Journal of Educational Research and Innovation

Volume 12 Article 1

2024

Input and Output in an English Classroom: A Young Learner Corpus of English (YoLeCorE)

Thomas Zapounidis
Aristotle University of Thessaloniki, thomasza@enl.auth.gr

Marina Mattheoudakis
Aristotle University of Thessaloniki, marmat@enl.auth.gr

Recommended Citation

Zapounidis, Thomas and Mattheoudakis, Marina (2024) "Input and Output in an English Classroom: A Young Learner Corpus of English (YoLeCorE)," *Journal of Educational Research and Innovation*: Vol. 12, Article 1.
Available at: https://digscholarship.unco.edu/jeri/vol12/iss1/1

Journal of Educational. Research and Innovation
2024, Vol. 12, No. 1

Input and Output in an English Classroom: A Young Learner Corpus of English (YoLeCorE)

Thomas Zapounidis
Aristotle University of Thessaloniki

Marina Mattheoudakis
Aristotle University of Thessaloniki

Despite research findings that underline the value of both input and output in instructed second language acquisition, there has been, to our knowledge at least, hardly any longitudinal research aiming to shed light on the relation between the language input students are exposed to and the subsequent development of their output within a given instructional setting. As there is a dynamic interaction between L2 input and output (Crossley et al., 2016), focusing on the relation between the input learners receive in a specific instructional setting and their written and oral output will provide us with significant data regarding the second language acquisition process in instructional settings. Also, taking into consideration the multitude of variables that impact the second language acquisition process as well as the final learning outcomes achieved, a longitudinal study that involves the same group of learners being taught by the same teacher over an extensive period of time might give us the opportunity to control at least variables related to the instructional setting (cf. Bestgen and Granger, 2014; Laufer and Waldman, 2011; Meunier, 2011; Paquot and Granger, 2012).

What is more, given the complementary nature of input and output in language learning (VanPatten, 2002), knowledge of both what learners have been exposed to and what they actually produce may be of value to a number of instructional practice stakeholders such as teachers and curriculum designers, among others. Similar corpora are expected to allow teachers to account for the input and output of language in terms of quantity and quality and evaluate whether certain input needs to be re-taught, further practiced or not. Similarly, curriculum designers may use such corpora to adjust or improve the content of the material taking into consideration learners' output in relation to the input they received. In other words, material found to be too challenging or inappropriate for learners' language level, in terms of the Common European Framework of Reference (CEFR) could be modified or changed.

The use of corpora in language education has been of great value in a number of ways. For instance, corpora have been used for (a) the accurate description of both target and learner language (Leńko-Szymańska, 2014), (b) the creation of language material (McCarthy et al., 2005; Lee and Swales, 2006), use by the learners in the language classroom (Aijmer, 2009; Campoy-Cubillo et al., 2010) just to mention a few. What is of great value in foreign language teaching and learning seems to be the approach of using learners'/teachers' data for raising language awareness. Indeed, Data Driven Learning (DDL) which consists of "using the tools and techniques of corpus linguistics for pedagogical purposes" (Quilquin and Granger, p.359) includes significant benefits. As O'Sullivan (2007) notes, the use of DDL entails a number of important skills that include "predicting, observing, noticing, thinking, reasoning, analysing, interpreting, reflecting, exploring, making inferences (inductively or deductively), focusing, guessing, comparing, differentiating, theorising, hypothesising, and verifying" (p.277). Above all, as Gabrielatos (2002) rightly mentions, DDL may transform teachers from mere 'skilled materials operators' to educators that "focus more consistently on research skills, as well as language analysis and its implications for ELT" (p.3). In this light, corpora have the potential to raise both teachers' and students' language awareness as well as research skills (Coniam,1997) through a number of particular corpora techniques and tools.

Thus, the examination of types (unique words in texts) as well as tokens (total number of running words in texts) give valuable information about the word frequencies produced in language input and output. The most important tool for conducting such searches is the concordance which displays the target language in one-example-per-line format (McEnery and Hardie, 2012). In this way, students and teachers view the target language within its closest and wider context and form assumptions or establish rules regarding the words' collocation.

The development of large learner corpora over the past 20 years, such as the International Corpus of Learner English (ICLE), has allowed us to acquire a better understanding of second language acquisition processes; however, we have not been able to relate students' output with the input they are exposed to, since to date there have been no corpora that combine both types of information. Even though pedagogic corpora are expected to contain "all the language a learner has been exposed to" (Hunston, 2002, p.16), there is hardly any pedagogic corpus that contains the full range of input students actually receive in a particular classroom. While there are noteworthy projects such as the Linguistic Barriers to Transition (University of Leeds) and the Grammar and Growth Project (University of Exeter) which include data from various schools, they still do not include the totality of spoken and written productions from each and every student for a whole school year.

Our study aims to fill this gap by presenting a comprehensive database that records all the instructional input received by a group of young learners, as well as the written and oral output they produced within a given formal instructional setting over an entire school year. The corpus compiled does not include any type of input those learners were exposed to outside the classroom setting. It is quite likely that most, if not all, of those students were exposed to English language input through social media, YouTube, or video games in English and such leisure activities that increase their input in English are also expected to impact their output as well. However, access to this input is not possible and therefore, this is a variable we cannot control.

The corpus is called *YoLeCorE (Young Learner Corpus of English)* and its content allows researchers to examine students' output and its relation to the input received. Such a development is positioned within the existing traditions of examining instructional input through the compilation of pedagogic corpora and learner output as it is recorded in learner corpora. We believe that the compilation of a new corpus that includes both the language input and the language output of instructional sessions is an important development in corpus research as it sheds light on the relation between the two – input and output – and allows us to draw valid inferences about second language acquisition within instructional settings.

Study

This paper aims to present *YoLeCorE,* a corpus which combines a learner and a pedagogic corpus since it records the input received and the output produced by 17 young Greek learners of

English as a Foreign Language (henceforth EFL) within a formal educational setting. Our paper will present the design of *YoLeCorE,* its rationale, as well as some quantitative data, in particular: (a) the total number of tokens and types students in this particular classroom were exposed to during a school year, that is, the quantity of EFL input, and (b) the total number of tokens and types produced in the particular classroom over a school year, that is, the quantity of EFL output (Zapounidis, 2017, p. 103).

Setting and Participants

YoLeCorE includes the recording and transcription of all class activities that took place in an EFL classroom at the 3rd Experimental Primary School of Evosmos (henceforth, 3rd EPSE) in Thessaloniki, Greece. The 3rd EPSE is a state school supervised by the Department of English Language and Literature, Aristotle University of Thessaloniki (AUTh), Greece. As the school is supervised by the English language Department, special emphasis is placed on the teaching of EFL; this translates into more hours of EFL instruction from grades one to six, as compared to other Greek state schools, but also into the use of more advanced language syllabi and instructional materials. At the 3rd EPSE, EFL instruction is introduced in grade one; First graders attend EFL classes on a daily basis and as of grade three, there is a gradual increase of EFL instructional time. In particular, in grades three and four, students attend EFL classes six hours a week and in grades five and six, the instructional time is eight hours a week. The school curriculum follows the Common European Framework of Reference (henceforth, CEFR) language standards but teachers use authentic materials, such as fairy tales and short stories written for native young speakers rather than EFL coursebooks or graded readers.

Participants included 17 students (seven girls and ten boys) in grade four (9-10 years of age); only two of them had a mixed ethnic background and only one of them was bilingual (Russian and Greek). All students received EFL instruction exclusively at the particular school and none of them attended private or extra-curricular classes in English. Of course, as already mentioned, we cannot know how variably each one of them engages with English outside of the classroom; this is a variable over which the researchers have no control.

The Design of *YoLeCorE*

YoLeCorE includes both the input those students were exposed to during their EFL class as well as their output, i.e., their oral and written productions. Their input includes the aural input, i.e. teacher language, other people's utterances in L2 when they visited the specific classroom and interacted with the students (i.e., other teachers, teacher trainees, researchers, parents, the school advisor, and the school Principal), all the listening activities and anything that was read aloud. Their input also includes the written input, i.e., all the printed or projected material students were required to read (worksheets, writings on the board, video subtitles, etc.).

The corpus was also designed to include the oral and written productions of all learners in the particular classroom. The former includes the total number of utterances produced by all learners and the latter includes all forms of written production, including writings on the board, in notebooks, tests and worksheets.

The Compilation of *YoLeCorE*

The compilation of *YoLeCorE* was a complex and time-consuming process. It followed a series of stages, all of which were performed by the researcher who was also the EFL instructor of this particular class. Familiarity with the teacher inspired trust vis a vis the decision to video-record the class for a whole school year.

Similarly, familiarity with students' families also facilitated the research procedure, as parental consent prior to the research was necessary and families felt quite comfortable to give their consent knowing that the researcher was also their children's teacher.

Instruments and Procedure

For the creation of *YoLeCorE,* we needed to purchase equipment for the recording and storage of data. Three High-Definition video cameras, memory cards of 32GB for each camera, long lasting batteries, tripods and hard disks of a total of 4TB capacity were obtained. Two cameras (the third one was used as an emergency backup) were placed on two opposite corners of the room so that they would not distract learners' attention. By placing them there, the researcher aimed to minimize the observer effect (or Hawthorne effect), in this case, the potential effect cameras might have on students' behavior (Scheyder, 2012). The decision to use multiple cameras instead of a single one was dictated by reasons related to the research design of our study. In specific, multiple cameras allowed us to discriminate between speakers in case of overlapping voices. Compared to other forms of oral data, classroom language often includes overlapping voices (e.g. choral repetition) and a single camera would not be able to capture all learners' oral contributions in class. Additionally, given that one of our aims was to measure learners' written input and output in the classroom, the camera placed at the back of the room allowed us to record what each learner, as well as the teacher, wrote on the whiteboard, in other words, learners' written output and teachers' written input.

The recording of the lessons started on October 1st, 2012, and was completed on June 15th, 2013. All students in the particular school attend at least 60 minutes of EFL instruction on a daily basis. In total, about 8,850 minutes of EFL instruction were recorded; this includes both teacher and student talk. As far as teacher talk is concerned, this is delivered almost exclusively in English; student talk, however, may also include utterances produced in Greek, students' L1. Each recording was codified according to the number of the camera and the date of the recording (e.g. CAM1_Jan_10, CAM2_Jan_10). Data was organized into folders according to the month of the recording. This allowed easy retrieval and comparison of the data collected. Additionally, written data was further divided into: (a) input, which included anything they read and was further categorized into sub-corpora depending on the reading source (e.g., short stories or readers, worksheets, projected material on the whiteboard), and (b) output (which included any activities written in their notebooks or written on the whiteboard and was equally represented in sub-corpora).

During all classes, the teacher-researcher recorded the names of absentee students and made sure that the input delivered to the rest of the class on that day would not be added to the input of the absent students. This allowed the accurate computation of input per student. The list also included temporary absentees; in particular, the teacher noted down the names of students who were absent for a short period of time from class (e.g. visit to the toilet) so that similar modifications would be made to the input received by this particular student, i.e., removal of the input they had missed due to their short absence.

The next stage was quite time-consuming as it entailed the transcription of all recorded material. Although there are various types of transcriptions available such as selective, comprehensive, clean or smooth, pure verbatim, special character, comment column (Mayring, 2014), the pure verbatim was chosen as it includes a word-for-word transcription and some fillers.

This choice aimed to ensure that the corpus would include all language instances occurring in this classroom so as to facilitate future research utilizing the particular data.

For every speaking turn, the transcriber indicated at the beginning of each utterance its source (e.g. teacher, the name of the student codified as S1, S2 etc.) so that it could be easily attributed to the corresponding person during the analysis stage. The transcription of the 177 instructional hours took the researcher about two years, approximating about 2,600 working hours. Although the transcription was a long and time-consuming process, this was undertaken exclusively by the researcher for two main reasons. First, as a teacher of this class, he was able, when transcribing, to assign the recorded utterances to the corresponding students. Additionally, he was the only person granted permission by parents to use and process the recorded data.

The whole transcription process was done manually without the use of speech recognition applications as the latter, during the period of transcription, were not accurate enough or were unable to discriminate against students' overlapping voices. Of course, the advent of more robust speech recognition applications is already taking place at present and this definitely enables researchers and teachers to create their classroom corpora with greater ease and within less time. If governments funded educational institutions for the purchase of commercial speech recognition software or if universities created such state-of-the-art software and distributed for free, then researchers and teachers would, of course, be more likely to produce similar types of corpora.

The transcription of the present study produced a Word file of 4,825 single spaced pages that included all language input and output – oral and written – produced by the particular group of fourth graders and their teacher during an entire school year. The researcher checked the reliability of the transcriptions by repeating the transcription for random parts of the corpus selected from the file of each month.

A Word file was chosen for the transcription of the recorded data mainly for practical reasons. In particular, it was easier for the transcriber to use the word processor to record the transcribed data and, as a Word file is easily converted to other corpus-software friendly formats (e.g. txt), this was considered to be an appropriate and convenient choice. What is more, a single Word file that includes the total classroom language is more useful to researchers who are interested in examining the total classroom language for a number of relevant topics such as turn-taking, teacher's or peers' corrective feedback and so on.

The next stage involved classifying the recorded data into input and output and matching it with the corresponding student. Initially, 17 folders were created (one for each student, see Figure 1), each one containing four distinct folders (one per language skill). In this way, the raw data produced two sub-corpora (a) the input sub-corpus (including the listening and reading skills sub-corpora, marked with L and R, respectively, in Figure 1), and (b) the output sub-corpus; (including the speaking and writing skills sub-corpora, marked with S and W, respectively, in Figure 1). Each language skill sub-corpus was further sub-categorized according to its source. More specifically, all teacher utterances formed one of the input sub-corpora ('teacher talk sub-corpus'); this was learners' *listening input* and as such it was placed in learners' listening folders. The rest of the listening input (e.g., audio material used in class, other people's utterances, etc.) was processed in the same manner, thus creating more listening skills sub-corpora ('audio material sub-corpus', 'other teachers sub-corpus', etc.) for the learners. In short, the listening folder of each learner included three distinct listening sub-corpora: ‘teacher talk’, ‘audio material’, and ‘other people’. The listening input is almost identical for all learners given that they were all exposed to the same oral input; there were only slight differences in the case of the absentees whose input on the day of their absence was not included in their folder.

Similarly, the *reading input* sub-corpora included all worksheets, tests and other reading material read by learners in the classroom or assigned for homework (e.g. prompts in worksheets and assigned readers). In order to discriminate between spoken and read material (as read aloud sentences are also uttered), all videos had to be viewed again so that only those utterances produced by learners while viewing and reading a text from a book (or something written on the board), would be included in the reading input folder.

As with the listening input, the reading input was also further subcategorized. The first reading sub-corpus included anything read inside the classroom, such as books, notes etc. The second source of reading input included material projected on the whiteboard that students were required to read (e.g. texts, song lyrics, etc.); the third source of reading input included the readers assigned for homework every week. Although there was no further discrimination between ‘spoken reading’ and ‘written reading’, the sub-corpora of only read materials (e.g. assigned readers) and read aloud materials (e.g. reading from whiteboard and textbooks) allows such comparisons.

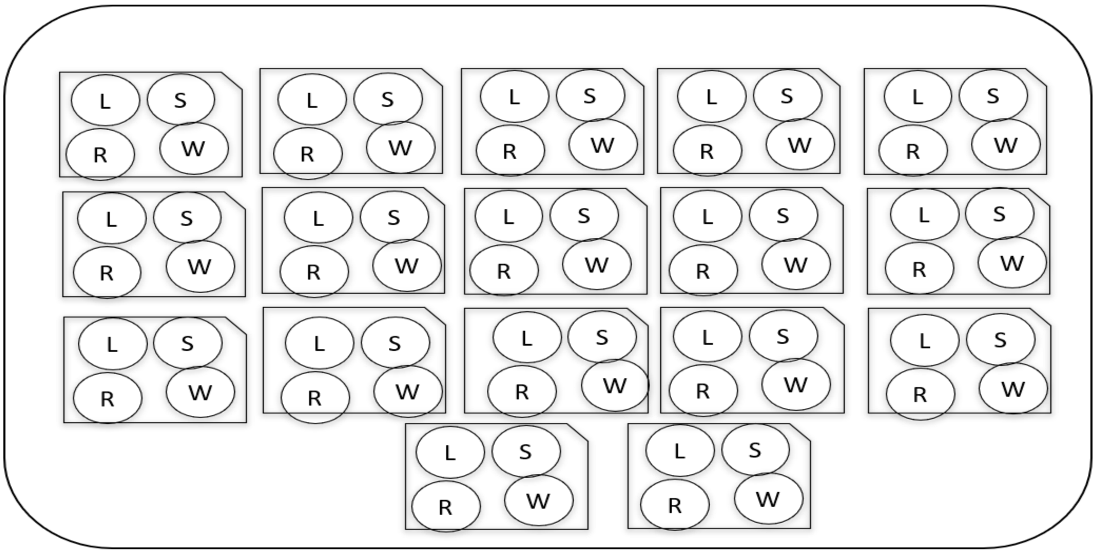

Figure 1. Students’ sub-corpora

With reference to students’ output, this was divided into two sub-folders, one for the spoken and one for the written output. The spoken output included the utterances of each learner during the whole school year; these were coded in files indicating (a) the skill, (b) the speaker, and (c) the date (e.g. Speaking_ S1_Feb_1). These files made up a series of spoken sub-corpora for each learner. As for the written output, this included each student’s written productions, including tests, notebook writings, and even words, sentences, etc. they wrote on the board. These were included in the appropriate sub-folder (written output) for each student. Although a tag for each of these genres or types of text might also characterize each type of written source, no such practice is followed at this stage as the corpus has not yet been extensively annotated.

To further discriminate between the different types of written output, three distinct sub-corpora of learners' written output were created. The first one (marked as 'conventional or typical' written output) included students' written productions in tests, paragraph writing and worksheets. The second sub-corpus (marked as 'classwork') included anything that learners wrote on the whiteboard or typed on their computer in class; the third one (marked as 'notebook notes') included students' written productions in their notebooks.

The full list of available sub-corpora per skill and source is provided in Figure 2 below. The organization and structure of *YoLeCorE* allows researchers to treat this data as representative of young learners' second language development in an intensive EFL instructional setting and thus use it in order to track the second language acquisition process as well as measure its growth over a school year. Additionally, and perhaps more importantly, researchers may look into individual students' output and compare their language development by examining differences in the quantity and quality of output depending on the quantity and quality of input. In this sense, *YoLeCorE* provides an excellent source of data for the study of variability in instructed second language acquisition settings.

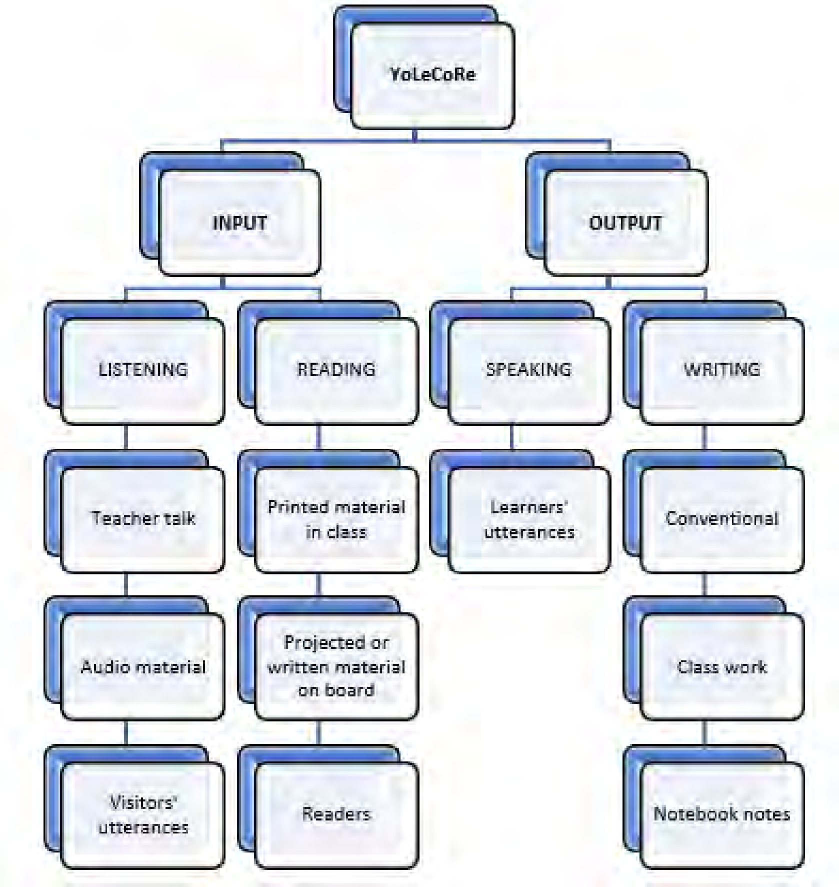

Figure 2. The structure of *YoLeCorE*

Corpus Analysis - Quantitative Data

The present section aims to provide quantitative data regarding the size of *YoLeCorE* and its sub-corpora. The tool used for measuring the number of types (unique words) and tokens (number of all words in a text regardless of how many times they are repeated) was *AntConc* (Anthony, 2014). The results indicated a total of 1,487,240 tokens (see Table 1). The inclusion of types in the above table and throughout this analysis is important as, compared to the total number of tokens, it gives additional information regarding the lexical diversity.

Table 1. The size of *YoLeCorE*

	INPUT		OUTPUT	
	Listening	***Reading***	***Speaking***	***Writing***
Types	4,905	2,937	4,640	3,356
Tokens	415,776	508,327	479,901	83,236
TOTAL:	**924,103**		**563.137**	
Total: 1,487,240 tokens				

Note: (Taken from Zapounidis, 2017, p.218)

An examination of learners' quantified input and output in Table 1 indicates that the input was greater than the output. Indeed, the total input includes almost a million tokens (924,103) compared to the output which totals a bit more than half a million tokens (563,137). This is natural, since output builds on input, and as participants are young EFL learners, their output is expected to be much lower than their input, especially when it comes to the writing skill.

There are also differences in the number of types and tokens within each skill of the input and output. Regarding input, overall, students' reading input is larger than their aural input. In other words, they receive more input from reading than from listening (508,327 vs 415,776). However, there is a more systematic recycling of words in the reading input compared to that of the listening input. That is why the types in the reading input are fewer than the corresponding number of types in the aural input. This is to be expected, of course, as it is easier for the teacher to manipulate students' reading input and modify it in order to suit learners' proficiency level rather than the aural input, which, quite often, is spontaneous or may not always derive from the teacher but also from other sources that cannot be monitored. With reference to the difference in types of the spoken and written output, the difference is very small (4,640 vs 3,356), which indicates that students have developed an active vocabulary of about 4,000 items which they can use both in speaking and in writing.

An examination of the number of types and tokens within the sub-corpora of each skill also renders interesting insights. For instance, Table 2 aims to compare the aural input students received from different sources. The last variable (other people) was used as part of the aim to discriminate between the teacher and any other input and it seems to indicate that even people visiting classrooms may have a slight impact on learners' output. Based on the analysis, learners listened to 415,776 tokens, the majority of which (91.7% see Table 2) originated from the teacher while 6.67% came from other media and the rest 1.62% were produced by other teachers or people visiting the classroom. With reference to the total number of 4,905 types, the overwhelming majority (89.17%) of them originated from the teacher, while the audio files also included a considerable percentage (33.7%).

Table 2. Sources of Listening Input

Sub-corpora	*Types*	*% types*	*Tokens*	*% tokens*
Teacher	4,374	89.17	381,278	91.70
Audio file	1,653	33.7	27,743	6.67
Other people	911	18.57	6,755	1.62
Total	4,905		415,776	

Note: (Taken from Zapounidis 2017, p.140)

The total number of types is not the sum of the rows above, as there is a number of overlapping types in the various sources. Given that types represent the unique words of a source, the total number of types (4,905) represents the unique words of the combined sources.

The types and tokens per reading source are presented in Table 3. Out of the 508,327 tokens, the *printed* sub-corpus, which includes reading students did in class, covers 22.57% of the tokens, while the *board* sub-corpus, which includes the reading of words or phrases written or projected on the whiteboard, covers 62.03%. Finally, the *readers* sub-corpus, which includes about 15 short story books learners were required to read, covers 15.4% of the total number of tokens. The percentages seem to be more or less the same in terms of types with the exception of the *reader* sub-corpus, as the percentage of types in this sub-corpus is 56.79% compared to the 15.40% of tokens.

Discussion and Concluding Remarks

YoLeCorE is an original and pioneering corpus because it is a database that includes all the input and the output produced within a specific EFL formal instructional setting by a group of 17 fourth grade students over a whole school year. Thus, *YoLeCorE* is a longitudinal corpus of L2 English, which includes spoken and written data produced by the same group of learners taught English as L2 by the same teacher within a primary school classroom. In this sense, we might suggest that *YoLeCorE* is a special type of a corpus, as it combines instructional input from various sources with learner output. Thus, we might suggest that it is both a learner corpus (a database of spoken and written texts produced by learners) and a pedagogic one (a database of language input they are exposed to within the EFL class).

Table 3. Sources of Reading Input

	Types	*% types*	*Tokens*	*% tokens*
Printed material	953	32.44	114,724	22.57
Board	1,763	60.02	315,310	62.03
Readers	1,668	56.79	78,293	15.40
Total	2,937		508,327	

Note: (Taken from Zapounidis, 2017, p.147)

Once again, the total number of types is not the sum of the rows above, as there is a number of overlapping types in the various sources. Given that types represent the unique words of a source, the total number of types (2,937) represents the unique words of the combined sources.

*YoLeCorE'*s unique characteristic is that it is a highly robust database, since it includes *all* the language that each and every one of the students listened to, read, or produced, in speaking and in writing, within a specific instructional setting over a whole school year. To our knowledge, this is the first L2 corpus that is based on the systematic video-recording of classroom instruction over a whole school year. Therefore, the major contribution of the corpus lies in the possibilities it offers for research in language development within formal educational settings where instructional input is expected to affect learners' output and where input and output interact in dynamic ways. In particular, *YoLeCorE* gives researchers the opportunity to trace the longitudinal development of lexical diversity and lexical density in learners' output and draw valid inferences about EFL learners' language development in similar educational settings which promote intensive EFL instruction. Such data are also expected to allow them to look into the impact of instructional input on students' output and make associations between specific instructional techniques (e.g. drilling, role plays, etc.) and development of students' fluency and accuracy. Additionally, as *YoLeCorE* consists not only of sub-corpora reflecting each language skill but also of sub-corpora reflecting different sources within each skill, comparisons between different types of input within the same skill are also possible (e.g. aural input from teacher versus aural input from YouTube or other audio material). Such comparisons are expected to shed light on the different impact of each source of input on students' language development.

Beyond studying the L2 development of a group of students, researchers are also interested in studying variability in L2 development, as second language acquisition is highly variable due to a variety of factors, including individual differences and exposure conditions, among others (Tagarelli et al., 2016). The contents and structure of *YoLeCorE* allow researchers to focus on the input and output of individual students in order to compare the written and/or oral output of learners who have received the same input over a specific period of time. This would allow them to focus on variables other than the input which may affect L2 development in instructional settings.

Finally, given the limited number of spoken learner corpora currently available, *YoLeCorE'*s database of oral output is a valuable source of data for the study of EFL learners' language development over a school year. As the corpus includes the input and output of a class of 17 young EFL learners in a Greek state school, we cannot claim that this is a representative database of EFL young learners' input and output, in general. However, the content and structure of *YoLeCorE* allows researchers to trace the pace of students' language growth at this age and make connections between specific types of language input and the quantity and quality of students' language output. As already mentioned, the particular school places emphasis on the development of students' oral communication skills and on the use of L2 in class - by both students and the teacher. This means that the corpus provides opportunities for research into the (pushed) output produced by students in their effort to communicate their messages as well as they can. The fact that this is a primary school EFL class with young L2 learners adds further value to *YoLeCorE,* as access to similar formal instructional settings for long periods of time is extremely rare due to learners' young age and the difficulty in obtaining the necessary permissions to access them. *YoLeCorE* is expected to be made available soon through the Department of English Language and Literature, Aristotle University of Thessaloniki, Greece.

While YoLeCorE may obviously be useful to a number of researchers, we believe that the present corpus, as well as others of similar type, may also have practical implications for classroom teachers. Given the importance of quantity of input (Flege, 2009), classroom teachers may have robust data regarding the amount of language input as well as knowledge of the particular units that are either easy or challenging for learners to learn. By knowing the quality of input (Jia and Aaronson, 2003), teachers can adjust the number of repetitions to the level of difficulty of the words or reduce the repetition of words learners have acquired in favor of others. This will in turn save valuable teaching time and safeguard against unnecessary repetition which might lower students' motivation (Nitta and Baba, 2014). What is more, given the importance of formulaic units in language learning (Ellis, 2006) and their high frequency in spoken and written production in class (Erman and Warren, 2000), teachers can examine whether the language used in class includes a number of such units and perhaps modify it accordingly. Finally, the digital form of both input and output allows for their comparison against the CEFR. Indeed, teachers may compare the syllabus against CEFR wordlists and determine the precise level of CEFR to which learners are exposed. In the same light, teachers can also examine learners' output and evaluate whether they meet the CEFR descriptors for each level of performance.

Dr. Thomas Zapounidis is currently the principal of the 3rd Experimental Primary School of Evosmos in Thessaloniki, Greece and a post-doc researcher. He can be reached at thomasza@enl.auth.gr.

Marina Mattheoudakis is a Professor in Applied Linguistics, School of English, Aristotle University of Thessaloniki, Greece. She is the director of the Foreign Language Teaching Lab at the Aristotle University of Thessaloniki and the director of Bilingualism Matters in Greece. She can be reached at marmat@enl.auth.gr.

References

Aijmer, K. (ed.) (2009). *Corpora and language teaching*. Amsterdam/ Philadelphia: John Benjamins.

Anthony, L. (2014*). AntConc (Version 3.4.3)* [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Bestgen, Y. & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28– 41. https://doi.org/ 10.1016/j.jslw.2014.09.004

Campoy-Cubillo, M. C., Bellés-Fortuño, B. & Gea-Valor, L. (2010). *Corpus-based approaches to English language teaching*. London: Continuum.

Coniam, D. (1997). A practical introduction to corpora in a teacher training language awareness programme. *Language Awareness, 6,* 199–207.

Crossley, S. A., Kyle, K., & Salsbury, T. (2016), A usage-based investigation of L2 lexical acquisition: The role of input and output. *The Modern Language Journal 100*(3), 702–715. doi:10.1111/modl.12344

Ellis N.C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics, 27*, 1–24.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, 20(1), 29-62.

Flege, J. E. (2009). Give input a chance! In

T. Piske & M. Young-Scholten (Eds.), Input matters in SLA (pp. 175–190). Clevedon, UK: Multilingual Matters

Gabrielatos, C. (2002). Grammar, grammars and intuitions in ELT: a second opinion. *IATEFL Issues, 170*, 2-3

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 359-370). London: Routledge.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Jia, G., & Aaronson, D. (2003). A longitudinal study of Chinese children and adolescents learning English in the United States. *Applied Psycholinguistics, 24*, 131–161.

Laufer, B. & Waldman (2011). T. Verb-noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning 61*, 647-672.

Lee, D. & Swales, J. (2006) A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*(1): 56–75.

Leńko-Szymańska, A. (2014). Is this enough? A qualitative evaluation of the effectiveness of a teacher-training course on the use of corpora in language education. *ReCALL, 26*(2), 260-278.

Mayring, P. (2014). *Qualitative content analysis: theoretical foundation, basic procedures and software solution.* Retrieved from http://www.ssoar.info/ssoar/handle/document/39517

McCarthy, M., McCarten, J. & Sandiford, H. (2005). *Touchstone: Student's Book 1.* Cambridge: Cambridge University Press.

McEnery, T. & Hardie, A. (2012.) *Corpus Linguistics: Method, Theory and Practice.* Cambridge: Cambridge University Press.

Meunier, F. (2011). Corpus linguistics and second/foreign language learning: exploring multiple paths. *RBLA, Belo Horizonte, 11*(2), 459-477.

Nitta, R. & Baba, K. (2014). Task repetition and L2 writing development. In R. Manchón & H. Byrnes, (Eds.), *Task-Based Language Learning: Insights from and for L2 Writing*(pp. 107–136). Amsterdam: John Benjamins.

O'Sullivan, I. (2007). Enhancing a process-oriented approach to literacy and language learning: the role of corpus consultation literacy. *ReCALL 19*(3), 269-86.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics, 32,* 130-149. https://doi.org/10.1017/S0267190512000098

Scheyder, E. C. (2012). *The impact of recordings on student achievement in critical language courses.* Dissertations available from ProQuest. AAI3509479. https://repository.upenn.edu/dissertations/AAI3509479

Tagarelli, K. M., Ruiz, S., Vega, J.L.M., Rebuschat, P. (2016). Variability in second language learning: The roles of individual differences, learning conditions, and linguistic complexity. *Studies in Second Language Acquisition*, *38*(2), 293-316.

VanPatten B. (2002). Processing instruction: An update. *Language Learning*, 52, 755–803.

Zapounidis, T. (2017). *Young Learners' L2*

Input and Output in the 3rd Experimental Primary School of Evosmos: The Young Learner Corpus of English (YoLeCorE).[Unpublished doctoral thesis]. Aristotle University of Thessaloniki.