) in the two contexts. The use of OCF types was found to be significantly different regarding pronunciation errors (p=.047), given that, even though the preferred OCF type was recast, the EFL teacher also resorted to other OCF types such as explicit correction or elicitation, while the few instances of OCF for pronunciation errors in CLIL, were recasts except for one case of repetition. L1 use was mainly corrected by recasts by both teachers, with very little attention given to this type of error by the CLIL teacher in spite of the great amount of L1 use in this classroom. Significant differences were found regarding the teachers' behavior towards L1 use (Fisher test, p=.032). Analyses performed without tallying L1 use showed significant differences as well, both in the proportion of correction by each of the teachers of each of the error types (CLIL p= 0; EFL p= 0) and in the comparison between the two teachers (p=.007).

Graph 3: Total number of OCF types depending on error type in CLIL



Graph 4: Total number of OCF types depending on error type in EFL



Data were triangulated by performing a qualitative analysis, which helped us to capture the nuances that were not reflected in the quantitative analyses, as had been previously done in comparative classroom observation studies (Milla & García Mayo,

2014). We found that the manner in which the teachers were using the types of OCF influenced the effect of these types, i.e. learners' uptake. For example, the EFL teacher used multiple feedback moves, which are said to lead to higher uptake, as explained above.

Looking at the data from a qualitative perspective we found that, although both teachers used recasts as the most frequent OCF type, these recasts were of a different nature. The EFL teacher used more explicit, shorter, and therefore more salient recasts, especially for pronunciation errors, as illustrated in example 5, while the recasts used by the CLIL teacher were more implicit or conversational in many cases. The CLIL teacher continued the topic after most of the recasts, preventing learners from acknowledging the correction or repairing the errors.

(5) Recast of an explicit type in EFL

Learner:	a moth lands on your *forehead [fɔ:head]...
Teacher:	forehead [fɔ:hed].
Learner:	forehead [fɔ:hed]. And then you hear laughter in the audience.

The CLIL teacher used a massive amount of recasts (71%) and very rarely resorted to other types of OCF, unlike the EFL teacher, who made use of elicitation and metalinguistic information, as well a great number of recasts. The CLIL teacher's use of recasts is illustrated in example 6, where he reformulates a grammar error but then continues with the topic:

(6) Recast with topic continuation in CLIL

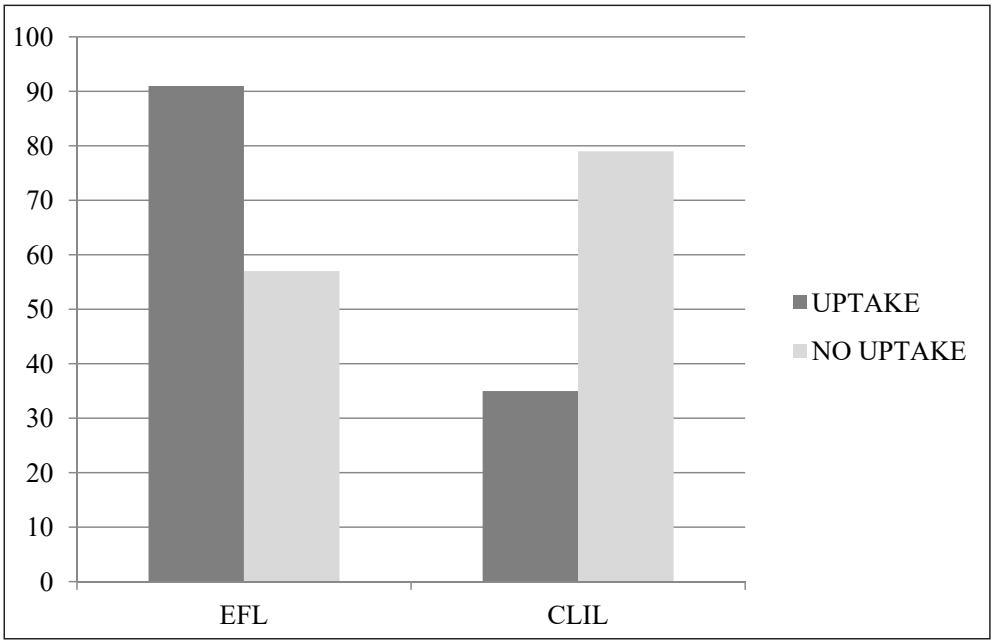
Learner:	*between departments...
Teacher:	OK, for example among the departments**.** When we meet all the departments, horizontal communication is in the same level of authority. OK, it is clear? That, we are going to see in a next day.

Finally, a remarkable number of CFEs with multiple feedback (12 out of 157 moves, 8%) was found in EFL, whereas only one instance (out of 120, 0.8%) was identified in CLIL. As seen above, the EFL teacher corrected a large amount of the errors and used the whole spectrum of types, the most explicit types and prompts quite frequently, while the CLIL teacher corrected only a small proportion of the errors and used mainly recasts. In this sense, this difference in the use of multiple feedback moves was expected since this technique is typical of teachers who are more concerned with accuracy.

The third research question wondered whether OCF would lead to more uptake in the CLIL or the EFL classroom, and whether learners would react differently to

various OCF types. On the basis of previous research, we expected higher rates of uptake in EFL due to the orientation to form in this type of lesson as well as the more salient OCF types provided. Graph 5 displays the findings of the Chi-square test, which showed that the learners' response was found to be different after OCF, with a significantly higher proportion of uptake (p=0) in EFL (62%) than in CLIL (32%). This finding could not be observed in previous comparative studies because the groups of students compared were different. The types of uptake were also significantly different (Fisher test, p=0), as more self-repair was found in EFL (56 out of 152 CFEs, 37%) than in CLIL (16 out of 116 CFEs, 14%).

Graph 5: Total number of uptake and no uptake moves in EFL and CLIL classrooms



The analysis of the uptake of the different OCF types showed that there were higher rates of uptake after recasts, elicitations and explicit correction moves in EFL but only significant differences were found for recasts (Chi-square test, p=0). Consequently, reformulations were also found to lead to higher uptake in EFL (p=0). Table 3 below displays these results:

Table 3: Total number and percentages of uptake moves to the different OCF types in CLIL and EFL

OCF TYPES	EFL	CLIL
RECASTS	42 (46.1%)	5 (14.4%)
CLARIFICATION REQUESTS	5 (5.5%)	6 (17.1%)
REPETITIONS	4 (4.4%)	6 (17.1%)
METALINGUISTIC CUES	27 (29.7%)	10 (28.6%)
ELICITATION	4 (4.4%)	7 (20%)
EXPLICIT CORRECTION	9 (9.9%)	1 (2.8%)
REFORMULATIONS	51 (56%)	6 (17.1%)
PROMPTS	40 (44%)	29 (82.9%)
TOTAL UPTAKE MOVES	91	35

Regarding multiple feedback, we found that a combination of CF types led to successful repair in 67% of the cases (8 out of 12), mainly after elicitations. Uptake was high (75%), probably because all multiple feedback types are by nature very explicit and, thus, salient.

3.5. Discussion

In this study we compared the behaviour regarding OCF of a CLIL teacher and an EFL teacher working with the same group of students and the students' reaction to the OCF provided. After the analysis of the lesson orientation in each of the classes, it was observed that EFL lessons were predominantly focused on language form, while CLIL lessons were oriented to meaning. The EFL teacher's behaviour was in line with the lesson orientation by showing great concern for learners' accuracy. She used a high proportion of CF and employed the whole spectrum of CF types with techniques such as multiple feedback, prompts, or explicit (didactic) recasts in order to promote learners' noticing of CF and self-repair, which she apparently achieved if one considers the high proportion of uptake in the EFL classroom. On the contrary, the CLIL teacher advocates for more implicit types of CF and addresses a very limited proportion of errors, which is coherent with the meaning-oriented context of his lessons. As seen above, a great deal of the errors in CLIL were related to the use of the L1, which the teacher himself considered inappropriate, as he often commented in the lessons. However, he chose not to address this type of error whenever learners were showing

fluency or content was being developed. In summary, there was a significant difference in the amount of errors addressed by the EFL teacher in comparison with her CLIL counterpart. As for differences in the type of OCF provided, both teachers preferred the use of recasts but whereas the EFL teacher used both types of reformulations, namely, recasts and explicit correction, together with some prompts such as elicitation, the CLIL teacher overwhelmingly used implicit (conversational) recasts.

These findings are in line with previous studies in FL settings, where teachers are more focused on form (Lochtman, 2007; Lyster & Mori, 2006) and use more explicit types of recasts and a higher amount of prompts. However, since there are no previous studies of OCF in secondary education levels in CLIL classrooms, we cannot establish a comparison with previous research. Llinares & Lyster (2014) included CLIL classrooms at primary education level, where CLIL teachers are typically language specialists, and hence, more concerned with language accuracy than the CLIL teachers in secondary education, who are subject specialists with much less linguistic background. CLIL teachers in this study showed a similar behaviour to those in SL settings, where attention to fluency and content is preferred and OCF is less frequently used. If OCF is provided, recasts are their choice.

Regarding the issue of whether the type of error would have an impact on the quantity and quality of OCF, our findings showed that the most striking differences between the EFL and the CLIL teachers were related to how they addressed grammar, pronunciation and L1 errors. Thus, the EFL teacher paid much more attention to all those types of errors, especially to grammar (as is usually the case in FL settings, Brown (2016), whereas the CLIL teacher focused his concerns on lexical errors, again in line with the focus on meaning of his lessons. As for the type of OCF used for each of the errors, recasts were again the type of OCF type preferred by both teachers but there is an important difference: whereas the CLIL teacher used implicit (conversational) recasts, the EFL teacher used more explicit (didactic) recasts and, what is more, she displayed multiple feedback moves, other types of reformulation such as explicit correction, and prompts such as elicitation. Recasts were also preferred by the two teachers to address pronunciation errors, which has been found to be beneficial as the learners can make comparisons of the erroneous and the target form (Lyster & Saito, 2010; Mackey, Gass & McDonough, 2000; Sheen, 2006). The fact that recasts are the preferred type of OCF by both teachers is in line with previous research, as illustrated in Brown's (2016) meta-analysis of 28 studies. Brown reported that recasts accounted for 57% of all CF provided while prompts comprised 30%.

As for OCF and learner uptake in the two types of lessons, the focus of our third research question, it seems that OCF was more effective in the EFL classroom where learners displayed a higher proportion of uptake than in the CLIL classroom.

Significant differences regarding uptake in the two classrooms were only found for recasts, that is, learners benefitted more from the use of recasts in the EFL classroom than in the CLIL classroom. This is in line with Lyster & Mori's (2006) Counterbalance Hypothesis, where they suggested that the use of more implicit types, such as recasts, are sufficient to obtain successful uptake in contexts where the learners' focus is on form, such as the EFL classroom in the present study. Besides, in more meaning-oriented classrooms, more explicit and output pushing OCF types would be preferable, in order to obtain learners' attention, since their focus in this type of classrooms is not on language form. This may account for the relatively high success of prompts in our CLIL classrooms.

However, in spite of the salience of some CF types, a few instances of multiple feedback moves that were not acknowledged by the learners were attested. Multiple feedback moves are clearly salient and in our study most of them very explicit. Therefore, our findings suggest that explicitness is not a guarantee for uptake and that eliciting the correct form is more effective if teachers are seeking immediate repair. Consequently, given the present results, successful uptake would be obtained by prompts more than by recasts, particularly in meaning-oriented classrooms.

This study has reported that teachers correct differently in EFL and CLIL lessons, possibly influenced by the lesson orientation to form or meaning, respectively, as well as by the teachers' academic background, particularly, previous training in FL teaching. As shown above, learners act accordingly, displaying a different behaviour in EFL and CLIL lessons. In EFL they do not make as many errors and they hardly ever make use of their L1 while in CLIL lessons, where they are focused on content, they show less concern for language form and resort to their L1 very often in order to try to communicate ideas. We were able to find this different behaviour on the part of the learners depending on the type of setting they were immersed in, showing a clear influence of the context on all the participants of the CFEs. This change in the learners' behaviour could not have been analysed so precisely in previous studies, since the groups of learners in the comparisons belonged to different schools and even different countries (Llinares & Lyster, 2014; Lyster & Mori, 2006; Sheen, 2004).

4. Conclusion

The aim of the present study was to analyze CFEs in two learning contexts, EFL and CLIL, and explore the potential of this variable on the teachers' OCF choices as well as on the learners' reaction to those choices (uptake). One of the main findings reported is that the difference in lesson orientation in CLIL and EFL contexts influences not only the teachers' amount and types of OCF provided but also the learners' behavior,

with different types and amount of errors and variations in the uptake depending on the lesson they are attending. Previous studies could not have analyzed this change in learners' behavior so precisely because the participants in the comparison groups belonged to different schools and sometimes to different countries. We have found that CFEs in the CLIL and EFL settings analyzed are different in number and in nature. Moreover, in this study we have found that, as predicted, CLIL classrooms in secondary education are oriented to meaning and not to form as was the case in primary education (Llinares & Lyster, 2014), probably due to the teachers' lack of linguistic background in our setting, since they are subject specialists and not language teachers.

There are several pedagogical implications deriving from the findings of the study. First, we suggest that CLIL teachers should try to strike a balance between form and meaning in their lessons and they would probably benefit from training on this matter (Lo, 2019). That is, attention to form should be considered in CLIL settings as well if the aim of CLIL programs is to foster second language acquisition together with the acquisition of content knowledge. Moreover, we believe that collaboration between teachers and researchers should be encouraged. Teachers should be involved in research and informed of the findings, as they might be unaware of their own practices. Research findings should go beyond the limits of the academic world and reach educators and policy makers so that they can make informed decisions.

The study has shortcomings that should be acknowledged. It was located in a very specific geographical area and in a particular school, so our findings, however interesting they might be, have to be taken cautiously and cannot be generalized. Future research in secondary education CLIL should include other content areas and use a larger sample. Moreover, written CF should be considered in further research in order to see whether the differences between the contexts found in OCF can also be identified in the written mode. Finally, teachers' and learners' perspectives on CF should be explored in order to assess their potential impact on classroom behavior (Brown, 2009; Kartchava, 2016; Kartchava & Ammar, 2014b).

Acknowledgements

The authors gratefully acknowledge funding from research grant IT904-16 from the Basque Government. We are extremely grateful to Botikazahar high school for participating in the study and especially to the teachers Susana Hernández and Iñaki Valencia for allowing data collection in their classrooms.

5. References

Ammar, A. & Spada, N. (2006). One size fits all? Recasts, prompts, and L2 learning. *Studies in Second Language Acquisition,* 28, 543–574.

Antón, M., & DiCamilla, F.J. (1998). Socio-cognitive functions of L1 collaborative interaction in the L2 classroom. *Canadian Modern Language Review*, 54, 314–342.

Basturkmen, H. (2012). Review of research into the correspondence between language teachers' stated beliefs and practices. *System,* 40 (2), 282–295.

Bryfonski, L. & Ma, X. (2020). Effects of implicit versus explicit corrective feedback on Mandarin tone acquisition in an SCMC learning environment. *Studies in Second Language Acquisition, 42*, 61-88.

Brown, A. (2009). Students' and teachers' perceptions of effective foreign language teaching: A comparison of ideals. *The Modern Language Journal* 93, 46-60.

Brown, D. (2016). The type and linguistic foci of oral corrective feedback in the L2 classroom: A meta-analysis. *Language Teaching Research*, 20(4), 436-458.

Council of Europe (2001). *Common European Framework of Reference for Languages*. Cambridge: Cambridge University Press.

Dalton-Puffer, C. (2011). Content and language integrated learning: from practice to principles. *Annual Review of Applied Linguistics,* 31, 182–204.

Dalton-Puffer, C. & Nikula, T. (2014). Content and language integrated learning (guest editorial). *The Language Learning Journal,* 42(2), 117-122.

García, O. (2019). Translanguaging: A coda to the code? *Classroom Discourse,* 10, 369–373. Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven SL learning. *Studies in Second Language Acquisition,* 34, 445–474.

Goo, J. & Mackey, A. (2013). The case against the case against recasts. *Studies in Second Language Acquisition*, 35, 127–65.

Gurzynski-Weiss, L. (2010). *Factors Influencing Oral Corrective Feedback Provision in the Spanish Foreign Language Classroom: Investigating Instructor Native/Nonnative Speaker Status, Second Language Acquisition Education and Teaching Experience*. Unpublished PhD dissertation, Georgetown University

Havranek, G. & Cesnik, H. (2001). Factors affecting the success of corrective feedback. In S. Foster- Cohen & A. Nizegorodzew (Eds.), *EUROSLA Yearbook*. Volume 1 (pp. 99–122). Amsterdam: Benjamins.

Kartchava, E. (2016). Learners' beliefs about corrective feedback in the language classroom: Perspectives from two international contexts. *TESL Canada Journal/Review TESL du Canada,* 33, 19-45.

Kartchava, E. & Ammar, A. (2014). The noticeability and effectiveness of corrective feedback in relation to target type. *Language Teaching Research,* 18 (4), 428–452.

Lasagabaster, D. & Sierra, J.M. (2010). Immersion and CLIL in English: more differences than similarities. *ELT Journal,* 64, 376–395.

Li, S. (2014). Oral corrective feedback. *ELT Journal Volume,* 68 (2), 196–198.

Li, S. (2018). Corrective feedback in L2 speech production. In J. Liontas et al. (Eds.), *The TESOL encyclopedia of English language teaching*. London: Blackwell.

Llinares, A. & Lyster, R. (2014). The influence of context on patterns of corrective feedback and learner uptake: a comparison of CLIL and immersion classrooms. *The Language Learning Journal,* 42 (2), 181–194.

Lo, Y. L. (2019). Development of the beliefs and language awareness of content subject teachers in CLIL: does professional development help? *International Journal of Bilingual Education and Bilingualism,* 22 (7), 818–823.

Lochtman, K. (2002). Oral corrective feedback in the foreign language classroom: how it affects interaction in analytic foreign language teaching. *International Journal of Educational Research,* 37, 271–283.

Lochtman, K. (2007). Die mündliche Fehlerkorrektur in CLIL und im traditionellen Fremdsprachenunterricht: Ein Vergleich. In C. Dalton-Puffer & U. Smit (Eds.), *Empirical Perspectives on CLIL Classroom Discourse* (pp. 119–138). Frankfurt: Peter Lang.

Loewen, S. (2004). Uptake in incidental focus on form in meaning-focused ESL lessons. *Language Learning,* 54 (1), 153–188.

Long, M. H. (2015). *Second language acquisition and task-based language teaching*. Oxford, England: Wiley-Blackwell.

Lyster, R. (2002). Negotiation in immersion teacher-learner interaction. *International Journal of Educational Research,* 37, 237–253.

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. *Studies in Second Language Acquisition,* 26 (3), 399–432.

Lyster, R. & H. Mori (2006). Interactional feedback and instructional counterbalance. *Studies in Second Language Acquisition,* 28 (2), 269–300.

Lyster, R., & Ranta, L. (1997). Corrective feedback and learner uptake: Negotiation of form in communicative classrooms. *Studies in Second Language Acquisition,* 19, 37–66.

Lyster, R. & Saito, K. (2010). Oral feedback in classroom SLA. A meta-analysis. *Studies in Second Language Acquisition,* 32, 265–302.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback. *Language Teaching,* 46 (1), 1–40.

Mackey, A. & Goo, J. (2007). Interaction research in SLA: A meta-analysis and research synthesis. In A. Mackey (Ed.), *Conversational Interaction in Second Language Acquisition* (pp. 407–472). Oxford: Oxford University Press.

Mackey, A., Gass, S., & McDonough, K. (2000). How do learners perceive interactional feedback? *Studies in Second Language Acquisition,* 22, 471–497.

McWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk.* 3rd edition. Mahwah, N. J. Lawrence Earlbaurn Associates.

Milla, R. & García Mayo, M. P. (2014). Corrective feedback episodes in oral interaction: A comparison of a CLIL and an EFL classroom. *International Journal of English Studies,* 14(1): 1-20.

Milla, R. & García Mayo, M. P. (in press). Teachers' and learners' beliefs about corrective feedback compared with classroom behaviour in CLIL and EFL. In Talbot, K., Mercer, S., Gruber, M.-T. & Nishida, R. (Eds.), *The Psychological Experience of Integrating Language and Content*. UK: Multilingual Matters.

Mori, R. (2002). Teachers' beliefs and corrective feedback. *JALT Journal,* 24 (1), 48–69.

Nassaji, H. (2020). Assessing the effectiveness of interactional feedback for L2 acquisition: Issues and challenges. *Language Teaching,* 53, 3-28.

Nassaji, H. & Kartchava, E. (eds.). *The Cambridge Handbook of Corrective Feedback in Language Learning and Teaching*. Cambridge: Cambridge University Press (forthcoming).

Oliver, R. & Grote, E. (2010). The provision and uptake of different types of recasts in child and adult ESL learners: What is the role of age and context? *Australian Review of Applied Linguistics,* 33 (3): 26.1–26.22.

Panova, I. & Lyster, R. (2002). Patterns of corrective feedback and uptake in an adult ESL classroom. *TESOL Quarterly,* 36, 573–595.

Pérez Cañado, M. L. (2012). CLIL research in Europe: Past, present and future. *International Journal of Bilingual Education and Bilingualism,* 15, 315-341.

R Development Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Russell, V.J. & Spada, N. (2006). The effectiveness of corrective feedback for second language acquisition: A meta-analysis of the research. In J. Norris & L. Ortega (Eds.), *Synthesizing Research on Language Learning and Teaching* (pp. 133–164). Amsterdam: John Benjamins.

Saito, K. (2013). The acquisitional value of recasts in instructed second language speech learning: Teaching the perception and production of English /ɹ/ to adult Japanese learners. *Language Learning,* 63 (3), 499-529.

Saito, K. & Lyster, R. (2012). Effects of form-focused instruction and corrective feedback on L2 pronunciation development of / ɹ / by Japanese learners of English. *Language Learning,* 62, 595–633.

Schmidt, R. (1990). The role of consciousness in L2 learning. *Applied Linguistics,* 11, 129–158.

Sheen, Y. (2004). Corrective feedback and learner uptake in communicative classrooms across instructional settings. *Language Teaching Research,* 8(3), 263–300.

Sheen, Y. (2006). Exploring the relationship between characteristics of recasts and learner uptake. *Language Teaching Research,* 10, 361–392.

Sheen, Y. (2011). *Corrective Feedback, Individual Differences and Second Language Learning*. New York: Springer.

Spada, N. & Fröhlich, M. (1995). *COLT. Communicative Orientation of Language Teaching Observation Scheme: Coding Conventions and Applications*. Sydney, Australia: National Centre for English Language Teaching and Research.

Tedick, D. J. & Cammarata, L. (2012). Content and language integration in K-12 contexts: Learner outcomes, teacher practices and stakeholder perspectives. *Foreign Language Annals,* 45 (1), 28–53.

Yang, J. (2016). Learners' oral corrective feedback preferences in relation to their cultural background, proficiency level and type of error. *System*, 61, 75-86.

Yang, Y. & Lyster, R. (2010). Effects of form-focused practice and feedback on Chinese EFL learners' acquisition of regular and irregular past tense forms. *Studies in Second Language Acquisition,* 32, 235–263.

Yilmaz, Y. (2012). The relative effects of explicit correction and recasts on two target structures via two communication modes. *Language Learning,* 62, 1134–1169.

Bayesian Vocabulary Tests

Paul M. Meara
Lognostics, Cardiff, UK
p.m.meara@gmail.com

Imma Miralpeix
Department of English Studies, Universitat de Barcelona, Spain
imiralpeix@ub.edu

Abstract

This paper proposes a new way of looking at productive vocabulary in L1 and L2 speakers. An experiment was conducted where 160 participants provided six words for five different picture prompts they were presented with. Data from this minimal vocabulary test was analysed using Bayesian statistics in order to decide whether a set of responses were generated by an L1 speaker or by an L2 advanced learner. Results obtained provide some interesting insights into the viability of minimal vocabulary tests (small sets of words can carry large amounts of information on vocabulary use), as well as some indications of how Bayesian methods could help us explore productive vocabularies of L2 speakers at different proficiency levels.

Keywords: Bayesian statistics, L2 learning, minimal vocabulary tests, productive vocabulary, vocabulary testing.

Resumen

Este artículo propone una nueva forma de considerar el vocabulario productivo en hablantes nativos y aprendices de segundas lenguas. Se realizó un experimento en el que 160 participantes proporcionaron seis palabras para cinco imágenes diferentes que se les presentaron. Los datos de esta prueba mínima de vocabulario se analizaron utilizando estadística bayesiana para decidir si un conjunto de respuestas fue generado por un hablante nativo o por un aprendiz de nivel avanzado. Los resultados obtenidos ofrecen ideas interesantes sobre la viabilidad de las pruebas mínimas de vocabulario (un pequeño número de palabras puede proporcionar gran cantidad de información sobre el uso del vocabulario), así como algunas indicaciones de cómo los métodos

bayesianos podrían ayudarnos a explorar los vocabularios productivos de hablantes de idiomas a distintos niveles de competencia.

Palabras clave: Estadística bayesiana, aprendizaje de segundas lenguas, pruebas mínimas de vocabulario, vocabulario productivo, test de vocabulario.

1. Introduction

In L2 vocabulary research, a distinction between receptive and productive vocabulary knowledge has often been made: we usually assume that receptive vocabulary involves being able to recognize and understand a word when it is encountered in listening or reading, while productive vocabulary means being able to use it in speech or writing. There is also a general agreement by the research community that receptive vocabularies tend to be bigger than productive vocabularies, as reception precedes production (e.g. see Melka, 1997; Webb, 2008). Receptive vocabulary size has been object of study for a very long time and several tests have been proposed to measure this dimension, namely multiple choice tests, for example the *Vocabulary Levels Test* (VLT: Nation, 1990; Schmitt, Schmitt, & Clapham, 2001; Webb, Sasao, & Balance, 2017) and the *Vocabulary Size Test* (VST: Nation & Beglar 2007; Coxhead, Nation, & Sim, 2014) or yes/no tests (Meara & Buxton, 1987) such as *V_YesNo* (Meara & Miralpeix, 2015b). However, productive vocabulary size has been largely unexplored, which is partly due to the need for new assessment methods. Different approaches have been taken in an attempt to measure productive vocabulary size. So far, (1) we are able to describe the sort of vocabulary learners produce in a speaking or writing task; (2) we can measure 'controlled productive vocabulary size' when learners are asked to provide a specific word given its first letters; (3) we can derive scores from word associations tests or lexical availability tasks that may give an idea of how big vocabularies are and (4) we can use mathematical methods typical of other fields, such as biology, to estimate the amount of words students of a language may know productively. Nevertheless, these approaches have also raised several concerns, as we will see in sections 1.1-1.5 below. Therefore, in section 2 we propose a new method to explore productive vocabulary sizes following Bayes' theorem. Our aim in this paper is to evaluate the Bayesian approach by analysing its potential for distinguishing between L1 and L2 speakers on the basis of very few words, which were produced to describe a set of picture prompts (sections 3-5).

1.1. The Lexical Frequency Profile

One of the first attempts to characterise learners' productive vocabulary is Laufer and Nation's Lexical Frequency Profile -LFP- (Laufer & Nation, 1995). The operation

of the LFP is essentially very simple: LFP takes a raw text as input and returns as output a profile of the text in terms of the frequency distribution of its words. Laufer and Nation suggest that a profile based on four frequency categories is useful - the four categories being based on Nation's earlier work on word lists for L2 learners (Nation, 1984). Category 1 consists of the 1000 most frequent words in English as defined by Nation's lists; category 2 consists in the second 1000 most frequent words; category 3 consists of words in the University Word List (Xue & Nation, 1984); category 4 includes any word not found in the previous three lists. It should be acknowledged, though, that Laufer and Nation's profiles are not particularly easy to work with: they describe a learner's output as a four point profile, rather than as a single measure, and it is difficult to summarise the data that they encapsulate in an economical and transparent way. Furthermore, rather than predicting testees' lexical proficiency, the LFP describes the kind of words testees use in any piece of oral or written data.

1.2. Controlled productive vocabulary

Laufer (1998) makes a distinction between controlled productive vocabulary and free productive vocabulary. She defined a test of controlled productive vocabulary as one that "entails producing words prompted by a task" (e.g. when the first two letters of a word in the context of a sentence are provided to the student), whereas free productive vocabulary "has to do with using words at one's free will, without any specific prompts for particular words" (1998: 257). Laufer and Nation (1999) developed a test to assess the former, by using the same words as in the VLT receptive vocabulary test. In this case, the test-takers responses are constrained by providing the first few letters of the expected response, as in the example below:

The house was su_____ by a big garden. *(surrounded)*

However, assessing free productive vocabulary is much more challenging, as research on the topic has clearly evidenced.

1.3. Vocabulary size from word association and lexical availability tasks

Although word association tests (Meara & Fitzpatrick, 2000) or lexical availability tasks (Roghani & Milton, 2017) were not first devised for this purpose, data from these tasks typically consist of L2 words that could be profiled using standard vocabulary assessment tools such as Range (Heatley, Nation & Coxhead, 2002). For example, in lexical availability tasks learners are asked to name as many words as they can from a prescribed category, such as *food*, *parts of the body*, *animals* or *transport* (Jiménez Catalán, 2014). The learner profiles obtained from learners' answers could provide a picture of

the scope of a testee's productive vocabulary, as shown above. However, more research is needed on the extent to which these profiles might be good indicators of productive vocabulary knowledge (Fitzpatrick & Clenton, 2017).

1.4. The capture-recapture method: V_Capture

V_Capture is a computer program based on the idea developed by biologists interested in counting the number of species in a particular area by capturing and then recapturing animals in traps on a number of different occasions. This process is similar to comparing the words produced in a particular task over several performances. Mathematics uses the proportion of animals captured (in the case of vocabulary that would equal words used) on both occasions to estimate the number of animals or species being studied (Meara & Olmos Alcoy, 2010). In spite of the fact that the program computes Petersen Estimates (and thus provides an estimate of vocabulary size), there are several problems with these estimates for continuous texts, and more plausible results can be obtained from wordlists rather than from texts (Meara & Miralpeix, 2017). Interesting work on the capture-recapture method can also be found in Williams, Segalowitz and Leclair (2014).

1.5. A productive vocabulary size estimator: V_Size

As it is problematic to assess total productive vocabulary size, it may be more suitable to compare relative vocabulary sizes, such as the vocabulary someone uses for a particular task compared to what others (e.g., NS or learners at different proficiency levels) use when performing the same tasks. A range of different tasks, such as cartoon storytelling or picture description, may be needed for this purpose, and tools using different estimation methods can help researchers obtain reliable estimates. Estimates by *V_Size* (Meara & Miralpeix, 2015a) are based on the Power Law, a ranked distribution found not just in language but also in other physical and biological phenomena like earthquake size or social network connectivity. The program assumes that certain words in language (in high frequency bands) are more frequent than others (in low frequency bands) and that there is a direct relationship between the number of times a word occurs in a corpus and its rank in a frequency list generated by the corpus. Thus, it allows researchers to go beyond the mere shape of the frequency profile generated by a text and enquires into what the profile tells us about the size of the productive vocabulary of the person who produced the text. As noted by Castañeda-Jiménez and Jarvis (2014: 501), *V_Size* "is the only freely available computer program we [the authors] know of that outputs estimates of learners' productive vocabulary based on the texts they produce".

While *V_Size* may give us good indications in the future of the vocabulary known productively for certain tasks, it would also be very useful for us to determine learners' proficiency level from a sample of words they know productively. A first step in this direction would be trying to distinguish between native speakers (NS) and advanced learners on the basis of very few words, which is what we will try to do in this paper.

2. Bayes theorem and its applicability to vocabulary assessment

As noted by Miralpeix (2020), among others, assessing productive vocabulary always involves eliciting a set of words from learners and inferring from this sample the size of the learners' repertoire, i.e. how many words they would be able to retrieve from memory without seeing them written or hearing the spoken forms. All estimations are based on probabilities, as it is impossible to elicit from learners all the words they know productively in an L2 (unless they are at the very first stages of learning a language and know very few words).

Up to now, the mathematical procedures that we have used for productive vocabulary measurement have relied on analysing learners' data using proportions (e.g. in the capture-recapture method, see section 1.4) or comparisons of rank distributions with curve-fitting (e.g. in *V_Size*, see section 1.5). It has also been observed that the more a learner speaks (or writes), the more capable we are of making inferences about his/her lexical knowledge, as every word introduces new information that can help us make guesses about his/her level. It would be really useful if we could formalise guesses of this type and one way of doing this is to apply Bayesian statistics to the data.

The immediate background to the work we present in this paper is an earlier study (Meara & Miralpeix, 2017) in which we asked L1 Spanish and Catalan learners of English to generate a set of ten adjectives in response to a cartoon stimulus like the one shown in Figure 1. This apparently simple task turned out to be quite difficult, even for advanced learners.

Figure 1: The cartoon figure used in Meara & Miralpeix (2017).

The main focus of our research at the time was how similar learners' response sets were. We will not discuss this work here other than to say that the cartoons generated a very large number of disparate responses, and that a response set generated by advanced L2 speakers would typically share just over two words with another response set from the same group of participants. Some examples are provided in Table 1. These responses were made by first language Catalan or Spanish learners of English.

Table 1: Some example response sets made to Figure 1.

NNS001,slim,intelligent,serious,tall,big_headed,angry,lonely,bad_tempered,strict,helpful
NNS002,smart,formal,strict,serious,slim,bad_tempered,intelligent,thoughtful,impatient,rude
NNS003,big_headed,tall,ugly,slim,bald,serious,shy,open_minded,young,impartial
NNS004,ugly,short,big_headed,evil,strange,bad_tempered,scary,angry,lonely,moody
NNS005,grumpy,surprised,big_headed,lunatic,creepy,dirty,stinky,weird,adult,ugly

While working with these data, we noticed that we were often able to decide whether a set of responses was generated by an L2 learner or an L1 speaker. Obviously, the L2 speakers sometimes produced responses that were easily identifiable as learner

errors, but leaving responses of this kind aside, we noticed that some words were more likely to be used by learners than by native speakers, and vice versa. Some words were almost exclusively used by one group rather than the other, while other words were somewhat more likely to be used by one group. For the stimulus picture in Figure 1, the two groups generated a total of 650 different words, which can be divided into three categories: words used by both groups of participants (shared words: 30%), words used predominantly by the L1 participants (L1 words: 46%) and words used predominantly by the L2 participants (L2 words: 24%). With practice, one becomes fairly good at guessing whether a data set was generated by an L1 speaker or and L2 learner.

One way to formalise this intuition is by means of Bayes' Rule (McGrayne 2011). Bayes' Rule is a mathematical procedure which allows us to change our estimate of something being true, in the light of additional evidence. This approach has not been popular in language acquisition studies, although recently Norouzian et al. (2018) introduced the application of Bayesian methods to various research designs, and Pearl and Goldwater (2016) and Zinszer et al. (2018) used Bayesian inference models to analyse L1 acquisition. Bayes theorem has mostly been used in situations where data is partial and difficult to interpret - naval searches, medical diagnoses, face recognition and spam filtering, to name but a few. This last example is particularly interesting for us, as the best spam filters rely on a comparison of the words typically used in spam emails, and the words used in bona fide emails - a comparison that is not a million miles away from the problem we are faced with when we try to assess the vocabularies of L2 speakers.

The approach is usually described as follows:

$$P(A|B) = [P(B|A)\ P(A)] / P(B)$$

Which tells us:	how often event A happens *given that B happens*, written **P(A|B)**,
When we know:	how often event B happens *given that A happens*, written **P(B|A)**
	and how likely A is on its own, written **P(A)**
	and how likely B is on its own, written **P(B)**

Although the approach has not been used in vocabulary assessment, Bayes' rule could help us predict proficiency levels from peoples' productive vocabularies. For instance, it can give us information on the chance a set of words being really from an L2 learner or a NS taking into account the words that they produce in a test, as there are words with higher chances of appearing in L2 learners' sets than in NS sets. In this case:

Our Event A: The set is produced by an L2 learner
Our Event B: The presence of certain words

Then:
P (L2 | words) = [P(words | L2) P(L2)] / P(words)

By making this filtering, based on updating probabilities, we could know, for example, if a set has an 85% chance of being produced by an L2 learner (then, it probably is) or if it has a 10% (then, it probably has been produced by a NS).

3. The study

3.1. Research question

In the light of the previous research on measuring productive vocabularies, it would be interesting to explore the potential of Bayesian statistics to correctly identify different proficiency levels from a set of words provided by participants at these levels. In this paper, the research question that we try to answer is the following:

How can a Bayesian approach help in distinguishing between NS and advanced English learners from a small set of words they provided for the same picture stimuli?

3.2. Method

3.2.1. Participants

Two groups of 100 test participants were assembled: a group of 18-21 year-old L1 English students at Swansea University (mean age 20.1), and a group of L1 Spanish/Catalan students of the same age range (mean 20.3) at an advanced level at the University of Barcelona. The final sample for the present study consists of 160 participants (80 per group). In the NS group there were 53 females and 27 males, and 59 females and 21 males in the L2 learners' group. These learners were in the third year of English Studies, a degree on English Linguistics and Literature taught in English since the first year. At the moment of data collection, they had a C1 level and a 30% of the sample had been abroad to English speaking countries for a month or less. Although no placement test was conducted for the purpose of this study, other cohorts at this level have been shown to have receptive vocabulary sizes between 6,500-7,000 words.

3.2.2. Instruments

A set of fifty cartoons was commissioned in the same style as the cartoon that appears in Figure 1. Ten of the cartoons depicted older men, ten depicted older women, ten depicted younger men, ten depicted younger women and ten depicted young children. Five cartoon pictures were selected: an older man, an older woman, a young man, a young woman and a child. The aim for selecting these five (shown in Figure 2) was that they looked as different as possible so that testees had a high chance to provide enough words for each, without repeating any.

Figure 2: The stimulus pictures used in the study.

3.2.3. Procedure

Participants were given the picture set and asked to provide six adjectives that might describe the person in each picture (no examples or training were provided). They were asked to write the words on the dotted lines after the prompts *'Neville is ...'*, *'Margaret is...'*, etc.

It should be noted that, in some ways, this approach is similar to that followed in the productive tests we mentioned earlier (e.g. lexical availability tasks), in the sense that we will be working with relative vocabulary sizes, i.e. the vocabulary someone uses for a particular task compared to what others use when performing the same task. Therefore, the approach will also be limited in the insight it provides about vocabulary proficiency 'in general'.

Not all participants managed to provide six answers for all the pictures (the L1 Ss in particular often produced a phrase rather than the single word that was requested). From the original 200 Ss, we managed to construct two sets of 80 Ss who generated six words for each of the five pictures. These groups were divided into two. For each L1 we

set up a group of 50 Ss whose data was used to establish a reference file for the group. The remaining 30 Ss were set aside to be used an evaluation group.

Next, for each reference group, we identified all the word types generated in their responses, and from this raw data we were able to identify word types used by both groups (shared words), words which were used only by the L1 group, words which were used only by the L2 group and singletons which occurred only once in the data set.

The question we then ask is whether these data can reliably predict the provenance of a new response set. In order to test this idea, we used the Bayesian approach described above to evaluate the 60 response sets that were left out of the analysis – 30 L1 speakers and 30 L2 speakers. For each response set, we estimated the probability of its being produced by an L1 speaker. Probabilities greater than 0.6 were taken to indicate that the data were produced by an L1 speaker, while probabilities below 0.4 were deemed to indicate that the data comes from an L2 speaker. Data sets where the final probability lies between 0.4 and 0.6 are deemed to be undetermined.

3.2.4. Data analysis

As Bayesian statistics is probably unfamiliar to most readers of this journal, we will explain the approach in some detail using as an example a data set of 10 words:

weird, unpredictable, old, angry, clever, worried, intense, interested, kind, thoughtful

Given raw data from suitable groups of participants, we can draw up a table which shows what we can expect of an L1 participant and an L2 participant in this task. In this particular case, we would expect about 75% of the responses generated by L1 participants to be shared words, about 20% of their responses to be L1 words and perhaps one of their responses to be a typical L2 response. For L2 speakers, we would expect 82% of their responses to be shared words, about 13% of their responses to be typical L2 words and perhaps one of their responses to be a typical L1 word. These response patterns look fairly similar (See Table 2), but taken together, the differences are large enough to allow us to ascribe an individual data set to the L1 group or the L2 group with a fair degree of confidence.

Table 2: The composition of the L1 and L2 data sets

	Shared Responses	L1 Responses	L2 Responses
L1 participants	75%	20%	5%
L2 participants	82%	5%	13%

Figure 3 shows how this works in practice. A given response set can either belong to an L1 speaker or an L2 speaker. Assuming that we do not know which answer is correct at the very beginning, we first assign both outcomes a probability of .5, as we start from a 50/50 hypothesis. Next, we look at word 1 in the response set (which is the word **weird** in example) and carry out the calculations shown in Table 3.

Table 3: Recomputing a probability in the light of new data

1. Find the appropriate column in Table 2. As an example we take **weird**, from the data set above. As it is an L2 word, so we work with the figures in column 3.

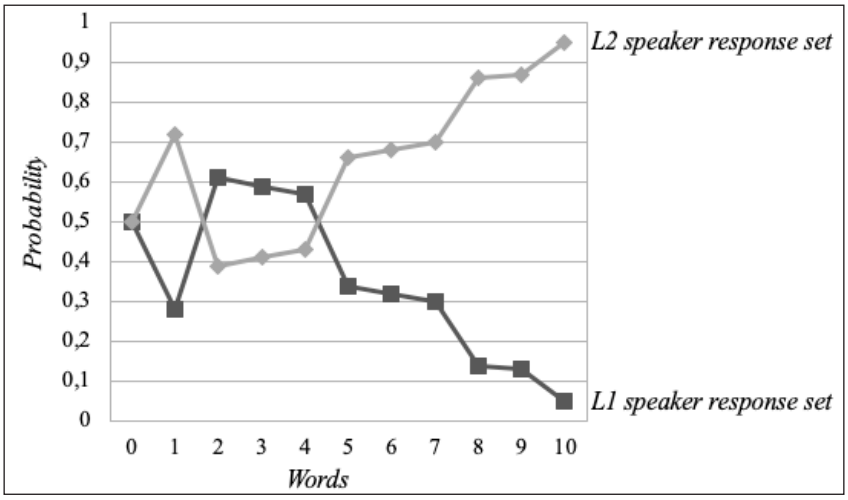
2. Multiply our current NS estimate by .05 .5*.05 = .025
3. Multiply our current NNS estimate by 0.13 .5*.13 = .065

4. Rescale the new probabilities so that they sum to 1. .025+.065 = .090

   new NS estimate: .025/.090 = .278
   new NNS estimate: .065/.090 = .722

Figure 3 shows, first of all, how the information provided by **weird** changes our assessment of whether this data set is generated by an L1 speaker or an L2 speaker: taking **weird** into account, it now seems slightly more likely that we are dealing with an L2 speaker.

Figure 3: Changes in confidence as more words are added to the data set



Next, we repeat the steps detailed in Table 3 using the new probabilities that resulted from step 3; that is, .278 (the new NS estimate) and .722 (the new NS estimate) instead of .5 (which was adopted for the first word in the data set). These steps are an implementation of *Bayes' Rule* (McGrayne, 2011; Stone, 2013). As our next word in the example set is **unpredictable**, which is a NS word, we use the probabilities in column 1 of Table 2. Applying the steps in Table 3, we get two new estimates: the L1 speaker estimate rises to .625 and the L2 speaker estimate falls to .375.

Finally, applying these steps to all ten of the words in the data set produces a convincing result: the probability that this response set is generated by an L2 participant is 0.95 (see Figure 3).

This is a pretty remarkable outcome. We started out with a mere sample of 10 words, generated to a simple cartoon, and we end up being 95% certain that the 10 words were generated by an L2 speaker. It is more than a little surprising that such a small data set can carry so much information, and allow us to make such confident assessments. It can also be observed in Figure 3 that by word 6 we start having a clear indication on whether the words were produced by an L1 or L2 speaker, that is why we opted for asking participants in the present study to provide six words for each stimulus. Therefore, these calculations were made using the six words in each of the sets that our participants produced.

3.3. Results

In this section we present the results obtained for each of the pictures used in the study. For each cartoon we provide some examples from the corpora we gathered, as well as the probabilities of response types according to the reference data (corpus), and a final assessment on the extent to which participants were classified as NS or learners using Bayesian statistics.

Table 4 shows the data for ***Shirley***, the cartoon of the young woman. Table 4a lists the words used to describe Shirley, divided into shared words, L1 responses, L2 responses and singleton responses.

Singleton responses make up a large proportion of the data (just over half the words fall into this category). Table 4b shows the probability of the different response types in the reference data set. Table 4c shows the way the test data are classified by the Shirley reference data set.

Table 4: Classified data for Shirley

Table 4a: Examples of words in the corpora for Shirley

Shared responses
SURPRISED NICE FREE ACTIVE SWEET CHEERFUL STYLISH THIN SINGLE POSITIVE WILD TALKATIVE LOVELY SEXY SMILING DANCER AFRO FRIENDLY HAIRY YOUNG CONFIDENT SKINNY FUNNY HAPPY SMILEY CURLY CHEEKY FASHIONABLE BLACK CRAZY PRETTY ENERGETIC SASSY TOOTHY DANCING EXCITED JOYFUL TALL FEMALE OUTGOING EXTROVERT COOL LIVELY

Typical L1 responses
ANNOYING DAME TEETH SINGER LOUD ENTHUSIASTIC JAZZY APPROACHABLE FUNKY HAIR SNAZZY BIG_HAIR BUBBLY SMUG TEETHY FUN

Typical L2 responses
BIG_MOUTHED HIGH POSH SHALLOW ATTRACTIVE SLENDER OPEN_MINDED UGLY SENSUAL GOOD_LOOKING SELF_CONFIDENT PLAYFUL FASHION CURLY_HAIRED SMART EXPRESSIVE ARTISTIC NERVOUS ELEGANT RICH SLIM BEAUTIFUL EXTROVERTED TRENDY WITTY

Table 4b: Probabilities of a specific response type in the Shirley reference data

	Shared responses	L1 responses	L2 responses	singletons
L1 participants	.223	.213	.050	.514
L2 participants	.194	.050	.247	.509

Table 4c: Discrimination between the 60 response sets for the Shirley reference file

Actual evaluation	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	23	6	1
L2 speaker response set	4	5	21

This cartoon is clearly very good at distinguishing between data sets generated by the two groups. Only one L1 participant is incorrectly classified, while 23 L1 speakers are correctly identified as such. For the L2 participants, the classifications are not quite so good: four L2 participants are incorrectly classified, but 21 are correctly identified as L2 speakers, with only five undecided. A chi squared analysis suggests that this distribution is very unlikely to have arisen by chance (χ^2=43.5; p<.001).

On the face of things, this looks like a very satisfactory result. Even when we simplify the test task by asking participants to produce only six words, the approach can correctly classify almost 75% of the response sets, and only one of the L1 speakers is misclassified. The L2 group contains a number of very high-level participants, and so we might expect the classifier to make some errors where an L2 speaker is judged to be performing like a L2 speaker on this task. The group of four L2 speakers who are misclassified seems like an allowable error.

Unfortunately, the results of the four remaining tasks are rather less compelling. Table 5 shows the data for ***Margaret*** – the older woman. The figures in Table 5b are quite close to the corresponding figures in Table 4b. The main difference is that both groups in the reference data set are about equally likely to generate one of the shared responses. This makes the classification rather more difficult, and Table 5c indicates that ***Margaret*** is indeed less good at discriminating between the groups than ***Shirley*** was. Here, 37 of the test cases were correctly identified as L1 or L2 speakers, but twelve L2 speakers were incorrectly classified as L1 speakers, and four L1 speakers were classed as an L2 speaker. Seven cases were undecided. A chi squared analysis suggests that this distribution is unlikely to have arisen by chance (χ^2=10.9; p<.01).

Table 5: Classified data for Margaret

Table 5a: The words participants use to describe Margaret

Shared responses
WAVING STUDIOUS FOREHEAD HARD_WORKING ENTHUSIASTIC ADULT SMILING CALM SENSIBLE QUIET HAPPY BLIND WELCOMING GLASSES GENEROUS INTELLIGENT NICE MOTHER SWEET OLD GENTLE CARING STRICT CHEERFUL FUNNY MIDDLE_AGED POSITIVE APPROACHABLE CLEVER WOMAN KIND SHY MOTHER_LIKE FRIENDLY SMART FEMALE CONSERVATIVE TEACHER

Typical L1 responses
BOOKWORM HAIR SMILEY WAVY GROOMED HELPFUL CASUAL DANCING KNOWLEDGEABLE PARTIALLY_SIGHTED KEEN LIBRARIAN CAT_LOVER PROFESSIONAL INNOCENT LOVING SIMPLE MOTHERLY

Typical L2 responses
SENSITIVE WISE, SHORT_SIGHTED PATIENT NERVOUS MATURE LOVELY RELAXED TALKATIVE EMPATHIC NAIVE EASY_GOING SHORT BEAUTIFUL OLD_FASHIONED CHARMING STRAIGHTFORWARD SYMPATHETIC POLITE FAMILIAR CURIOUS RELIABLE PRETTY MIDDLE_AGE OPEN_MINDED WELL_MANNERED RESPONSIBLE HONEST

Table 5b: Probabilities of a specific response type in the Margaret reference data

	Shared responses	L1 responses	L2 responses	singletons
L1 participants	.233	.166	.050	.551
L2 participants	.216	.050	.283	.451

Table 5c: Discrimination between the 60 response sets for the Margaret reference file

Actual evaluation	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	21	5	4
L2 speaker response set	12	2	16

Neville (Table 6) was fairly good at classifying the data sets, but identified a large proportion of undecided cases (15), and incorrectly classified five L1 speakers and seven

L2 speakers. Again, the overall distribution was not likely to have occurred by chance (χ^2=10.7; p<.005), but the relatively large number of undecided cases, particularly for the L1 speakers, is a problem.

Table 6: Classified data for Neville

Table 6a: Examples of words in the corpora for Neville

Shared responses
MAN MOUSTACHE TIRED SERIOUS MALE FUNNY ANGRY BORED MIDDLE_AGED STRICT SHORT RETIRED ARROGANT BOLD BIG_HEADED LONELY OLD_FASHIONED SHY CREEPY BAD ELDERLY WELL_DRESSED OLD WEALTHY NARROW_MINDED HAIRY RICH WISE CONCERNED GRUMPY CLEVER TRADITIONAL SLEEPY INTELLIGENT INTIMIDATING BALD TEACHER SAD NICE THOUGHTFUL BORING MARRIED

Typical L1 responses
PROPER OLDER QUIET BUSHY PROFESSIONAL STERN BIG_HEAD AWKWARD SUIT BALDING MISERABLE INQUISITIVE EYEBROWS GRANDAD RUSSIAN SMART SNOBBY

Typical L2 responses
UNFRIENDLY DISTANT RESPECTFUL RESPONSIBLE RUDE ELEGANT FAT UGLY EXHAUSTED CLOSE_MINDED THINKING WEIRD FORMAL BAD_TEMPERED

Table 6b: Probabilities of a specific response type in the Neville reference data

	Shared responses	L1 responses	L2 responses	singletons
L1 participants	.250	.206	.050	.494
L2 participants	.183	.050	.140	.627

Table 6c: Discrimination between the 60 response sets for the Neville reference file

Actual evaluation	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	16	9	5
L2 speaker response set	7	6	17

Table 7 shows the data elicited by ***Kevin,*** the picture of a young man. This cartoon was not very good at classifying the data sets. The usual chi squared test finds that the classifications were on the whole correct (χ^2=8.7; p=<.05), but a substantial number of cases were classified incorrectly (fully twelve of the L2 cases were classified as L1 speakers, and six of the L1 speakers were classified as L2 speakers. Eight of the L2 speakers were undecided. The distinguishing feature here seems to be that the L1 speakers produced a very low number of singleton responses, and were very likely to produce a response which was also used by L2 speakers.

Table 7: Classified data for Kevin

Table 7a: Examples of words in the corpora for Kevin

Shared responses
SERIOUS SMART UGLY MYSTERIOUS INTELLIGENT EXTROVERT SUSPICIOUS RUDE HAPPY INTERESTED WEIRD JUDGMENTAL SCARY SEXIST SKINNY MISCHIEVOUS WORKER HUNCHBACK ANGRY MANIC COMFORTABLE POOR BORING STARING RURAL FUNNY INTIMIDATING ADULT FARMER WORKING BIG_HEAD MAN ODD SMILING GRUMPY CREEPY SHY SILLY QUIET UNHAPPY

Typical L1 responses
GRITTY CONTENT SMIRKING UNKEMPT DANGEROUS INQUISITIVE FRINGE LAD TOUGH SIMPLE LONELY CONFIDENT HARD STRANGE COMMITTED SLY HIGH_TROUSERS GREASY UNTRUSTWORTHY UNTIDY LOST PAINTER OLDER SARCASTIC DICEY SHADY RUGGED SCRUFFY UNFRIENDLY

Typical L2 responses
CURIOUS INTROVERTED TALL FRIENDLY HAIRY BLONDE THINKING STUBBORN BIG_HEADED PLOTTING LAZY ARROGANT BORED EASY_GOING SELFISH HONEST SHORT WHITE NAIVE GOOD_LOOKING UNTRUSTING STUPID INTERESTING PROUD MIDDLE_AGE ILLITERATE POLITE SAD HANDSOME

Table 7b: Probabilities of a specific response type in the Kevin reference data

	Shared responses	L1 responses	L2 responses	singleton
L1 participants	.423	.236	.050	.291
L2 participants	.277	.050	.264	.409

Table 7c: Discrimination between the 60 response sets for the Kevin reference file

Actual evaluation	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	16	8	6
L2 speaker response set	12	2	16

The final set of results was generated in response to the child cartoon, ***Cory***. The data is shown in Table 8. This picture was by far the worst of the five cartoons, in that it failed to make clear decisions for more than half the response sets (34 response sets were classified as "undecided"). Only six cases were wrongly classified by Cory: two L1 speaker response sets were incorrectly classified as L2 speakers, and four L2 response sets were identified as L1 speakers. Again, a chi squared analysis suggests that the distribution of the classifications is unlikely to be due to chance (χ^2=8.02; p=<.05), but the overall success rate can only be described as poor.

Table 8. Classified data for Cory

Table 8a: Examples of words in the corpora for Cory

Shared responses
CHEERFUL CREEPY MALE EXCITED LAUGHING SMART NAUGHTY TEENAGER SURPRISED CURIOUS CUTE CHILDISH BOLD IMMATURE OUTGOING ACTIVE ENTHUSIASTIC MISCHIEVOUS VULNERABLE INNOCENT EARS LIVELY EXTROVERT SHORT JOYFUL ENERGETIC FRIENDLY PLAYFUL YOUNG NAIVE HAPPY SMALL CHILD UGLY FUNNY SMILEY CHEEKY

Typical L1 responses
MESSY FRECKLES ANNOYING TROUBLE BOY SPIKY_HAIR BIG_HEADED BIG_FOREHEAD JOLLY INQUISITIVE NUISANCE SPORTY LOUD WILD FUN

Typical L2 responses
SMILING EXTROVERTED EASY_GOING CRAZY SPORT SCARY CHATTY SYMPATHETIC STUBBORN HANDSOME SWEET CHARMING BLUE_EYED OPEN STRANGE GOOFY KIND MEAN NERVOUS SILLY THRILLED LITTLE IMPATIENT

Table 8b: Probabilities of a specific response type in the Cory reference data

	Shared responses	L1 responses	L2 responses	singletons
L1 participants	.173	.173	.050	.604
L2 participants	.173	.050	.183	.591

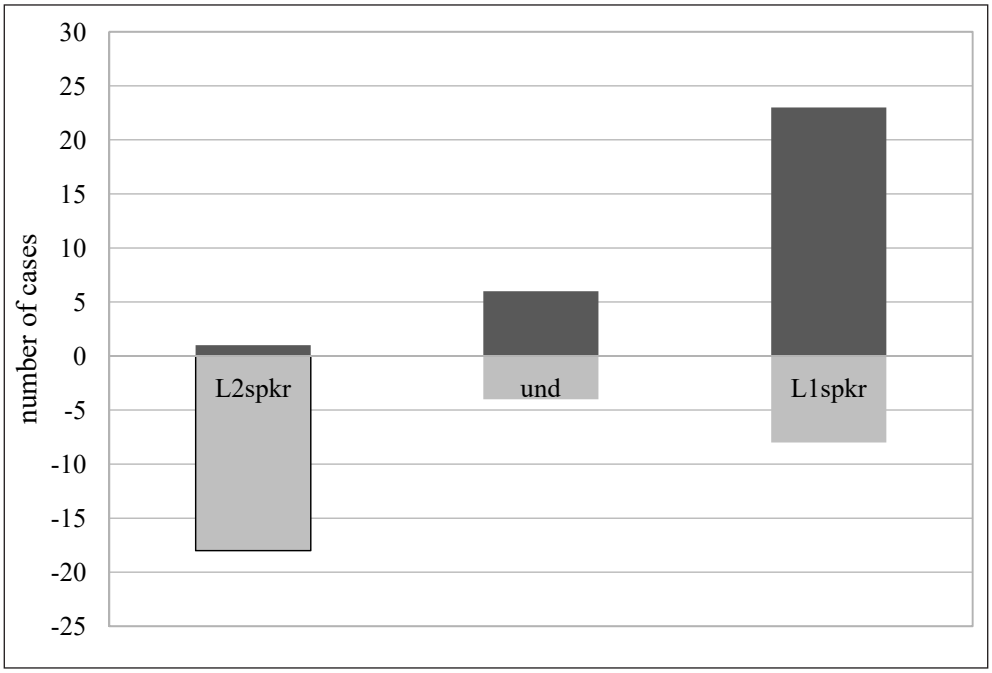
Table 8c: Discrimination between the 60 response sets for the Cory reference file

Actual evaluations	Judged to be L1 spkr.	undecided	Judged to be L2 spkr.
L1 speaker response set	10	18	2
L2 speaker response set	4	16	10

4. Discussion

The present study was set to explore how a Bayesian approach could help in distinguishing between NS and advanced English learners from a small set of words provided for the same picture stimuli. According to our results (see Tables 4-8 above), the picture stimuli in the study were actually very inconsistent in how they categorised the response sets. We originally expected that the pictures would tend to categorise an individual test-taker's response sets in the same way, but again this was not the case. None of the sixty test cases was consistently categorised by all five tasks – largely because the Neville and Cory pictures produced relatively large numbers of undecided classifications. If we disregard these two sets of results, we can combine the data from the three best classifiers into a "majority verdict" for each test case. This data is shown in Figure 4. Using this approach, the 60 test cases were generally well-categorised: most response sets were correctly ascribed to the correct group, with only one L1 response set being classified as an L2 example. A small number of response sets were classified as undecided. This distribution is better than chance (χ^2= 25.1; p<.001), but it is clearly not as good as we might have hoped. The main problem here seems to be that a small number of the L2 speakers seemed to generate response sets that were reliably classified as coming from L1 speakers, that is, their choice of words is more characteristic of the choices made by the L1 group. This may not in fact be such a serious problem as we first thought – it may just be a reflection of the very high standard of proficiency enjoyed by some of the L2 participants. If that is the case, then the real problem cases are the instances where the program classifies an L1 participant as an L2 speaker. Figure 4 shows that only 1 L1 speaker was mis-classified in this way when the three best data sets are taken into account. We should also bear in mind that there is always some error rate in estimations, especially if we conceive 'nativelikeness' as a binary category (Vanhove, 2020).

Figure 4: The "majority verdict" from the three good discriminator cartoons (Shirley, Neville and Margaret). L1 response sets are shown in black, L2 response sets are shown in grey



The present study also confirms the idea that small sets of words carry very large amounts of information about L2 speakers' vocabulary use. However, the approach used does not work as well as we expected it to do. With hindsight, it seems that the decision to ask participants to provide only six words in response to the picture stimuli was a tactical error. It resulted in a fairly high number of cases where the analysis was unable to make a confident categorisation. It also allowed the confidence judgments to be strongly influenced by a single instance of an "inappropriate" response. For example, if an L2 speaker generated just one response that was normally generated by L1 speakers, then the confidence estimate would be skewed in the direction of an L1 assessment. If this "inappropriate" word was introduced as the fifth or sixth response, then further evidence would not be available to correct this error. With a larger number of words in the response sets, errors of this sort are normally corrected. Some simulation work with artificially created responses sets suggests that responses sets consisting of 10 words are considerably more powerful than smaller response sets: they almost always result in a definite decision one way or the other. This is a question on the viability of minimal vocabulary tests that should be further explored.

It should be borne in mind that minimal vocabulary tests of this kind are constrained by the task we ask testees to perform. Therefore, the more information we have about how participants approach the task and the type of output it produces in large populations, the better it will be for the interpretation of the scores. In this study we chose these five pictures because we thought that the caricature cartoons would generate a fairly narrow range of responses, especially when we instructed participants to supply us with single-word adjectives. This turned out not to be the case – about 50% of the responses were singleton responses generated by only one participant. More importantly, perhaps, a number of respondents gave us descriptors which focussed on the style of the cartoon, rather than the person who was being depicted. BIG HEADED and BIG EARED both appeared surprisingly often in the response sets. It is not clear whether the same problem would arise if we used other kinds of visuals. Equally surprising was the finding that the cartoons differed quite markedly in the kinds of words that they elicited. We had originally thought that the stylistic similarities between the pictures would result in response sets that were to a large extent comparable, but again this turned out not to be the case. For all five stimulus pictures, the number of singleton responses was considerably larger than we had found in our pilot studies, and consequently, the number of response words that could be classified as typical L1, typical L2 and shared words was correspondingly reduced. Typical L1 words, for example, accounted for only 20% of the L1 speaker responses, and typical L2 responses accounted for only 18% of the L2 speaker responses. Shared responses accounted for around 22% of the responses. However, there was considerable variation around these means: 42% of the responses that L1 speakers made to the *Neville* picture were shared responses, and only 29% of their responses were singletons. And both groups generated about 60% of singleton response for the *Cory* picture. Clearly, there is an issue of stimulus consistency here which needs to be investigated.

It can also be possible that the two groups may have approached the task from a different point of view or used different strategies to provide answers. For example, among the NS responses, we have a number of "awkward" responses (e.g. for Shirley: *singer, teethy*), which feel as though they belong to an informal register, whereas some of the NNS responses seem to be more "literary". When checking the items for frequency and range, we see that words produced more often by learners (1) appear more frequently in the frequency lists, such as the JACET List (Ishikawa et al., 2003) (e.g. *young, shy, thin*), (2) can be more often found in students' textbooks (e.g. *cheerful, talkative, open-minded, friendly*) and (3) can be often cognates (e.g. *elegant, relaxed, attractive, extroverted, modern...*) or borrowings (e.g. *fashion*). NSs sometimes produce words that learners do not typically know or that appear less often in textbooks (e.g. *stern, scruffy, bubbly...*), but we do also find words that tend to be very basic (e.g. *quiet, grandad, boy*). So both groups use a mixture of high and low frequency items, it is not just a matter of frequency or

range: there does not appear to be a reliable significant difference between the groups in respect of these features. Register does seem to be an important feature, and some individual responses are strongly marked for this. However, we do not find that the majority of the responses generated by a single individual are characterised in this way.

We also think that the study raises some interesting questions about the use of a Bayesian approach to linguistic data of this kind. We have observed there are some technical issues that will need to be resolved in the future. In this study, we started out with two collections of response sets each generated by 80 test-takers. Each collection was split into two: fifty responses sets were used to define a reference corpus, and thirty response sets were held back to be used as test cases. Of course, these two numbers are arbitrary: we could have split the data in other ways. For example, we could have used a set of 25 response sets to establish the reference corpus, and this would have left us with 55 response sets to be used for evaluation. Or we could have used a bigger number of response sets to establish the reference corpus, and evaluated only a handful of test cases. Ideally, we would like to work with a small but reliable reference corpus, since this makes it considerably easier to build the reference corpus, and allows us to evaluate a larger number of test cases. Unfortunately, we do not know how the size of the reference corpus affects the evaluations. We might expect that increasing the number of response sets that are used to build the reference corpus would increase the number of singleton responses, but exactly how this interaction would work is unclear. We might expect that a larger reference corpus would affect the number of response words that are "typically" L1 responses or "typically" L2 responses, but it is difficult to assess this characteristic in practice. We are currently assembling some very large data sets which will allow us to answer these questions with some confidence (see Meara & Miralpeix, in preparation).

A more important issue concerns the way we have characterised the four types of words in the response sets, particularly the singleton responses. In this paper, we have treated any word which appeared only once in the relevant reference corpus as a "singleton", and we have lumped together into a single class words that were generated by a single L1 speaker, or a single L2 speaker. Any new word that appeared in the test response sets, but not in the reference corpus was classified as a singleton, regardless of its characteristics. It is probable that this classification is just too broad, and that a closer examination of the singletons generated by the L1 group and by the L2 group would reveal some subtle differences between the groups. For example, the L1 singletons tend to be infrequent words, whereas the L2 singletons are sometimes invented words based on cognates. We have not explored this avenue here, as it is difficult to automate the process of distinguishing the different types of singletons. However, a closer examination of these words would be worthwhile. Simply ignoring the singleton problem, and treating words that appear in the reference corpus as L1

words or L2 words regardless of how many times they occur would be an even simpler solution. This approach would have the added advantage of increasing the proportion of "typical" L1 and "typical" L2 words, but again, it is not clear how this approach would affect the performance of the program.

A related issue has to do with our criterion for classifying a word as a "typical L1 word" or a "typical L2 word". In this study, all words which occur at least two times but only in response sets generated by L1 speakers were identified as L1 words, and all words which occur at least two times but only in response sets generated by the L2 speakers were identified as L2 words. However, once again, we are dealing with an arbitrary cut-off here. We could have used a rather stricter criterion, in which case the number of words identified as "typical" cases would have been much smaller, and we would need to introduce a new category of "words which do not occur often" - say, all words which occur only once or twice in the reference corpus. We think that this would make the classifier program rather less accurate than it is currently. Alternatively, we could lower the threshold for describing a word as "typical", and include all words which are generated by only one of the groups. This would increase the number of "typical" words, and would allow us to eliminate the entire class of singleton words by subsuming them into the "typical L1 word" and "typical L2 word" categories. Our guess is that this might be a good way to go in future research of this kind.

The last technical issue concerns a feature that we have not commented on before, but will doubtless have been noted by astute readers. Given that the reference corpora used in this study are samples, and not comprehensive lists, there will always be occasions when, for example, an L2 speaker uses a word which has formally been defined as a "typical L1 word" because it has not appeared as a "typical L2 word" in the reference corpus. The question which arises here is how should we deal with these cases. The logical solution would be to say that, by definition, L2 speakers do not use "typical L1 words", and therefore the probability of an L2 speakers using a word of this type is nil. The problem with this obvious solution is that with these assumptions, and working through the steps in Table 3, a L2 response set that contains a single instance of a "typical L1 word" will return a value of zero despite the fact that the response set was actually generated by an L2 speaker. And once this zero value is found, it cannot be changed by any later data because of the way the mathematics works. Obviously, we need to avoid this over-determination, and we do this by setting a non-zero value to the probability that an L1 word will be generated by an L2 speaker and vice-versa. In tables 4-7 we have set these values to 0.05 – i.e. we anticipate that an L2 speaker might produce a "typical L1 response" from time to time: usually, slightly fewer than one response of this type per response set. This value of 0.05 is actually quite strict, and it severely penalises a test-taker who produces "the wrong sort of word". Ideally, we would like this non-null parameter to be an empirically based one, rather than

an arbitrary choice. As usual, we do not know how changing the non-null parameter from 0.05 to a rather higher figure would affect the way the classification program works. We think it should result in fewer incorrect classifications, but that it might generate more undecided classifications. This is an issue that we can address using the simulation approach mentioned earlier.

5. Conclusion

To sum up, this paper presents an empirical way of measuring productive vocabulary. Data from a minimal vocabulary test taken by NS and advanced EFL learners was analysed following a Bayesian approach, which was used to decide whether the data was generated by an L1 or an advanced L2 speaker. In theory, it seemed a good way to put this method to the test, as at high proficiency levels the differences in lexis between learners and native speakers may not be obvious (Hellman, 2008), and even less in this case with sets of just six words. Therefore, results from the current study can inform future research on the suitability of this method to distinguish between learners at different proficiency levels, where differences in productive vocabulary are more remarkable. This form of assessment could also be very helpful for teachers: by obtaining this information from minimal vocabulary tests, they could more easily identify students' weaknesses in vocabulary skills. We think the format might be particularly useful in situations where testing events need to be administered frequently, as administration of the test in its current format requires only a very short time. This is a considerable advantage over more traditional vocabulary tests. Despite this, the data the test provides appears to be rich, and the test format is challenging, even for advanced test-takers.

Finally, the work we have reported here has turned out not to be as straightforward as we expected and we have identified a number of technical issues that we failed to anticipate. In spite of this, we think the idea of assessing productive vocabularies using minimal vocabulary tests and Bayesian statistics might be worth of further exploration. In particular, we can speculate whether the Bayesian probabilities generated by the program correlate with scores generated by the productive vocabulary tests that we discussed in our introduction. Work of this sort clearly lies outside the scope of this paper, but we think that it would be worth doing work of this kind in future.

Acknowledgements

The authors would like to thank Dr Rhian Meara for her help in data collection, as well as all the students who participated in the study. Thanks also to the editor and reviewers of VIAL for their comments on the article.

6. References

Castañeda-Jiménez, G., & Jarvis, S. (2014). Exploring lexical diversity in second language Spanish. In K.L. Geeslin (Ed.), *The Handbook of Spanish Second Language Acquisition* (pp.498-513). Chichester: Wiley Blackwell.

Coxhead, A., Nation, P., & Sim, D. (2014). Creating and trialling six forms of the Vocabulary Size Test. *TESOLANZ Journal*, 22, 13-26.

Fitzpatrick, T., & Clenton, J. (2017). Making sense of learner performance on tests of productive vocabulary knowledge. *TESOL Quarterly,* 51(4), 844-867.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). *Range* [Computer software]. Retrieved from http://www.victoria.ac.nz/lals/staff/paul-nation/nation.aspx

Hellman, A. (2008). *The limits of eventual lexical attainment in adult-onset second language acquisition*. PhD thesis, School of Education, Boston University.

Ishikawa, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N. & Tono, Y. (2003). *JACET 8000: JACET List of 8000 basic words*. Tokyo: JACET.

Jiménez Catalán, R.M. (2014). *Lexical Availability in English and Spanish as a Second Language.* Berlin: Springer.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics*, *19*(2), 255-271.

Laufer, B., & Nation, I.S.P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*(3), 307-323.

Laufer, B., & Nation, I.S.P. (1999). A vocabulary size test of controlled productive ability. *Language Testing, 16*(1), 33-51.

McGrayne, S.B. (2011). *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines & emerged triumphant from two centuries of controversy*. Boston MASS.: Yale University Press.

Meara, P.M., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*(2), 142-154.

Meara, P.M, & Fitzpatrick, T. (2000). Lex30: an improved method of assessing productive vocabulary in an L2. *System*, *28*(1), 19-30.

Meara, P.M., & Miralpeix, I. (2015a). *V_YesNo.* Cardiff: Lognostics.

Meara, P.M., & Miralpeix, I. (2015b). *V_Size.* Cardiff: Lognostics.

Meara, P.M., & Miralpeix, I. (2017). *Tools for Researching Vocabulary*. Bristol: Multilingual Matters.

Meara, P.M., & Miralpeix, I. (in prep.). *Minimal vocabulary tests for English language learners.*

Meara, P.M., & Olmos Alcoy, J.C. (2010). Words as species: An alternative approach to estimating productive vocabulary size. *Reading in a Foreign Language*, *22*(1), 222-236.

Melka Teichroew, F.J. (1997). Receptive vs. productive aspects of vocabulary. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp.84-102). Cambridge: CUP.

Miralpeix, I. (2020). L1 and L2 vocabulary size and growth. In Webb, S. (Ed.). *The Routledge Handbook of Vocabulary Studies* (pp. 189-206). New York: Routledge.

Nation, I.S.P. (1984). *Vocabulary Lists: Words, affixes and stems*. Victoria University of Wellington: English Language Institute. Occasional publications 12.

Nation, I.S.P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I.S.P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(1), 9-13.

Norouzian, R., de Miranda, M., & Plonsky, L. (2018). The Bayesian revolution in second language research: An applied approach. *Language Learning 68*(4), 1032-1075.

Pearl, L., & Goldwater, S. (2016). Statistical learning, inductive bias, and Bayesian inference in language acquisition. In J.L. Lidz, W. Snyder, & J. Pater (Eds.), *The Oxford Handbook of Developmental Linguistics* (pp.664-695). Oxford: OUP.

Roghani, S., & Milton, J. (2017). Using category generation tasks to estimate productive vocabulary size in a foreign language. *TESOL International Journal*, *12*(1), 128-142.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55-88.

Stone, J.V. (2013). *Bayes' Rule: A tutorial introduction to Bayesian Analysis*. Sheffield: Sebtel Press.

Vanhove, J. (2020). When labelling L2 users as nativelike or not, consider classification errors. *Second Language Research, 36*(4), 709-724.

Webb, S. (2018). Receptive and productive vocabulary size of L2 learners. *Studies in Second Language Acquisition, 30*(1), 79-85.

Webb, S., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL - International Journal of Applied Linguistics, 168*(1), 34-70.

Williams, J., Segalowitz, N., & Leday, T. (2014). Estimating second language productive vocabulary: A capture-recapture approach. *The Mental Lexicon, 9* (1), 23-47.

Xue, G., & Nation, I.S.P. (1984). A University Word List. *Language Learning and Communication* 3 (2), 215-229.

Zinszer, B.D., Rolotti, S.V., Li, F., & Li, P. (2018). Bayesian word learning in multiple language environments. *Cognitive Science* (42, Suppl.2), 439-462.

Contents:

Universidade Vigo