RiCL Research in Corpus Linguistics

The Primary Education Learners' English Corpus (PELEC): Design and compilation

Zeltia Blanco-Suárez – Francisco Gallardo-del-Puerto – Evelyn Gandón-Chapela
University of Cantabria / Spain

Abstract – This paper describes the process of design and compilation of the *Primary Education Learners' English Corpus* (PELEC), a learner corpus which includes written (14,577 words) and spoken materials (47,032 words) from Primary Education learners in the Autonomous Community of Cantabria. It is composed of data from a total of 252 students in the fourth and sixth grade of Primary Education (aged 9–10 and 11–12, respectively) who were studying in five different state schools which followed either a Content and Language Integrated Learning (CLIL) or an English as a Foreign Language (EFL) approach.

Keywords – PELEC; learner corpora; Content and Language Integrated Learning (CLIL); English as a Foreign Language (EFL); young learners; Primary Education

1. INTRODUCTION[1]

Over the past three decades, the role of computer learner corpora, i.e. "systematic computerized collections of texts produced by language learners" (Nesselhauf 2004: 125), has become of paramount importance in the field of Second Language Acquisition and Teaching. Thanks to the technological advances currently available, it is now possible to store and tag language learners' oral and written productions in electronic format, as well as to retrieve and analyse them automatically. This fact opened up a wide range of research possibilities given that scholars could access learners' data much more quickly and easily, instead of having to store and consult them in "shoe boxes" (Díez-Bedmar 2009: 920). It is in this light that numerous learner corpora have emerged, especially in Europe (see, among many others, the *International Corpus of Learner English* (ICLE),

[1] The authors would like to thank Dr. Julia Williams Camus for her invaluable help in the design and compilation of PELEC. We would also like to thank two research assistants (Míriam Fernández Arenal and Silvia Mendiguchía Pérez) for their participation in the data gathering procedure. For generous financial support, we are grateful to the Vice-rectorate for Research and Transfer of Knowledge from the University of Cantabria (grant ref. UC2016-GRE-10). Special thanks go to Alberto San Emeterio Bolado for the storytelling panels. We would also like to thank two anonymous reviewers for their helpful comments and suggestions.

initiated by Granger (Catholic University of Louvain) in 1990 and first published in 2002; see Granger *et al*. 2002). In the particular context of Spain, several English learner corpora have been compiled with data from primary, secondary and university students (see Section 2 for more details).

This paper presents the characteristics of *The Primary Education Learners' English Corpus* (henceforth PELEC), which was compiled in 2018 by a team of researchers at the University of Cantabria with the aim of gathering data from primary students in the Autonomous Community of Cantabria. This corpus includes both written and spoken materials from Primary Education learners of English as a second language and totals 61,609 words. At the time of data collection, the participants were enrolled in the fourth and sixth grade of Primary Education at five different state schools in Cantabria, which offered either traditional English as a Foreign Language (EFL) or Content and Language Integrated Learning (CLIL) programmes. All the schools selected were located in the outskirts of Santander and their choice was motivated by the rather homogeneous socio-economic status of the families in these institutions, in contrast to urban schools, which provided a sharper contrast in that regard.

As is well-known in the field of learner corpora, collecting data from young learners from this level of education poses a great number of additional challenges because of the technical difficulties involved. To mention but a few, data collection from this population involved obtaining ethical consent from the young learners' parents or legal guardians and the compliance with the current data protection regulations, especially so because the students were both audio recorded and videotaped when performing the oral tasks. However, even if the procedures of data collection and processing were rather complex and time-consuming, they were worth the effort in that they offer the possibility of analysing both the written and oral data from the same participant. This fact alone is innovative and opens a number of research avenues which have been relatively unexplored so far. Moreover, given that PELEC contains data from CLIL and non-CLIL state schools in Spain, the results gathered from its analysis may help to advance not only in the field of Second Language Acquisition research, but also in the realms of language teaching, language planning and language policies.

In the remainder of this paper, we will set forth the process of compilation and main characteristics of PELEC, a learner corpus featuring spoken and written data from L1-Spanish young learners of English. To this end, Section 2 first provides an initial

overview of the extant learner corpora which are most relevant to PELEC. Section 3 is concerned with the process of design and compilation of the corpus, including the participants of the study and the types of data collected. Finally, in section 4 the potential applications of PELEC and issues for further research are presented.

2. Learner corpora: An overview

Learner corpora offer manifold research possibilities, although they have been mainly used for two purposes, namely to gain insights into second or foreign language acquisition and to create tailor-made pedagogical materials based on the students' most frequent errors (Granger 2008: 259). This section will offer a brief account of the learner corpora available with data from L1-Spanish learners of English. As will be shown, these learner corpora have targeted different levels of education, ranging from primary through to secondary and university levels. Some of them contain both oral and written data, whereas some others have focused only on written data. In what follows, a classification of the levels and types of data included in the different corpora is offered. Firstly, we will present the characteristics of the corpora based on data obtained from primary and secondary students, to then focus on those corpora that only contain data from university level pupils.

Among the corpora that compile data from primary and secondary students, one can find *The Barcelona Age Factor* (BAF) *corpus*, *The Barcelona English Language Corpus* (BELC) and the corpora used by the research teams *Research in English Applied Linguistics* (REAL) and *Language and Speech Laboratory* (LASLAB), based at the University of Barcelona and the University of the Basque Country, respectively.[2] In particular, the BAF corpus and the BELC corpus gathered both oral (oral narratives, role plays, oral interviews) and written data (written compositions) to analyse the effect played by the age at which students started to learn English (2,063 participants in total). The BELC corpus, stemming from the BAF corpus, allows to track subjects longitudinally over a period of seven years (see Muñoz 2006). The participants in these corpora were all bilingual in Spanish and Catalan and studied in state schools in Catalonia, differing only in the starting age of English instruction (8, 11, 14 and 18).

[2] For further information on the BAF and BELC corpora, see https://slabank.talkbank.org/access/English/BELC.html.

The age factor was also the focus of a number of projects stemming from REAL and LASLAB. This team of researchers gathered data from bilingual (Basque and Spanish) students enrolled at various primary and secondary schools in this region, which offered either CLIL programmes or traditional EFL classrooms. The participants, who had started learning English at different ages (4, 8 and 11), were asked to do several written and oral tasks, as well as different reading and listening activities, so as to measure their overall language proficiency. This variety of data enabled researchers to shed light on a wide range of aspects relevant to the acquisition and learning of English, including the maturational effects in L3 English, the learning context and the type of task performed (see, among many others, García-Mayo and García-Lecumberri 2003; Lasagabaster and Doiz 2003; Gallardo-del-Puerto and Gómez-Lacabex 2013, 2017; García-Mayo and Imaz-Agirre 2019).

Several learner corpora have also sourced from secondary and university students. This is the case, for instance, of the *Santiago University Learner of English Corpus* (SULEC), the *Universidad Autónoma de Madrid* (UAM) *Corpus de Interlenguas Escritas* and the *Corpus of English as a Foreign Language* (COREFL).

The first of these resources, SULEC, is a monitor corpus compiled at the University of Santiago de Compostela under the direction of Ignacio Palacios-Martínez (see Palacios-Martínez 2005). The project intended to compile 1,000,000 words of oral and written English by primary, secondary and university students with different proficiency levels. At present, the corpus comprises around 500,000 words[3] and it only contains data from secondary (first and second year *Bachillerato*) and university students from the degree in English Philology with intermediate and advanced levels of proficiency. The written data were collected in the form of 500-word compositions, whereas the oral component was collected through instruments such as oral presentations, oral exams and personal interviews.

The second corpus which also includes data from secondary and university students, the so-called *UAM Corpus de Interlenguas Escritas*, is divided into three subcomponents. One of them contains 210 essays written by secondary school pupils during class time. This subcomponent includes 174 texts that were written by 87 students from the first, second and third years of *Bachillerato* and the former pre-university *Curso*

[3] Palacios-Martínez, January 2020, personal communication.

de Orientación Universitaria (COU) before and after a pedagogical intervention. The remaining 36 texts belong to students from the same levels of education but do not have a pre- or post-task counterpart (Barrio-Luis 2005: 64–65; Díez-Bedmar 2009: 925). The main results of this intervention programme were published in a book by Martín-Úriz and Whittaker (2005). Another subcomponent is formed by a collection of essays written by 119 pre-university students from several high-schools who wrote about three composition topics and answered a cloze test (Martín-Úriz and Whittaker 2005; Díez-Bedmar 2009: 925). The third subcomponent of this corpus is composed by essays on the same topic as the first subcomponent, which were written by students in their first year of the degree in English Philology at the Autonomous University of Madrid (Díez-Bedmar 2009). The process of error-tagging of this corpus and the toolbar used to tag errors and retrieve them have been described in Barrio-Luis (2005).

The COREFL, on its part, is a corpus that is currently being compiled at the Universities of Granada and Bremen. It contains L2 English written and spoken data from L1 Spanish and L1 German learners at secondary and university levels with different proficiency levels (A1-C2 in the Common European Framework of Reference for Languages) and ages (from 12 years onwards). As of September 2019, according to Díaz-Negrillo *et al.* (2019), the corpus contained approximately 1,612 texts (189 oral and 1,423 written texts) sampling four different types of narrative tasks. In the future, this corpus will also feature two control corpora of the learners' L1, that is, Spanish (including Peninsular and Latin American varieties) and German (under compilation). Interestingly, the L1 Spanish-L2 English subcomponent includes data from learners in different types of instructional contexts: secondary school bilingual programmes (CLIL) vs. mainstream EFL classrooms, on the one hand, and university English as a Medium of Instruction (EMI) learners vs. university Spanish as a Medium of Instruction (SMI) learners, on the other.

By contrast, some other corpora restricted their samples to university students, as is the case of the *Madrid Corpus* (MAD), the *Written Corpus of Learner English* (WriCLE), the *English Written Interlanguage* (ENWIL) and the *Non-native Spanish Corpus of English* (NOSE).

MAD was compiled at the Complutense University of Madrid by the SPAINWRITE research group and is subdivided into three components. The first one is composed of a collection of argumentative essays written by over 200 students of English

as a foreign language from the degree in English Philology in their first and fourth years. The second subcomponent includes argumentative essays by the same students in their native language, and finally, the third subcomponent consists of a control corpus of essays written by third-year American students of the degree in Spanish Philology that formed part of the Middlebury Programme in Madrid (Díez-Bedmar 2009: 923).

The *Written Corpus of Learner English* (WriCLE; Rollinson and Mendikoetxea 2010), on its part, includes two subcomponents. The first, WriCLEformal, contains a set of 752 essays written by Spanish university students (from all levels of proficiency) in their first and third years of the degree in English Studies at Autonomous University of Madrid (around 750,000 words in XML format). The second subcomponent, WriCLEinf, is the informal, non-academic counterpart of the WriCLEformal, featuring texts from blogs, emails, autobiographical pieces of writing, narratives, descriptions, poems, among many others, amounting to 1,140 texts and totalling around 8,000 words.[4]

Another learner corpus drawing upon collections of university students' essays is the *English Written Interlanguage* (ENWIL) corpus, created in 1997. ENWIL includes essays written by first-year students of English Philology at the University of Alcalá de Henares (Valero-Garcés *et al*. 2000). In line with the *UAM Corpus de Interlenguas Escritas*, it has also been error-tagged and the ultimate aim was to create more tailor-made materials targeting the students' needs, which resulted in a resource book for Spanish learners of English on how to write successfully in the realm of academic writing (Valero-Garcés *et al.* 2003).

Likewise, the *Non-native Spanish Corpus of English* (NOSE), of about 300,000 words from 1,000 samples of 250-300 words, is a collection of descriptive and argumentative essays written by approximately 500 Spanish students of English at the universities of Granada and Jaén. The corpus is also error-tagged and is available with a corpus tool, which allows to search for a number of variables, including the informants' profiles, topics and text types (Díaz-Negrillo 2012). This has not only permitted researchers to assess and diagnose the written competence of the learners, but also to "propose remedial work to counteract students' difficulties" (Díaz-Negrillo 2012: 43).

In addition to the aforementioned L1-Spanish English learner corpora, there also exist a number of learner corpora with different mother tongue backgrounds. Among such

[4] For more detailed information on this corpus, visit http://wricle.learnercorpora.com/.

initiatives we find the *International Corpus of Crosslinguistic Interlanguage* (ICCI), which constitutes an international joint project initiated by Prof. Yukio Tono (Tokyo University of Foreign Studies) in 2007 (Hong 2012: 47). This corpus contains 6,700 transcripts of argumentative and descriptive essays written by students from grades 3 to 12, i.e. from primary and secondary education levels. This over 500,000-word corpus represents 6,700 subjects from seven different countries (Austria, China, Hong Kong, Israel, Poland, Spain and Taiwan), thirty-five mother tongues and different proficiency levels (see Hong 2012: 50–51 for more details). Likewise, Sylviane Granger launched another project in Europe, the *International Corpus of Learner English* (ICLE), whose first version was published on CD-ROM in 2002 and which is available since 2009 as an expanded version, ICLEv2 (see Granger *et al.* 2002). ICLE comprises argumentative essays by intermediate to advanced learners of English from languages as diverse as Bulgarian, Chinese, Dutch, Finnish or Turkish, totalling about 3.7 million words. The compilation of the Spanish subcomponent of ICLE, known as SPICLE, was undertaken by the compilers of MAD (Martínez-Osés and Neff-van Aertselaer 2001). ICLE has a spoken counterpart, the *Louvain International Database of Spoken English Interlanguage* (LINDSEI), which includes interviews with EFL learners with different mother tongues, for a total of about 100,000 words (Gilquin *et al*. 2010). Following the same principles and guidelines in ICLE and LINDSEI, the Centre for English Corpus Linguistics at the Catholic University of Louvain additionally created the *Longitudinal Database of Learner English* (LONGDALE), which, unlike the spoken and written corpora, is not exclusively focused on interviews or argumentative essays, but contains a wide variety of data, including grammaticality judgement tests.[5] At a much larger scale than ICLE and LINDSEI, Cambridge University Press and Longman have devised two commercial mega-corpora which are constantly being expanded: the *Longman Learners' Corpus* and the *Cambridge Learner Corpus* (CLC). These corpora contain over 10 million words and represent countless mother tongue backgrounds (Granger 2008: 261).

Having provided a broad overview of the main corpora of L1-Spanish learners of English, in the following section we will offer a detailed description of the design of PELEC and its compilation, including data regarding the participants, the different

[5] For more information on LONGDALE, see https://uclouvain.be/en/research-institutes/ilc/cecl/longdale.html.

questionnaires and tests used to gather the data, as well as an overview of the written and spoken components of the corpus.

3. CORPUS DESIGN AND COMPILATION

3.1. Participants

The corpus gathers data from a total of 252 students from the fourth and sixth grade of Primary Education (aged 9-10 and 11-12, respectively) in five different state schools in Cantabria. In compliance with current data protection regulations, the participants' parents or legal guardians were asked to sign a consent provided by the University of Cantabria before the data compilation process started in each of the schools. Three of these schools offered a CLIL approach, while in the remaining two learners received regular EFL courses. Therefore, in the non-CLIL groups students were exposed to three weekly hours of English, as required in the curriculum. In turn, in the CLIL groups learners benefited from two extra CLIL hours of English in addition to the compulsory ones, in subjects such as Natural Sciences, Physical Education, Arts and Crafts or Music. Table 1 shows the distribution of the students in the corpus according to the type of approach, gender, grade and number of hours of instruction.

	Students	**Gender**	**English exposure**	
			Grade 4	**Grade 6**
CLIL	124	F: 46.77% (n=58) M: 53.23% (n=66)	EFL 361h CLIL 307 h	EFL 617 h CLIL 462 h
Non-CLIL	128	F: 53.12% (n=68) M: 46-88% (n=60)	EFL 361 h	EFL 617 h

Table 1: Description of participants in PELEC

3.2. Types of materials

3.2.1. Questionnaires and tests

In order to obtain a more comprehensive picture of their English learning profiles, all the students were asked to complete an initial questionnaire in Spanish consisting of biographical information (parents' or legal guardians' occupations, L2 learning onset time, extramural exposure, etc.) and of the compensatory strategies pursued when attempting to overcome a communicative gap, such as L1 use, appeal for assistance, paraphrasing, etc. (see, among others, O'Malley and Chamot 1990; Gallardo-del-Puerto

and Gómez-Lacabex 2017). Together with these initial questions, they were also asked to answer a motivation questionnaire in Spanish, which was based on Gardner's (1985) Attitude/Motivation Test Battery (AMTB) and adapted to this type of learners, in line with previous motivation studies for young apprentices (Kiss and Nikolov 2005; Carreira 2006; Cid *et al*. 2009; Lasagabaster and Sierra 2009; Fernández-Fontecha 2014, 2015). This test comprised a total of 34 items measuring factors such as their reported effort and self-efficacy, their willingness to integrate in the target language community, their anxiety levels, the degree of parental support, as well as their intrinsic and extrinsic motivation. All the statements in these questionnaires had to be marked on a 5-point Likert scale, from the lowest (*I do not agree at all* ⛔⛔) to the highest degree (*I completely agree* 👍👍). The results from the motivation survey revealed that the motivation profiles of the CLIL and non-CLIL students were rather similar (see Gallardo-del-Puerto and Blanco-Suárez forthcoming): both groups exhibited particularly high levels of extrinsic motivation (the external factors to learn the language), which should not be surprising given the foreign, non-naturalistic learning setting. Interestingly, CLIL students reported being more encouraged by their parents or families to learn English than their EFL counterparts. As for the effect of gender, no differences were found in the motivation scores of male vs. female young learners in the CLIL group. Nonetheless, in the non-CLIL context girls outperformed boys in the overall and intrinsic motivation. This motivation questionnaire, together with the one on background information and the one on compensatory strategies, were completed during a 50-minute session.

In addition to the aforementioned questionnaires, the students' competence in English was examined by means of a series of language tests (see Table 2). Thus, they had to do a listening, reading and a use of English test, for which purposes we drew on materials from the Cambridge English A1 Movers and A2 Flyers tests. The listening comprehension test consisted of two multiple choice exercises (with five items each) and the reading comprehension included three short texts which described one student each and their daily routines. In total, they had to answer ten multiple choice questions related to these characters and their lives. For the use of English test, our young learners had to complete a cloze test with ten gaps in two emails which were exchanged between a Spanish and a Chinese student. The blanks related to grammatical contents such as the article use, the third person present tense inflection *–s* and the use of pronouns or

prepositions, in accordance with the curriculum for those educational levels. These language tests were conducted on two separate days during a 50-minute session.

	Use of English (max=10)	**Listening (max=10)**	**Reading (max=10)**
Non-CLIL Grade 4	4.24	5.39	3.36
CLIL Grade 4	4.87	5.25	3.49
Non-CLIL Grade 6	6.55	7.23	4.76
CLIL Grade 6	6.69	7.02	5.33
Non-CLIL all	**5.42**	**6.29**	**4.07**
CLIL all	**5.52**	**5.84**	**4.13**

Table 2: Mean scores in the language tests

As can be seen in Table 2, the mean scores obtained by CLIL and non-CLIL learners were rather similar, more differences being observed when grade 4 vs. grade 6 students are compared, either in the CLIL or the non-CLIL samples.

3.2.2. Written component

The written component of PELEC comprises 246 compositions of L1-Spanish learners of English, totalling 14,577 words (6,398 and 8,179 words from fourth and sixth grade, respectively), as shown in Table 3. The average length of these writings is 58.08 words, ranging from a minimum of 4 to a maximum of 191 words, and the standard deviation is +/-33.401.

	4th grade	**6th grade**	**Both grades**
	No. of words	No. of words	No. of words
Non-CLIL	2,870	4,629	7,499
CLIL	3,528	3,550	7,078
Total	**6,398**	**8,179**	**14,577**

Table 3: Written component of PELEC

For this part, students were asked to write a short letter to an English pen friend, Tom, telling him about their favourite things and what they do on a normal day. They had to complete this task in approximately 20 minutes. Moreover, they were asked to write the same letter in Spanish some weeks later, which would additionally allow us to verify their level of written competence in their first language.

All the writings were scanned and later transcribed and saved in separate .txt files, so that the texts could be prepared for subsequent analysis with a concordance software and with the CLAN tool from *TalkBank* software.[6] Each file contains the body of the text

[6] For more information on CLAN and *TalkBank*, see https://childes.talkbank.org/.

(a student writing) with its corresponding header, which includes the following metadata: filename, school name, grade and group, date of collection, as well as the student's name(s) and surnames. The filename additionally identifies the type of task, school and year. Data regarding the students' language competence in the other language tests (reading, listening and use of English) and questionnaires were stored on a separate Excel spreadsheet with all the participants. The corpus has thus far not been tagged for part of speech (POS) and no linguistic annotation has yet been added to mark any relevant lexical or morphosyntactic features. Nonetheless, corpus annotation would be possible in the .txt files with XML-language, in compliance with the TEI guidelines, and in CLAN. The following excerpt from one of the students' written compositions serves to illustrate the written component of PELEC:

(1) *CHI: at eight o'clock I go at the kitchen to have lunch.
*CHI: at half past eight I'm wash my teeths and I dress up.
*CHI: I take the bag and I go to school.

3.2.3. Oral component

As shown in Table 4, the spoken component of PELEC includes a total of 47,032 words, 24,863 words from fourth-grade student spoken productions and 22,169 words of spoken materials from sixth-grade students. The average length of students' oral productions is 181.05 words,[7] with a standard deviation of +/-89.457, a minimum of 11 words and a maximum of 572. The oral corpus contains a total of 7,771 utterances with a mean of 32.66 utterances per student, the minimum and maximum being 4 and 82 utterances, respectively. The average length of individual utterances is of 6.46 words, the standard deviation being +/-3.981 and the minimum and maximum ranging from 1 to 39.8 words per utterance.

	4th grade	**6th grade**	**Both grades**
	No. of words	No. of words	No. of words
Non-CLIL	11,092	13,435	24,527
CLIL	13,771	8,734	22,505
Total	**24,863**	**22,169**	**47,032**

Table 4: Spoken component of PELEC

[7] Given that PELEC consists of three different tasks and that the data for each of them were collected on various days, the number of students who accomplished each task may differ.

For the spoken data of PELEC, our young learners were taped and video recorded performing two separate tasks. In the first one, they were given a set of pictures and asked to tell the story individually (see Figure 1). On average, the students completed this task in five minutes. Additionally, they were asked to tell the story in Spanish some weeks later, which would allow us to compare these children's spoken behaviour in the L1 and the L2.

Figure 1: Speaking task I: Storytelling

The drawings show a friendly dog which seems to be lost on a rainy day. Luckily, the protagonist of the story, a young boy holding an umbrella, runs into the dog and they walk home together. There they are welcomed by the boy's parents and the dog becomes a new member of the family. The students were asked to perform this task following their own resources, without any help. The researcher who was recording the spoken productions did not intervene at any moment, unless required by the child in cases of appeal for assistance or in cases in which the students had a mental block and were not able to continue with the activity. Example (2) provides an extract of a student's oral production in this task.

(2) *CHI: the person with the umbrella eeeh (3.) takes the dog.
*CHI: a:nd (3.) travel with him to her house.
*CHI: and they take dog for us.

The second was a spot-the-differences task which was done in pairs, in line with an exercise in the Cambridge Young Learners English Test (Movers, A1 level). By asking each other questions, dyads had to collaborate to find five differences in their respective photos (see Figure 2), which took them approximately ten minutes. Since there was a barrier between them, they could not see each other's photos, so they were forced to rely

on the linguistic resources at hand to discover the differences. Both oral production activities were recorded and later transcribed by two coordinated research assistants. As in the case of the storytelling, the researcher(s) present during the recording session did not take part unless specifically requested by the students or in those cases in which they became too nervous and were unable to follow.

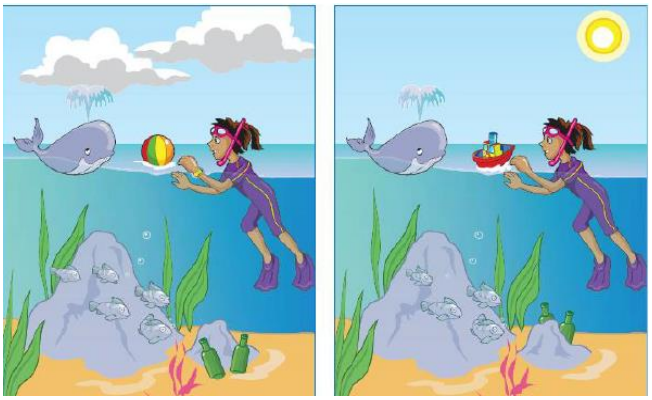

Figure 2: Speaking task II: Spot the differences[8]

For the data transcription conventions, we issued some guidelines based on other spoken corpora, including LINDSEI, the Lancaster Corpus and VOICE, as well as on CLAN and *TalkBank*. Thus, for each transcription we marked the participants' names, including that of the researcher who was present during the recording session, the filename, the transcriber's name, the date of the recording and then the beginning and end of each of the interventions. Moreover, we marked each speaker's turn, any overlaps in the different turns, pauses or lengthening of an utterance, laughter and non-verbal elements (e.g. coughing or body language), L1 use and the use of fillers. Example (3) illustrates part of the interaction between two students in this task:

(3) *CH1: aaam (7.) in your pictures the bottles is green?
*CH2: the eeeh yes.
*CH2: eeem eeeh i:n your picture the girl wear a ponytail?
*CH1: yes.

4. CORPUS APPLICATIONS AND ISSUES FOR FURTHER RESEARCH

PELEC opens up a wide range of research possibilities. Firstly, the potentialities of a corpus of this kind in the field of language acquisition research are immense, since it

[8] Taken from the online A1 Movers sample test at https://www.cambridgeenglish.org/Images/young-learners-sample-papers-2018-vol1.pdf.

would allow numerous studies on phonological, lexical, morphosyntactic and discourse aspects, thereby detecting the most problematic areas for learners in both CLIL and non-CLIL learning contexts. In this regard, the first study derived from the investigation of the present corpus analysed verb number agreement errors and null subjects (Fernández-Pena and Gallardo-del-Puerto 2019) in Primary Education Grade 6 schoolchildren. This investigation did not find striking differences between the CLIL and non-CLIL samples examined concerning these two aspects of English grammar. Both groups omitted expletive subject pronouns (**is raining* vs. *it is raining*) to a considerably larger extent than referential subject pronouns (**is a dog* vs. *it is a dog*), the latter being dropped minimally. Similarly, they both produced omission errors more frequently in affixal (**the boy sleep with the dog* vs. *the boy sleeps with the dog*) than in suppletive inflection (**the boy sleeping with the dog* vs. *the boy is sleeping with the dog*). However, non-CLIL learners omitted auxiliary *be* (**the boy sleeping with the dog* vs. *the boy is sleeping with the dog*) more frequently than copula *be* (**the dog in a bedroom* vs. *the dog is in a bedroom*), which, together with their greater use of null expletive subjects and of placeholders (**the boy is sleep with the dog* vs. *the boy sleeps with the dog*), was indicative of an earlier stage of acquisition. Conversely, although the presence of commission errors (**they goes to bed* vs. *they go to bed*) was minimal in the data, CLIL learners' rate of incorrect inflection supply in copula *be* contexts (**your eyes is brown* vs. *your eyes are brown*) was surprisingly higher than that of non-CLIL learners. In addition to the analyses reported by Fernández-Pena and Gallardo-del-Puerto (2019), PELEC would permit the analysis of, for instance, measures of amount of production (type and token ratios), density of production (total number of tokens per utterance, etc.) and compensatory strategies, such as appeals for assistance or L1 use, in line with the studies by Gallardo-del-Puerto and Gómez-Lacabex (2013, 2017).

Secondly, cross-linguistic studies would also be possible with PELEC, given that, as was the case with MAD and COREFL, PELEC records L1-data in the writing and in the storytelling tasks. In addition, the analysis of the spot-the-differences task in English would allow a comparison with the results from previous investigations on collaborative interaction such as García-Mayo and Imaz-Agirre (2019) and Martínez-Adrián (2020). The former discovered that young CLIL learners' occurrence of language-related episodes depended on the type of task, whereas the latter found that older schoolchildren

resorted to previously known languages more frequently than younger ones to keep the flow of interactive speech.

Thirdly, the analyses derived from the students' L2-productions in this corpus would be highly beneficial to educators, as another possible application would be the creation of more targeted and tailor-made materials based on the most recurrent errors, in line with the aims of some of the corpora mentioned in Section 2. Importantly, the analysis of the output of the CLIL vs. non-CLIL learning contexts would also enable the contribution to the realms of language planning and language policies in Spain in the long run.

Finally, although the corpus presents several limitations in terms of the student sample and its size, it could be expanded by including representation from all grades in Primary Education and additional schools in Cantabria, both from CLIL and non-CLIL approaches. Furthermore, the process of data collection and transcription could be extended to other educational levels such as Secondary and Tertiary Education. This would of course allow us to obtain a more comprehensive picture of the productive skills of the English learners in this northern region in Spain.

REFERENCES

Barrio-Luis, María. 2005. Diseño del corpus de interlenguas de textos escritos en inglés lengua extranjera. In Ana Martín-Úriz and Rachel Whittaker eds. *La Composición como Comunicación: Una Experiencia en las Aulas de Lengua Inglesa en Bachillerato*. Madrid: Ediciones de la Universidad Autónoma de Madrid, 61–75.

Carreira, Junko Matsuzaki. 2006. Motivation for learning English as a foreign language in Japanese elementary schools. *JALT Journal* 28/2: 135–158.

Cid, Eva, Gisela Grañena and Elsa Tragant. 2009. Constructing and validating the foreign language attitudes and goals survey (FLAGS). *System* 37/3: 496–513.

Díaz-Negrillo, Ana. 2012. Learner corpora: The case of the NOSE corpus. *Systemics, Cybernetics and Informatics* 10/1: 42–47.

Díaz-Negrillo, Ana, Cristóbal Lozano and Marcus Callies. 2019. Introducing the Corpus of English as a Foreign Language (COREFL): A bimodal, multi-task corpus for SLA research. Paper presented at the *5th Learner Corpus Conference* (12–14 September 2019). University of Warsaw.

Díez-Bedmar, María Belén. 2009. Written learner corpora by Spanish students of English: An overview. In Pascual Cantos-Gómez and Aquilino Sánchez-Pérez eds. *A Survey of Corpus-based Research*: *Proceedings of the AELINCO Conference*. Murcia: Asociación Española de Lingüística de Corpus, 920–933.

Fernández-Fontecha, Almudena. 2014. Motivation and gender effect in receptive vocabulary learning: An exploratory analysis in CLIL Primary Education. *Latin American Journal of Content and Language Integrated Learning* 7/2: 27–49.

Fernández-Fontecha, Almudena. 2015. Motivation and vocabulary breadth in CLIL and EFL contexts: Different age, same time of exposure. *Complutense Journal of English Studies* 23: 79–96.

Fernández-Pena, Yolanda and Francisco Gallardo-del-Puerto. 2019. Number agreement errors and subject omission in CLIL vs. non-CLIL learners of English in Primary Education. Paper presented at the *43rd Conference of the Spanish Association of Anglo-American Studies* (13–15 November 2019). University of Alicante.

Gallardo-del-Puerto, Francisco and Zeltia Blanco-Suárez (forthcoming). Foreign Language Motivation in Primary Education students: The effects of additional CLIL and gender. *Journal of Immersion and Content-based Language Education*.

Gallardo-del-Puerto, Francisco and Esther Gómez-Lacabex. 2013. The impact of additional CLIL exposure on oral English production. *Journal of English Studies* 11/1:113–131.

Gallardo-del-Puerto, Francisco and Esther Gómez-Lacabex. 2017. Oral production outcomes in CLIL: An attempt to manage amount of exposure. *European Journal of Applied Linguistics* 5/1: 31–54.

García-Mayo, María Pilar and María Luisa García-Lecumberri eds. 2003. *Age and the Acquisition of English as a Foreign Language*. Clevedon: Multilingual Matters.

García-Mayo, María del Pilar and Ainara Imaz-Agirre. 2019. Task modality and pair formation method: Their impact on patterns of interaction and LREs among EFL primary school children. *System* 80: 165–175.

Gardner, Robert C. 1985. *Social Psychology and Second Language Learning: The Role of Attitudes and Motivation*. London: Edward Arnold.

Granger, Sylviane. 2008. Learner corpora. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics: An International Handbook* (Volume 1). Berlin: Mouton de Gruyter, 259–275.

Gilquin, Gaëtanalle, Sylvie De Cock and Sylviane Granger eds. 2010. *LINDSEI: Louvain International Database of Spoken English Interlanguage*. Louvain: UCL Presses.

Granger, Sylviane, Estelle Dagneux and Fanny Meunier. 2002. *The International Corpus of Learner English*. Louvain: Université Catholique de Louvain.

Hong, Huaqing. 2012. Compilation and exploration of ICCI corpus for learner language research. In Yukio Tono, Yuji Kawaguchi and Makoto Minegishi eds. *Developmental and Crosslinguistic Perspectives in Learner Corpus Research*. Amsterdam: John Benjamins, 47–62.

Kiss, Csilla and Marianne Nikolov. 2005. Developing, piloting and validating an instrument to measure young learners' aptitude. *Language Learning* 55/1: 99–150.

Lasagabaster, David and Aintzane Doiz. 2003. Maturational constraints on foreign-language written production. In María Pilar García-Mayo and María Luisa García-Lecumberri eds. *Age and the Acquisition of English as a Foreign Language*. Clevedon: Multilingual Matters, 136–160.

Lasagabaster, David and Juan Manuel Sierra. 2009. Language attitudes in CLIL and traditional EFL classes. *International CLIL Research Journal* 1/2: 4–17.

Martín-Úriz, Ana María and Rachel Whittaker eds. 2005. *La Composición como Comunicación: Una Experiencia en las Aulas de Lengua Inglesa en Bachillerato*. Madrid: Ediciones de la Universidad Autónoma de Madrid.

Martínez-Adrián, María. 2020. The use of previously known languages and target language English during task-based interaction: A pseudolongitudinal study of primary-school CLIL learners. *EuroAmerican Journal of Applied Linguistics and Languages* 7/1: 59–77.

Martínez-Osés, Francisco and Joanne Neff-van Aertselaer. 2001. Corpus analysis of prepositional patterns and non-native university writing. In Carme Muñoz-Lahoz, María Luz Celaya-Villanueva, Marta Fernández-Villanueva, Teresa Navés and Oliver Struck eds. *Trabajos en Lingüística Aplicada*. Barcelona: Univerbook, 139–147.

Muñoz, Carmen ed. 2006. *Age and the Rate of Foreign Language Learning*. Clevedon: Multilingual Matters.

Nesselhauf, Nadja. 2004. Learner corpora: Learner corpora and their potential for language teaching. In John McH. Sinclair ed. *How to Use Corpora in Language Teaching*. Amsterdam: John Benjamins, 125–152.

O'Malley, J. Michael and Anna Uhl Chamot. 1990. *Learning Strategies in Second Language Acquisition*. Cambridge: Cambridge University Press.

Palacios-Martínez, Ignacio. 2005. Las nuevas tecnologías y la investigación en el campo de la adquisición de segundas lenguas. In Mario Cal-Varela, Paloma Núñez-Pertejo and Ignacio Palacios-Martínez eds. *Nuevas Tecnologías en Lingüística, Traducción y Enseñanza de Lenguas*. Santiago de Compostela: Servizo de Publicacións da Universidade de Santiago de Compostela, 203–224.

Rollinson, Paul and Amaya Mendikoetxea. 2010. Learner corpora and second language acquisition: Introducing WriCLE. In Jorge Luis Bueno-Alonso, Dolores González-Álvarez, Úrsula Kirsten-Torrado, Ana Elina Martínez-Insua, Javier Pérez-Guerra, Esperanza Rama-Martínez and Rosalía Rodríguez-Vázquez eds. *Analizar Datos > Describir variación/Analysing Data > Describing Variation*. Vigo: Servizo de Publicacións da Universidade de Vigo, 1–12.

Valero-Garcés, Carmen, Guzmán Mancho-Barés, Carmen Flys-Junquera and Esperanza Cerdá-Redondo. 2000. Evolución de la interlengua y análisis de textos: *ENWIL* y el análisis de errores en la expresión escrita en EFL. In Francisco José Ruiz-de-Mendoza-Ibáñez, Lorena Pérez-Hernández, Mercedes Fornés-Guardia and Juan Manuel Molina-Valero eds. *Panorama Actual de la Lingüística Aplicada: Conocimiento, Pensamiento y Uso del Lenguaje. Vol. 3: Adquisición y Aprendizaje de Lenguas. Diseño Curricular. Lengua con Fines Específicos*. Logroño: Mogar Linotype, 1840–1860.

Valero-Garcés, Carmen, Guzmán Mancho-Barés, Carmen Flys-Junquera and Esperanza Cerdá-Redondo. 2003. *Learning to Write: Error Analysis Applied*. Universidad de Alcalá: Servicio de Publicaciones.

Corresponding author
Zeltia Blanco-Suárez
Faculty of Education (Department of Philology)
University of Cantabria
Avenida de los Castros s/n
39005 Santander
e-mail: zeltia.blanco@unican.es

received: February 2020
accepted: March 2020