Dialogue and Argumentation for Cultural Literacy Learning in Schools

DIALLS

Multilingual Data Corpus

This project has received funding from the European Union's Horizon 2020 research and innovation Programme under grant agreement No 770045.

The sole responsibility of this publication lies with the author. The European Union is not responsible for any use that may be made of the information contained therein.

Contents

1. Description of the data

The Multilingual corpus is a deliverable of the DIALLS project (for more information see www.dialls2020.eu) and consists of a dataset of transcripts of classroom interactions of 5 to 15 years old students. The topic of the project and the classroom interactions is cultural literacy through dialogue and argumentation in school children.

A 15-session programme designed for three age groups:

Age group A	5-6 years old
Age group B	8-9 years old
Age group C	14-15 years old

The implementation was carried out in schools of the following countries of the project consortium:

UK	England - University of Cambridge
PT	Portugal - Nova University of Lisbon
DE	Germany - University of Munster
LT	Lithuania - Vilnius University
ES	Spain - University of Barcelona
CY	Cyprus - University of Nicosia
IL	Israel - Hebrew University of Jerusalem

The same programme was followed by the participants, pursuant to the Cultural Literacy Learning Programme. Two sessions implemented by a selection of participating classes were recorded. The recorded classes are distinguished as follows:

KL1	Key-point Lesson 1. Session number three of the Cultural Literacy Learning Programme
KL2	Key-point Lesson 2. Session number eight of the Cultural Literacy Learning Programme

2. Comparability of the data: Information on the lessons

The same session structure was implemented in every school. Every lesson had a cultural text (a short film or a picturebook) from which the activities were planned. For KL1, age A students watched the film *Ant*, age B students watched the film *Papa's Boy* and age C students analysed the picture book *Excentric City*. The theme explored in KL1 was 'Living Together', and the sub-theme changed according to the age group. KL2 had the same cultural text for all age groups (short film *Baboon on the moon*) and its main discussion theme was 'Being European' (more information on these and other cultural texts, and on the cultural analysis framework can also be found at the project's website). KL2 also had activity through the project platform with other participating classes. Table 1 shows the association between each Key-point Lesson and its cultural texts.

Age Group_Key-point Lesson	**A_KL1**	**B_KL1**	**C_KL1**	**A_KL2**	**B_KL2**	**C_KL2**
Theme	Living Together			Dispositions; Being European		
Sub-theme	Democracy	Equality	Celebration of diversity	Empathy; Belonging		
Cultural Text (type)	*Ant* (short film)	*Papa's Boy* (short film)	*Excentric City* (picture book)	*Baboon on the Moon* (short film)		
Year	2017	2010	2014	2002		
Author	Julia Ocker	Leevi Lemmetty	Béatrice Coron	Christopher Duriez		

Table 1: Cultural texts used, theme and sub-theme per session and per students' age group. The title of the cultural text has an embedded link to the project's online library.

3. Corpus description

Each partner selected the school and classes to be recorded, transcribed, and later translated. For each class, together with the recording of the whole class discussion, the interactions of two separate small groups (the focus and the backup) were recorded; only one of these two groups had to be mandatorily transcribed. The classroom sessions were recorded by DIALLS research team members with audio and video recorders. The transcriptions were made by the researchers or external transcribers and the translations by the researchers or external contracted translators.

The corpus of the recorded classes consists of 202 records. Due to the Covid limitations, the country in which the recording was most affected by the sanitary measures and the lockdown was Cyprus, with 19 records. The multi-lingual corpus has the structure presented in Table 2 below.

Country		UK		PT		DE		LT		ES		CY		IL		**Total**
Key-point Lesson		KL1	KL2	KL1	KL2	KL1	KL2	KL1	KL2	KL1	KL2	KL1	KL2	KL1	KL2	
Age Group	A	6	6	6	4	1	1	3	4	5	4	9	4	9	6	68
	B	5	5	5	4	5	5	5	5	4	4	6	-	7	5	65
	C	4	4	7	5	10*	7	10**	8	2	4	-	-	4	4	69
Sub-total		15	15	18	13	16	13	18	17	11	12	15	4	20	15	
Total		30		31		29		35		23		19		35		202

* 1 of them originally in English.

** 2 of them originally in English.

Table 2: Number of transcripts organized by country of origin, students' age group, and session.

The multilingual corpus had two specific purposes:

1. Allowing researchers working with one or more of the 7 languages included in the corpus or one or more countries involved in the data collection to carry out analyses on the specific data related to a specific country/language.
2. Allowing researchers to compare the data of different countries (cross-comparisons).

The multilingual corpus represented in Table 2 achieves the first purpose. However, to allow speakers of different languages to analyze (and compare, if needed) the data of all the countries involved, an English translation was needed. Considering the purpose no.2, the ideal number of transcripts to be translated per country was 19 (the highest number of transcripts collected by the country with the lowest number of transcripts). To this purpose, the teams of the other non-English speaking countries were requested to select a representative sample of a minimum of 19 transcripts to translate into the lingua franca for research, English. The original language transcripts together with their translations – aligned with the source files – led to a second corpus, which can be considered as "multilingual parallel corpus" which includes 123 bilingual transcripts. Table 3 displays the translated transcripts by country of origin, age group and Key-point Lesson.

Country		PT		DE		LT		ES		CY		IL		**Total**
Key-point Lesson		KL1	KL2	KL1	KL2	KL1	KL2	KL1	KL2	KL1	KL2	KL1	KL2	
Age Group	A	5	4	1	1	3	2	5	4	9	4	9	1	48
	B	2	2	5	1	2	3	4	4	6	-	7	1	37
	C	4	4	9	2	7	3	2	4	-	-	1	2	38
Sub- total		11	10	15	4	12	8	11	12	15	4	17	4	
Total		21		19		20		23		19		21		123

Table 3: Number of translated transcripts organized by country of origin, students' age group, and session.

The combination of the multilingual and the multilingual parallel corpus leads to a final corpus of 202 transcripts organized as follows (Table 4).

Country	**UK**	**PT**		**DE**			**LT**			**ES**	**CY**	**IL**		**Total**
Number of Transcripts	30	31		29			35			23	19	35		202
Language	**EN**	**PT-EN**	**PT**	**DE-EN**	**DE**	**EN**	**LT-EN**	**LT**	**EN**	**CA-EN**	**GR-EN**	**HE-EN**	**HE**	**Total**
Number of Transcripts	30	21	10	19	9	1	20	13	2	23	19	21	14	202

Table 4: Number of transcripts displayed by language or set of languages (in the case of bilingual transcripts).
EN=English; PT=Portuguese; DE=German; LT=Lithuanian; CA=Catalan; GR=Cypriot Greek; HE=Hebrew.

The diverse structure of the corpus, encompassing both translated and untranslated transcripts, is the result of two objectives that should be understood in two different axes, a vertical and a horizontal one. The vertical axis meets the goal of inclusiveness and reflects the multilingual dimension of the corpus, allowing

for a wider range of use of the data, for instance for analysis in each of the native languages. The horizontal axis meets the goal of cross-comparison, and therefore presents a significant share of the corpus in English (c. 77% of the corpus). A minimum of 19 transcripts translated into English was established for comparison purposes. Additional transcripts in or translated into English allow for further analysis, for instance, for binary comparison among countries.

4. Socio-economic information

DIALLS participant schools have different origins and socio-economic backgrounds. This heterogeneity represents a key point indicator for reaching the DIALLS objectives. Each country was responsible for gathering its own school level data. Although every category aims at a common objective, each country presented the data according to their own national policies, indicators, and criteria. There is nonetheless a common structure for presenting the data. School code refers to a code each team has given to the school in order to anonymize it. Age Group refers to the age group of the recordings from that school and not the sole age group of the entire school. Session in Corpus refers to the session that was recorded, allowing for analysts to check contextual data of the session.

The socioeconomic data can include the following: 1) Number of students in the classroom; 2) percentage of male students; 3) percentage of female students; 4) Socio-Economic Status (SES) criteria as defined by each country; 5) the percentage of ethnicity, specified for each country in the caption; 6) the location (whether in rural or urban areas); and 7) the legal status of the school, namely whether public of private.

These data are captured in tables. Each country's data, as well as descriptive summaries regarding these inputs, can be found below.

A. England

All participant schools in England were public and evenly distributed in rural and urban areas. Despite this mixed distribution, none of the schools in the age group C (14-15 years old) was in a rural environment. Considering the gender distribution, all schools represent a balanced sample between male and female students, the only exception being AI school, which was a single-sex school. Participant schools portray a high level of diversity amongst its students, shown by the percentage of students

classified as minority ethnic groups in the table below. Students eligible for social support measures are present in all schools not representing, however, the majority.

School Code	Age Group	Number of students in the school	% Male	% Female	*% SES	**% Ethnicity	Rural/ Urban	Public/ Private	Session in corpus
AA	A&B	279	48.7	51.3	3.6	14.7	Rural	Public	7, 8, 22, 23
AB	A&B	159	49.1	50.9	14.5	17.6	Rural	Public	1, 16
AD	A&B	100	47	53	7.2	15	Rural	Public	2, 9, 17, 24
AE	A&B	431	50.7	49.3	6	21.4	Urban	Public	3, 18
AG	A&B	416	46.6	54.4	5	49	Urban	Public	4, 19
AH	A&B	611	52.1	47.9	3.8	20	Urban	Public	10, 25
AI	C	1104	95.3	4.7	28.5	90.6	Urban	Public	12, 13, 27, 28
AJ	A&B	225	51.1	48.9	4.4	8	Rural	Public	11, 26
AK	A&B	415	50.4	49.6	8.2	18.3	Urban	Public	5, 20
AL	C	2084	49.1	50.9	2.7	11.7	Urban	Public	14, 15, 29, 30
AM	A	415	49.6	50.4	13.7	42.4	Urban	Public	6, 21

*SES = Social and Economic Status. % of students eligible for free school meals, training tools, etc.

**Pupils classified as minority ethnic (i.e., not white British origin). This can be compared with a national average in the population of 12.8%.

Table 5: Sociodemographic information for England.

B. Portugal

In Portugal, the schools where the sessions were recorded and transcribed are all located in urban settings. Almost all schools are public (i.e., state-owned), only one being private. Five schools are pre-primary, four of them primary (two hosting both age levels).

The average number of students participating in the project per school is 35, most of the schools (8) with less than thirty students, and one with over one hundred. The average of male students is 51%, with seven out of thirteen schools with more than half of students being male. Regarding ethnic/cultural diversity, three schools had no students meeting the criteria. The average of students is 22%, with five schools with cultural diversity up to 20%, three around 35%, and two schools where most of the participating students

(60 and 70%) met at least one of the cultural diversity criteria (foreign nationality, parents with foreign nationality, diverse cultural background).

School Code	Age Group	Students participating in the project*	% Male	% Female	**% Ethnicity	Rural/ Urban	Public/ Private	Session in Corpus
EA	A	24	42%	58%	0%	Urban	Public	5
EB	B	3	67%	33%	0%	Urban	Private	9
EC	C	113	55%	45%	0%	Urban	Public	16, 17, 18, 27, 31
ED	C	26	19%	81%	12%	Urban	Public	15, 30
EE	A	26	42%	58%	35%	Urban	Public	1, 20
EF	A	20	65%	35%	60%	Urban	Public	2, 21
EG	C	31	48%	52%	10%	Urban	Public	14
EH	C	57	40%	60%	3%	Urban	Public	12, 13, 28, 29
EI	B	20	65%	35%	75%	Urban	Public	10, 25
EJ	A	25	36%	64%	20%	Urban	Public	6, 19
	B	26	62%	38%	15%			7, 23
EK	B	22	50%	50%	36%	Urban	Public	8, 24
EL	A	15	47%	53%	20%	Urban	Public	3
	B	23	57%	43%	9%			11, 26
EM	A	20	65%	35%	35%	Urban	Public	4, 22

* Number of students present at the recorded sessions.

**% of students that have foreign nationality, and/or parents of foreign nationality, and/or having different cultural background.

Table 6: Sociodemographic information for Portugal.

C. Germany

All schools in the German sample are public and are in an urban environment. Most participant schools that implemented DIALLS belong to the age group C (14-15 years old) students, AD being the only school with age group A (5-6 years old) students. Since not all schools provided its information regarding gender, only a general framework can be extrapolated. On this regard, male and female students are well balanced in all schools, AF being the only exception with its percentage of males slightly higher than females.

Levels of ethnicity and students eligible for SES cannot be correlated, despite being very similar throughout all sample. The average of SES' percentage (students of the district under 25 that benefit from federal education subsidies) is 11.44%.

School Code	Age Group	Number of students in the school	% Male	% Female	*%SES	**% Ethnicity	Rural/ Urban	Public/ Private	Session in Corpus
AA	C	1418	44.70%	55.30%	10.17	37.4	Urban	Public	7, 8, 25, 26
AB	C	1046	NA	NA	12.89	43.5	Urban	Public	9, 10, 23
AC	B	267	52.40%	47.60%	8.25	36.6	Urban	Public	2, 3, 18, 19
AD	A	NA	NA	NA	10.45	NA	Urban	Public	1, 17
AE	C	NA	NA	NA	10.45	NA	Urban	Public	11, 12, 24, 27
AF	C	774	64.90%	35.10%	17.36	14.2	Urban	Public	13, 16, 28, 29
AG	B	180	46.10%	53.90%	6.78	27.8	Urban	Public	4, 5, 20, 21
AH	B	530	54%	46%	10.45	NA	Urban	Public	6, 22
AI	C	NA	NA	NA	16.15	NA	Urban	Public	14, 15

*Proportion of people in the district under 25 that benefit from federal education subsidies for school. AF, the value had to be estimated because the age categories that the district provides did not include "up to 25."

** Proportion of school's students with a "migratory background" (themselves or at least one parent with a foreign nationality).

NA = Data Not Available

Table 7: Sociodemographic information for Germany.

D. Lithuania

Lithuanian participant schools did not provide ethnicity percentages. This sample is a balanced mix of public and private schools, all private schools are in urban settings. The average number of students per school is 470. With an average of 47.8% male students and 52.1% female students, participant schools in Lithuania reveal that there's a gender balance across the sample. Rural schools are the ones with the highest percentages of SES numbers, although these differences are not significant.

School Code	Age Group	Number of students in the school	% Male	% Female	*%SES	Rural/ Urban	Public/ Private	Session in Corpus
DB	C	430	44%	56%	8%	Rural	Public	9, 10, 28
PA	C	987	51%	49%	2%	Urban	Private	17, 12, 33, 34
PM	C	950	49%	51%	10%	Rural	Public	13, 14, 15, 29, 30, 35
LU	C	567	55%	45%	1%	Urban	Public	16, 18, 31
NE	C	466	48%	52%	1%	Urban	Public	11, 32
KK	B	135	48%	52%	26%	Rural	Public	6, 7, 24, 25
LN	B	251	53%	47%	0%	Urban	Private	4, 5, 8, 23, 26, 27
EE	A	333	48%	52%	0%	Urban	Private	3, 22
AG	A	111	35%	65%	0%	Urban	Private	1, 2, 20, 19, 21

*Pupils eligible for free school meals, learning/training aids, etc.

Table 8: Sociodemographic information for Lithuania.

E. Spain

All schools in the Spanish sample are in urban areas and the majority are private, IMF school being the only exception. Gender distribution is evenly distributed amongst all schools in all age groups (average of 44% male and 47% female) and the ethnicity percentages are low, revealing sparse levels of diversity in this sample (the average of ethnicity rate in this sample is 1.5%).

School Code	Age Group	Total number of students	% Male	% Female	**% Ethnicity	Rural/ Urban	Public/ Private	Session in Corpus
LAG	A	450	59%	41%	0%	Urban	Private	1, 2, 12, 13
	B		57%	43%	1%			6, 16
	C		52%	48%	0%			10, 11, 20, 21
IMF	C	1400	51%	49%	5%	Urban	Public	22, 23

PAU	A	1125	49%	51%	0%	Urban	Private	3, 4, 5, 14,15
	B		49%	52%	0%			7, 8, 9, 17, 18, 19

** Percentages provided by schools.

Table 9: Sociodemographic information for Spain.

F. Cyprus

All Cypriot schools were public and most of them were in rural settings, except for AM and AS schools (AO school didn't provide this data). Cyprus only had A and B age groups students implementing DIALLS and these are evenly represented when it comes to gender distribution amongst all schools (average of 52.4% male and 47.5% female students). Schools in urban locations have similar ethnicity rates (average of 10%), but in rural institutions this indicator is more unbalanced (average of 13.2%).

School Code	Age Group	Number of students in the school	% Male	% Female	*% SES	**% Ethnicity	Rural/ Urban	Public/ Private	Session in Corpus
AA	B	327	51.99	48.01	12.23	6.12	Rural	Public	11
AE	A	124	41.13	58.87	NA	25.81	Rural	Public	5
AF	A	23	52.17	47.83	17.39	4.35	Rural	Public	6, 17
AJ	A	91	56.04	43.96	49.45	12.09	Rural	Public	1, 4
AL	A	49	51.02	48.98	0	12.24	Rural	Public	7, 9, 18
AM	A	123	53.66	46.34	7.32	10.57	Urban	Public	2
AN	A	43	48.84	51.16	0	6.98	Rural	Public	19
AO	B	NA	NA	NA	NA	NA	NA	Public	13
AP	B	152	61.84	38.16	11.84	10.53	Rural	Public	10, 14
AQ	B	205	49.27	50.73	4.88	5.85	Rural	Public	15
AR	B	90	52.22	47.78	13.33	41.11	Rural	Public	12
AS	A	74	48.65	51.35	2.7	9.46	Urban	Public	3
BC	A	85	62.35	37.65	5.88	7.06	Rural	Public	8, 16

* % of students who are entitled to free school meals.

** % of students who have at least one parent of a foreign nationality.

NA = Data Not Available

Table 10: Sociodemographic information for Cyprus.

G. Israel

Data from the schools in Israel refers only to the students that participated in the CLLP sessions. The average number of students per school is 74.9. Only one school had more than one hundred students partaking in the project, and four out of ten had less than fifty students. Female students represent 50% or higher of the student participants in eight schools. Regarding the territorial setting, only two of the schools, both of A and B age group, were rural. All the schools are public. Two schools had sessions with C age group students, representing 12.14% of students from Israel engaged in the DIALLS.

School Code	Age Group	Students participating in the project	% Male	% Female	Rural/ Urban	Public/ Private	Session in Corpus
ET	A&B	83	51%	49%	Urban	Public	9, 13, 14
EP	A&B	49	43%	57%	Rural	Public	6, 12, 21, 30
EO	A&B	56	58%	42%	Rural	Public	5, 16, 29
EN	A&B	45	49%	51%	Urban	Public	4, 11, 24, 27
EF	A	46	39%	61%	Urban	Public	3, 23
EA	A&B	98	64%	36%	Urban	Public	1, 2, 10, 22, 28
EZ	A&B	190	50%	50%	Urban	Public	15, 26, 31
ES	A&B	91	46%	54%	Urban	Public	7, 8, 25
JS	C	51	49%	51%	Urban	Public	17, 20, 33, 35
JJ	C	40	43%	58%	Urban	Public	18, 19, 32, 34

Table 11: Sociodemographic information for Israel.

5. Organization of the transcripts

The transcripts were produced in Excel (.xlsx) files following a pre-established structure. The transcripts are therefore organized in different columns. Below in Table 5 are displayed the items of each column.

Line number	Time	Activity	Gender	Speaker	Speech	Translation

Table 12: Items of the transcript.

Below the explanation of each item:

Line number: Sequential number of turns or lines written (for instance, with contextual information by the transcriber). In the excel file, the line number is sequential in the entire sheet, i.e., does not restart at 1 in a new transcript session. The last line of each session is in bold to provide a visual division from the next session.

Time: Time stamp at the beginning, end and whenever there was a change in Activity, or the transcriber felt necessary. For the ones from England, the time stamp was done for every turn.

Activity: The activity type was coded using the following pre-defined codes: WC, when a teacher-guided whole class discussion activity takes place; SM, when a small-group discussion activity takes place; TG, when teacher is having separate discussions with our pre-selected group during a small-group activity. SM and TG are mutually exclusive, and they differ by the presence of the teacher interacting with the group. SM and TG refer only to the selected group, i.e., only the previously selected group interactions were transcribed.

Gender: M (male), F (female), U (Unapplicable - whenever many students were talking at the same time in whole class discussions), [blank] (whenever it was not possible to identify the gender in individual talks).

Speaker: R (the researcher that was present at the session); T (or T1, T2, …) refers to teachers; S1, S2, …, S_n refers to students; S refers to an unidentified student; Ss refers to more than one student speaking at the same time.

Speech: Transcribed speech following the conventions and transcription guidelines established by the project team members, see Table 13.

Translation: Translated speech into English (only for the bilingual transcripts).

In .csv files, an extra column, after ‘Translation’, has an indication of overlapped speech. Whenever there is overlapping, in that column there is the note [overlapped] in the lines where there is overlapped speech.

Below in table 13 the convention system used (adapted from the Jefferson system) and description of each item.

Symbol	Phenomenon
[…]	Long pauses. At the end of a turn indicates a gap before the next speaker speaks.
-	Abrupt cut.
SPEech	Higher volume/emphasis of a (part of a) word or phrase. If the language has no distinction between lowercase and capital letter (Hebrew), it is signaled with bold letter (**SPEECH**).
[ ]	Overlapping speech. The two or more cells containing overlapping speech, i.e. when two or more speakers talk at the same time are also coloured in yellow. [ ] indicates the part of text being overlapped. In .csv files, only the symbols [ ] are visible.
{ }	Inaudible segment not transcribed, only partially transcribed or reconstructed, or omitted transcriptions. Omitted transcriptions could be: {unclear}: inaudible speech. {speech}: partially transcribed and/or reconstructed speech by the transcriber. {off-task}: students' misbehavior (talking among themselves on irrelevant topics, goofing around, …) {class management}: the teacher talks about other activities not related with the class (e.g. a visit to a museum planned for next week), teacher reprimand, etc.
(writes)	Description of a relevant nonverbal activity or *relevant* noise that allows understanding the activity (gasps, chatter sounds…), relevant indications allowing the understand the addressee of speech, relevant gestures important for understanding the conversation and meaning of turn.
Punctuation	'.' at the end of the turns (when not interrogative or exclamative sentences). ',' used to separate sentences (especially when it would be confusing to read this part of the transcript without a comma), and for lists of facts. '?' used to indicate questions. '!' used to indicate orders and sentences used to express *surprise* and similar expressions. '?!' used to indicate *astonishment*.

Table 13: Convention system use and description.

6. Anonymization of data

Students and teachers' names were anonymized as described above (see speaker description) to identify speakers. School mentions were anonymized with a code (e.g., EC, ED, LU, etc) which was not related to the school acronym. Only students and teachers' name anonymization follows the same structure for each partner (S for students and T for teachers). Each partner was responsible for anonymizing their own data, which was not shared with the other country partners, in order to follow General Data Protection Regulation procedures.

7. Labeling of the records (transcripts and translations)

The dataset is organised in two main sections: an Excel file, and a zip folder with .csv files matching the excel file.

The excel file is divided into several sheets, each sheet representing a country. Each country page contains, organised by session (starting at KL1), the transcripts and translations of the implemented sessions (first the age group A, then B, then C). The first sheet in the dataset presents the index of all the sessions; the index has hyperlinks to the transcripts for easier navigation.

Files are named following the structure: Country_Number of Session_Age group_Key-point Lesson_Class number_language(s) (e.g. UK_1_A_KL1_C1, UK_16_A_KL2_C1, DE_1_A_KL1_C1_DE&EN, DE_17_A_KL2_C1_DE&EN).

Country refers to the countries of recording of sessions: UK= England; PT=Portugal; DE=Germany; LT=Lithuania; ES=Spain; CY=Cyprus; IL=Israel.

Number of sessions is the sequential number of transcripts within each country.

Age group refers to three school stages of the classrooms: A=pre-primary, 5-6 years old; B=primary, 8-9 years old; C=secondary, 14-15 years old.

Key Lesson refers to one of the two Key Lessons transcribed. KL1 =Key Lesson 1; KL2= Key Lesson 2.

Class number refers to class/group of students. Class number does not change for the same class/group of students from KL1 to KL2 (e.g. PT_1_A_KL1_C1_PT&EN --> PT_20_A_KL2_C1_PT&EN). A match of the class number indicates it refers to the same class/group of students.

Language(s) indicates the language of transcript and if it is translated.

The zip folder has the .csv files. Each session is a separate file, and the sessions are organized by country, which are separate sub-folders. Next to the sub-folders is an .xlsx file with the index with hyperlinks for the files for easier navigation; the index file matches the index file of the .xlsx major file. The naming of the files matches the ones in the index of the .xlsx major file (the corpus in .xlsx).

8. Format and software required to read the files

The corpus was built in excel (.xlsx) and then also saved as .csv files. The corpus and the index files of the .csv section are in .xlsx. xlsx files can be opened with spreadsheet software (e.g., Apple Numbers, OpenOffice Calc, LibreOffice Calc, Google sheets, …) but for optimal use Microsoft Excel should be preferred. .csv files can be opened with notepad or spreadsheet software. .zip files and folders can be

opened with the default option (e.g. in Windows, right click, then the option ‘extract’ or with .zip creater or opener software (e.g. Winzip, WinRAR, 7Zip, ZipExtractor, …).

9. Changes from Version 1 to Version 2

Version 2 of the corpus incorporates changes and corrections. The most relevant changes to report are:

- Indication if sessions/pairs of sessions refer to the same class/group of students (information available in the index);
- Addition of one session: ES_15_A_KL2_C4_CA&EN;
- Revision of translations of all sessions of ES;
- Change of labelling of sessions ES_15 – ES_22 (version 1) --> ES_16 – ES_23 (version 2);
- Change of reference of language: Spanish (ES) --> Catalan (CA), Greek --> Cypriot Greek;
- Revision of content of transcripts UK_14_C_KL1_C14, UK_30_C_KL2_C14, LT_27_B_KL2_C5_LT, LT_28_C_KL2_C9_LT&EN, ES_8_B_KL1_C8_CA&EN, ES_14_A_KL2_C3_CA&EN, ES_21_C_KL2_C11_CA&EN, CY_4_A_KL1_C4_GR&EN, CY_5_A_KL1_C5_GR&EN, CY_11_B_KL1_C11_GR&EN, CY_12_B_KL1_C12_GR&EN;
- Minor revisions/corrections throughout the corpus with no significant impact on the content of the corpus.

10. Licensing of the corpus