RiCL Research in Corpus Linguistics

The compilation of a developmental spoken English corpus of Turkish EFL learners

Ece Genç-Yöntem – Evrim Eveyik-Aydın
Yeditepe University / Turkey

Abstract – Although compiling a spoken learner corpus is not a recent enterprise, the number of developmental learner spoken corpora in the field of corpus linguistics is not satisfactory. This report describes the compilation of the *Yeditepe Spoken Corpus of Learner English* (YESCOLE), a 119,787-word corpus of Turkish students' spoken English at tertiary level. YESCOLE was compiled to generate a developmental corpus of spoken interlanguage by collecting samples from learners of different English proficiency levels at regular short intervals over seven months. In order to shed light on the laborious methodology of compiling the developmental spoken learner corpus, this paper elucidates the steps taken to build YESCOLE and discusses its potential benefits for research and instructional purposes.

Keywords – learner corpus; spoken corpus; corpus compilation; developmental corpus; EFL

1. INTRODUCTION

Data collection timing has been an important criterion in learner corpus research. In learner corpora studies, data can be collected either at one point in time or repeatedly over time depending on the purpose of the research study. The former are called synchronic corpora and the latter are diachronic corpora (Gilquin 2015: 13). Most of the learner corpora are synchronic and have a cross-sectional design (e.g. the *International Corpus of Crosslinguistic Interlanguage*, ICCI; Tono and Díez-Bedmar 2014). Since it is difficult to follow the same learner or group of learners over time, compiling diachronic corpora (e.g. longitudinal and developmental corpora) is quite challenging for many researchers. An example to a longitudinal learner corpus is the *Longitudinal Database of Learner English* (LONGDALE) in which learners were tracked over three years by collecting data once a year (Meunier 2016). Some learner corpora are called developmental when the data are collected more densely. In such corpora, "learner performance is documented at close intervals or at all points of production" (Belz and Vyatkina 2008: 33). In the study

Research in Corpus Linguistics 10/1: 45–62 (2022). Published online 2021. ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.10.01.03

by Belz and Vyatkina (2008), the research corpus, *Telecollaborative Learner Corpus of English and German* (Telekorp), was defined as developmental on the grounds that learners were followed over a two-month period. In this vein, both longitudinal and developmental corpora are rich sources that display progress of learners (Gilquin 2015: 13).

The present paper reports on the compilation steps of the *Yeditepe Spoken Corpus of Learner English* (YESCOLE), which is also a developmental corpus that has the potential to fill a significant gap in the field for a number of reasons. First of all, according to the information obtained from the *Centre for English Corpus Linguistics* (CECL) at Université catholique de Louvain,[1] it should be indicated that, except for some corpora containing samples of learners' spoken production (e.g. the *Louvain International Database of Spoken English Interlanguage*, LINDSEI),[2] the existing learner corpora in the field are in the written mode. Secondly, while existing corpora include language samples from learners with different first language (L1) backgrounds, the number of spoken corpora that involve language produced by L1 Turkish learners of English is found to be rather limited, as shown in Table 1.

Corpus Name	L1	Mode	Size in words
The Turkish component of the LINDSEI (LINDSEI-TR)	Turkish	Spoken	80,813
Corpus of Learner Monologues (CLM)	Turkish	Spoken	6,151
Turkish Corpus of Spoken Learner English (TC-SLE)	Turkish	Spoken	1,500 (in progress)

Table 1: List of spoken learner corpora in which Turkish is L1 and English is the target

Among the few existing corpora, the Turkish component of the LINDSEI (LINDSEI-TR) was compiled by Kilimci (2014), and it includes almost 50 advanced level English learners' interviews each lasting about 15 minutes. The tasks within the interviews range from telling a story to answering a question and describing a picture. Another spoken learner corpus of Turkish EFL learners, the *Corpus of Learner Monologues* (CLM), was compiled by Demirel and Şahin (2015) to identify lexical problems in spoken English, and it comprises 35 participants' two-minute talks on selected *International English Language Testing System* (IELTS)[3] speaking topics. The proficiency levels of the English learners contributing to CLM range from intermediate to upper-intermediate. The last corpus, the *Turkish Corpus of Spoken Learner English* (TC-SLE), was compiled by

[1] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
[2] https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html
[3] https://www.ielts.org/

Demirel and Kazazoğlu (2015), and it consists of two-minute talks of Turkish learners of English whose levels of proficiency range from intermediate to advanced on two selected IELTS speaking topics. TC-SLE is still in progress and intends to include learners' monologues and language use in classroom contexts and group work activities.

Despite their merits, the above-mentioned spoken corpora, compiled in a Turkish EFL context, do not comprise longitudinal data from the same learner or learners over time. As stated previously, one of the main reasons for the scarcity of longitudinal corpora is that it is rather challenging for researchers to track the same learner or group of learners over time. Moreover, many researchers face data attrition issues in a way that the same participant stops contributing to the corpus over time, which discourage them to conduct studies that will take time. Therefore, not only does a low number of scholars find opportunities to collect learner data longitudinally, but also such learner corpora in the field are mostly in the written mode. However, developmental and longitudinal learner corpora can shed light on the changes in interlanguage and difficulties that are experienced during the course of language learning. This was also pointed out by Thewissen (2013), who analyzed the error-tagged quasi-longitudinal corpus of 223 essays obtained from the *International Corpus of Learner English* (Granger *et al.* 2009) and demonstrated the accuracy development of B1, B2, C1 and C2 learners of English by examining the corpus-driven learner errors. This implies that there exists an ever-increasing need for developmental and longitudinal learner spoken corpora in foreign language contexts.

In view of this, we have compiled YESCOLE, a specialized spoken corpus of L1 Turkish learners of English as part of a doctoral study to investigate the spoken interlanguage of English-major students who attend a language preparatory program at Yeditepe University, a foundation university in Istanbul, Turkey. This paper covers a) in Section 2, the steps taken to compile YESCOLE, and b) in Section 3, a discussion of the research potential that such a learner spoken corpus offers. Hence, not only does the paper present a guiding methodology for those who pursue studies in a similar vein, but it also highlights how such corpora enable studies on learners' interlanguage.

2. THE COMPILATION OF YESCOLE

2.1. Representativeness and descriptive features

A learner corpus should be large and representative enough to address the target research questions (Granger 2004: 125). The design criteria (e.g. population, proficiency level, tasks, mode and timing) should be selected carefully before the corpus compilation. As highlighted by Rea Rizzo (2010: 3), the corpus can also be specialized when data are collected from specific groups (e.g. a spoken learner corpus) or genres (e.g. a corpus of English as a Lingua Franca in academic settings) and when it is aligned with corpus builders' own research purposes (e.g. studying the changes in language use of Spanish learners of English). Specialized corpora have lately become a preferred way to answer specific research questions, especially in EFL contexts.

Considering this, as displayed in Table 2, YESCOLE is a specialized corpus that includes spoken performance of young adult Turkish EFL learners at Yeditepe University, Turkey. It was compiled to generate a developmental corpus of spoken interlanguage by collecting spoken samples from learners at three different levels of English proficiency (A2, B1 and B2 according to the *Common European Framework of Reference for Languages*, CEFR)[4] to investigate grammatical and lexical errors over time in learners of English. YESCOLE represents the academic genre because the data collection setting is a university preparatory school, and data were elicited from the learners as part of their oral exams. The oral exams, which include tasks that require learners to make speeches to discuss the causes/effects or advantages/disadvantages of something, are used quite often in that prep school context. Such oral exams are typical examples of classroom genres that represent spoken academic discourse.

Type	Specialized corpus / learner corpus/ developmental corpus
Mode	Spoken
Population	Young adult English-major Turkish EFL learners at a foundation university preparatory school in Turkey
English proficiency level	A2 (pre-intermediate), B1(intermediate) and B2 (upper-intermediate)
Genre	Academic/ Oral exam/ (monologue/ non-interactional)

Table 2: Features of YESCOLE

YESCOLE comprises four sub-corpora: 1) YESCOLE-A2 (corpus of pre-intermediate level Turkish EFL learners), 2) YESCOLE-B1 (corpus of intermediate level Turkish EFL learners), 3) YESCOLE-B2 (corpus of upper-intermediate level Turkish EFL learners),

[4] See CEFR manual (Council of Europe 2005) for detailed information regarding each level of language proficiency.

and 4) YESCOLE-LONG (developmental corpus of A2, B1 and B2 level Turkish EFL learners). The corpus is not tagged by part-of-speech (POS); however, it is error-tagged (see 2.3.6.). The details related to the spoken corpus such as total tokens, types and utterance counts were computed using *AntConc* 3.5.8 (Anthony 2019). These are shown in Table 3.

	YESCOLE-A2	YESCOLE-B1	YESCOLE-B2	YESCOLE-LONG	YESCOLE-Total
Total word	13,806	50,571	19,088	36,322	119,787
Total type	1,462	3,066	1,901	2,053	3,922

Table 3: Size of YESCOLE and its sub-corpora

In total, YESCOLE comprises 119,787 words. YESCOLE-B1 is the largest in size (50,571 words), and it is followed by YESCOLE-LONG (36,322 words) and YESCOLE-B2 (19,088 words).

2.2. Participants

The spoken data used to generate the specialized spoken corpus of learner English were collected through convenient sampling from 105 Turkish young adult EFL learners who study in the English language preparatory program specifically intended for the students of Translation Studies (TRA), English Language and Literature (ELIT) and English Language Teaching (ELT) at Yeditepe University. Participation was voluntary; therefore, a consent form requesting students' permission to allow their instructors to record their speech during exams was prepared. 105 learners out of 112 granted permission and filled in a learner profile questionnaire, which provided demographic information related to age, gender, language background (e.g. their L1 and when they started learning English), and whether they had any health problems (e.g. hearing or speaking impairment or learning disability). The demographic data revealed that the learners' average age was 19. Out of 105, 80 participants were female and 25 were male. The average age at which they started learning English as a foreign language was nine. They were all L1 speakers of Turkish and none of the students had health problems.

2.3. Steps taken to compile YESCOLE

2.3.1. Checking the proficiency levels of the participants

Since the initial purpose of collecting learner English spoken data was to examine the learners' progress over time, the participants were placed into three groups on the basis of the *Oxford Quick Placement Test* (OQPT) (2001), which can be used to provide information about the proficiency level of learners. The OQPT also offers a chart of equivalent levels to be interpreted with respect to the levels of the CEFR (2005). Table 4 summarizes the number, gender, and proficiency level of learners in the corpus.

Gender	A2 level	B1 level	B2 level	Total
Female	19	44	17	80
Male	3	12	10	25
Total	22	56	27	105

Table 4: Number of participants according to gender and proficiency levels

According to the classification offered by OQPT, participants who received a score between 18 and 29 out of 60 were placed into A2; those who received a score between 30 and 39 were placed into B1; and those whose score ranged from 40 to 47 were placed into B2. These results indicated that 27 participants were B2 level (upper-intermediate), 56 were B1 level (intermediate), and 22 were A2 level (pre-intermediate).

2.3.2. Selection of the prompts to elicit oral data

Different techniques to elicit L2 oral data have been reported in the literature. Some of these ways include learner monologues on a given topic (e.g. Demirel and Şahin 2015; Yıldız 2016), tasks of oral argumentation (e.g. Kormos and Dörnyei 2004), picture description (e.g. Yuan and Ellis 2003; Ellis and Yuan 2004), role-plays (e.g. Ting *et al*. 2010; de Jong *et al*. 2013), oral interviews (e.g. Boers *et al*. 2006; Huang 2011), story-telling (e.g. Khan 2011) and course presentations (e.g. Aşik and Cephe 2013). Although different tasks require different cognitive demands (Skehan 1998), asking for oral argumentation was acknowledged to be an effective way to elicit L2 learners' speech in the literature (e.g. Masrom *et al.* 2015). Therefore, speaking prompts that elicit students' opinions on different topics were used to compile YESCOLE because a) the students' performance is assessed with these questions in the program, and b) the oral elicitation

technique should be similar to oral argumentation so as to benefit from the advantages that this task brings with it (e.g. a substantial amount of spoken performance data).

Thus, the corpus consists of Turkish EFL learners' speaking test performance, which is based on their monologic talks on given speaking prompts during their oral exams. In accordance with the requirements of the prompt, the participants talk about causes or effects of a particular topic, offer solutions or suggestions for a particular problem, talk about advantages or disadvantages of a particular topic, and provide reasons for their opinion. These prompts elicit students' opinions on various topics discussed in their courses. Some prompts elicited from their course materials can be exemplified as:

(1a) Attendance at university should not be obligatory. To what extent do you agree or disagree with this idea? State your reasons.

(1b) Obesity has become the main concern for many young people. What are the effects of this situation?

2.3.3. Preparing the setting for oral recordings

After taking students' voluntary consent, the next step was to collect audio-recordings of spoken data. The English monologues on different topics of the 105 participants were audio-recorded in three speaking exams, including one quiz and two achievement exams during the 16-week semester. At the end of the semester, 29 participants from the initial 105 were enrolled in the summer school, so their spoken data were also recorded in three speaking exams during the 12-week summer semester. As can be seen in Figure 1, spoken data were collected at short, regular intervals from the same students to see their progress in line with the aim of the study. This longitudinal design was necessary to reach a rich source of data.

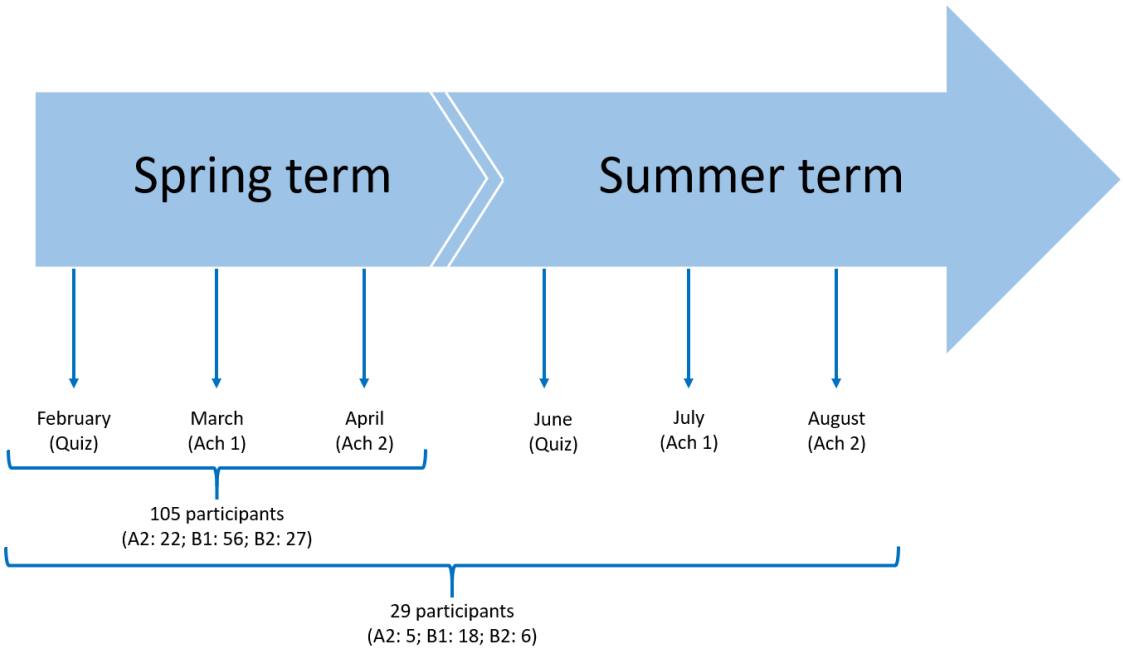

Figure 1: Timescale of recording spoken data

Speaking exams given to each participant were administered in private sessions in an exam room by two instructors and recorded with a professional quality audio-recorder. The students, who blind picked one of the speaking prompts on separate pieces of papers, were given one-minute planning time to think about their response. The importance of pre-task planning time was pointed out by Ahangari and Abdi (2011) in terms of its benefit to increase the quantity and complexity of the oral performance. Likewise, Skehan (1998) and Ortega (1999) highlighted the positive effects of planning time on L2 tasks, which balance the cognitive load and reduce speaking anxiety.

2.3.4. Transcription of audio-recordings

Approximately 20 hours of recording was done in the spring semester and five hours of recording was gathered during the summer school. After the recordings, the 402 sound files were checked for their quality and saved in the researcher's computer. Each participant was given a number, from one to 105, both to facilitate data storage and to keep data anonymous for ethical considerations. The audio files were also grouped according to the participants' proficiency levels and time of data collection. In order to code the files, the time of the data collection was indicated with a combination of the letter T and a number. In other words, T1, T2 and T3 were used to refer to the first, second and third data collection periods during the spring semester, respectively. T4, T5 and T6,

however, were used to indicate the fourth, fifth and sixth data collections in summer. Hence, to illustrate, the first recording of participant 9 at B2 proficiency level was coded as 9B2T1.

As the amount of spoken data to be transcribed was large, the workload of transcription was shared with a doctoral student in English Language Education. To transcribe the spoken data efficiently and consistently, a set of rules was followed. For example, as the focus in the corpus was on learner errors, spoken data were transcribed without correcting any error, false starts, repetitions and reformulations. To ensure consistency throughout spoken data transcription, the chat transcription conventions commonly used in the field (MacWhinney 2000) were adapted. To illustrate, fillers (e.g. *uh* and *uhm*) and unintelligible speech were marked with specific codes: [&-um] was used for fillers and [xxx] was used to transcribe unintelligible speech. Pauses were also indicated with [.] or [..] (if the duration is more than 7 seconds). The audio files were transcribed in *Microsoft Word* files. Then, these files were converted into plain text format in *Notepad* using *AntFileConverter* 1.2.1 (Anthony 2017) so that *AntConc* 3.5.8 (Anthony 2019) could be used for corpus description and analysis.

2.3.5. Identifying the utterances as units of analysis

In spoken learner language studies, one of the first decisions to be made is whether to count sentences or utterances as units of spoken language. In fact, researchers prefer the term ‘utterance’ or ‘C-unit’ (conversational unit) to refer to the basic unit of spoken language. Within the spoken corpus construction and analysis process, it is important to pinpoint utterances appropriately and consistently because analyses are conducted on the basis of utterances (Yaman *et al.* 2008). However, identifying utterances is very difficult in spoken production. There have been some techniques used to determine them, including intonation (Traum and Heeman 1997), probabilistic language models (Stolcke *et al.* 1998) and speech intervals as input units (Worm 1998).

In this study, the participants’ spoken production is monologic and academic, so it includes unscripted monologues. However, it also reflects the features of spoken language such as disfluencies, repetitions, retraces, and incomplete sentences. Before corpus-based linguistic data analysis, utterances were identified as they are the basic units of analysis

in this study (see Table 5 for utterance numbers). Both intonation and speech intervals were used as techniques to detect utterances in YESCOLE.

	YESCOLE-A2	YESCOLE-B1	YESCOLE-B2	YESCOLE-LONG	YESCOLE-Total
Utterance count	1,509	3,550	1,420	2,652	9,131

Table 5: Utterance numbers in YESCOLE

2.3.6. Corpus annotation and analyses

After the long transcription process, labels or tags were added to the corpus so that it can be automatically analyzed using a corpus analysis tool. This step is called 'corpus annotation' or 'coding'. In this process, some information such as tags or labels is inserted into the original transcriptions. There are some tools that annotate the native corpus data automatically (e.g. SPPAS, Bigi 2015); however, due to the nature of learner language, some problems (e.g. inconsistent annotation) might occur in learner corpus annotation. Researchers can develop their own tags and coding schemes depending on their objectives. Tags are generally hand-coded on the transcripts via corpus analysis tools such as *AntConc* (Anthony 2019) and *Computerized Language Analysis* (CLAN) (MacWhinney 2000). Yet, reliability tests should be conducted in order to make sure that the data have been consistently tagged.

In accordance with the purpose of our research, YESCOLE was error-tagged by adding the tags to show grammatical and lexical errors in the data, and this is called 'error annotation'. To describe the error types appropriately, Dulay *et al.*'s (1982) surface strategy taxonomy (omission, addition, misformation [misselection] and misordering of linguistic elements) was followed. To use a standard format for error tagging and to make it consistent, an annotation scheme was developed, and each error type was marked with a different tag. The tags for errors in YESCOLE were created by adapting those used in the CLAN manual (MacWhinney 2000). To illustrate, an error was identified with an asterisk in square brackets in the text after the error. For example, the omission of plural *-s* error in English was coded as: [*ms:a:0s]. In this code, **ms* stands for morpho-syntactic error, *a* indicates that it is an agreement error, and *0s* stands for omission of plural *-s*. In another example from the same linguistic level, addition of plural *-s* was coded as [*ms:a:+s]. In this code, + stands for addition of plural *-s*. Examples of an error-tagged utterance are given in Table 6.

Linguistic level	Error type	Example	Error Tag
Morpho-syntax	Omission of comparative *-er*	*He is short [*ms:0er] <shorter> than his sister.*	[*ms:0er]
Morpho-syntax	Addition of plural *-s*	*He gave me [*ms:a:+s] <advice>.*	[*ms:a:+s]

Table 6: Examples of an error-tagged utterance

Moreover, *AntConc* (Anthony 2019) was used to annotate corpus data by inserting tags or labels to facilitate word search and corpus analysis. These tags can be hidden or shown in the search results.

3. Potential benefits of YESCOLE for research and EFL instruction

YESCOLE includes spoken samples of A2, B1 and B2 level (according to the CEFR) Turkish learners of English at tertiary level, and it will be expanded by collecting spoken data from A1 and C1 level learners at regular time intervals. After the spoken samples have been collected to describe the continuum of language proficiency, the corpus will be made available for researchers. The potential of YESCOLE for research and instruction is a good example for those willing to compile such developmental corpora, which are truly lacking in the field. According to McEnery and Gabrielatos (2006: 49), corpora have assisted language-related inquiry in four main aspects: 1) depiction of language and creation of reference materials; 2) lexicogrammatical analysis of language; 3) EFL instruction; and 4) noticing changes in a language. In addition, regarding spoken learner corpus building, Du Bois (1991: 73) states that

> a transcription of spoken discourse can provide a broad array of information about these and other aspects of language, with powerful implications for grammar, semantics, pragmatics, cognition, social interaction, culture, and other domains that meet at the crossroads of discourse.

Considering this, compiling a specialized corpus of learners' developmental spoken English will offer many benefits and areas of application for research and EFL instruction.

Since YESCOLE is a developmental corpus, spoken data collected at different times reflect potential changes in learner English. As learner language has been claimed to have its own rules and developmental patterns (Selinker 1972), learner corpus research has been of great help to describe and track the progress across different language proficiency levels, observe the difficulties learners face in the process of learning and call for action (Thewissen 2013). One of the best ways to identify the problems which learners

face is to analyze the language used by them. By understanding the specific issues that a certain group of students from the same L1 language background face, researchers can more precisely understand and compare the different stages of that group's acquisition of English through the compilation of learner corpora in their own contexts. In spite of the difficulty in collecting spoken data repeatedly from a learner or a group of learners over time, developmental/longitudinal studies using corpus-approach to track EFL spoken performance should contribute more to the field.

De Cock (2010) highlighted that there is a need for studies using spoken learner corpora in the classroom. This holds especially true in the Turkish EFL context where there is a great need for diachronic or developmental/longitudinal spoken learner corpora to gain a better understanding of learner English. Such learner corpora can also be used in the classroom in order to support EFL instruction. As EFL learners find it challenging to speak accurately in English, learner corpora in spoken mode might show not only the weaknesses and the strengths of English language learners but also the progress of reducing language errors. Even specific language structures, such as the use of passive voice or adjectival clauses, can be investigated and specific treatments can be developed to improve EFL instruction.

Moreover, spoken learner corpora such as YESCOLE can be used in data-driven learning (DDL; Johns 1991), syllabus and material design, and language testing. On the one hand, teachers and students can directly use such corpora via computer software to search for examples of learner language use. Although many teachers may not be willing to integrate corpus directly in their lessons due to time limitations, lack of corpus knowledge and technical constraints such as the absence of computer facilities (Farr 2008; Hedayati and Marandi 2014; Ebrahimi and Faghih 2017), corpus-driven data can become quite beneficial to work on. On the other hand, corpus-driven data can be used indirectly to prepare ELT materials, course syllabi, language tests and exercises. In contrast to the direct use of corpus, corpus consultation may be more favorable especially in technology-poor contexts (Hedayati and Marandi 2014). Concordance lines extracted through a corpus tool can be used in activity preparation. For example, error-tagged corpora such as YESCOLE can be used to create error-correction activities in the classroom or the structures that students have difficulty with can be taught and then tested in the exams. The use of concordance lines obtained via the corpus analysis software, as illustrated in Figure 2, can bring students' attention to common word formation errors (coded as

[*m:affix] in YESCOLE). The concordance lines can be presented either as screenshots or used in an error-correction exercise.

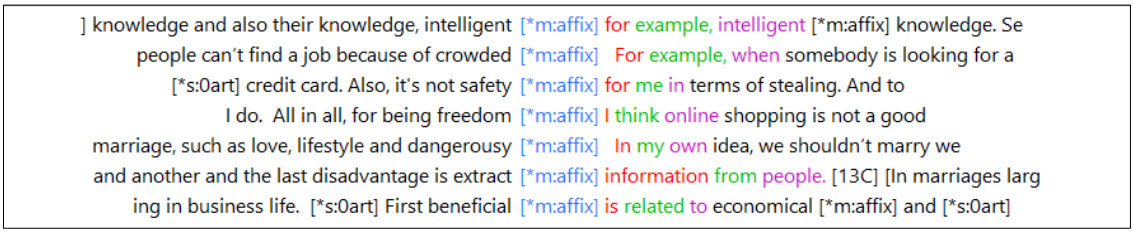

Figure 2: Screenshot of sample concordance lines showing some word formation errors in *AntConc*

Accordingly, teachers can give corpus-driven feedback on learners' speaking errors and output, so students will notice and be aware of the most common errors committed. Below is a sample exercise prepared with concordance lines that include word formation errors from YESCOLE.

Error-correction Exercise

The following concordance lines have been taken from YESCOLE, a spoken corpus of Turkish learners of English. Read the lines and correct the words with an asterisk (*).

1. … because of their tastes and cheap*, also their preparation to an. As a ...
2. … and, staying away from crowdness* and they can live small ...
3. … people's thoughts and their speaks*. And also she can do home chores...
4. … may feel themselves more energic*. And they can enjoy their life their ...
5. … is related to economical* and second is ...
6. … in big cities, but it's crowd* and most of the companies…
7. … trust issue is one of the nature* aspect of unhappy marriage. For example…
8. … compared to villages. Like sport* centers, parks, stuff like that. And people…
9. … First of all, it has a deterrent* effect on the society. Many people in…
10. … beautiful thing to do before die*. Everybody should keep a pet, they should...

Another example may be the use of learners' erroneous spoken performance found in the corpus to prepare awareness-raising activities (e.g. Hobbs 2005). For this purpose, transcripts obtained from recordings of students' spoken English can be used as consciousness-raising activities. Below is a sample speaking task prepared with transcript 5B1T1 obtained from YESCOLE. Similarly, transcripts with audio files can be used to detect mispronunciation. Although pronunciation errors have not been tagged in YESCOLE, such spoken corpora can be used to raise learners' awareness.

Speaking Task

A. **Imagine that you are asked to decide whether to continue your education online or in traditional classes next semester. In order to decide, you need to list the characteristics of online classes and traditional classes.**

Online Classes	Traditional classes
- online	- face-to-face
- …	- …

B. **A Turkish learner of English talks about whether online classes are better than traditional classes. Read the transcript of the talk taken from YESCOLE and correct all the errors. How many errors did you find? Compare your answers with those of your classmates.**

Hello. Today I'm going to talk about online classes are ... if they are better than conventional classes. I think I... it's not. I don't agree with that idea.

Because, first of all, I cannot make a conversation with my teacher face-to-face. And to explain, when I'm [xxx] in reality not in online, I feel more comfortable and ask whatever I want. Secondly, I cannot ask any questions while I'm not in classes, online classes. Mainly I have some question marks in my head. And I do not ask it to you because we are talking on skype or something other social media applications. What else? Finally, online classes are works with electric like computers and other electronic devices. If electric goes out I cannot keep up with the classes. So maybe I should go.

So online classes are very useless, in my opinion. I cannot make any face-to-face conversation. I cannot ask any question while I'm not in class, online class, and if electric goes out I cannot keep up with my classes. And I can miss some classes.

C. **Prepare a similar speech on the same topic. Do you think online classes are better than traditional classes? Justify your answer.**

De Moraes (2018) also points out that a spoken learner corpus can be used to teach speaking through creating instructional activities tailored to the needs of a specific learner group. Furthermore, as suggested by Gilquin *et al.* (2007), a specialized spoken corpus can fill the gaps in English for Academic Purposes (EAP) pedagogy to create teaching and testing materials. To illustrate, wordlists can be made through the corpus tool and students' vocabulary profile might be observed; review classes might be organized considering the resistant errors; and quality distractors can be selected from the error-tagged corpus while preparing language tests. In this vein, building and making use of learner corpora provide opportunities for teachers, curriculum/course designers, and test developers to prepare teaching and testing materials using the spoken language of learners.

Lastly, not only the features of spoken English, such as discourse markers, ellipsis, headers and tails but also communication strategies (e.g. compensation speaking strategies) of learners at different levels of proficiency can be investigated with the help

of learner spoken corpora. As can be seen, developmental spoken learner corpora, such as YESCOLE, can contribute to the field of ELT in many respects.

4. CONCLUSION

This paper introduced YESCOLE as a representative, specialized and developmental learner corpus of spoken English. It is a novel source of learner data in the Turkish EFL context which, to the best of our knowledge, is currently the only developmental/longitudinal spoken corpus of English-major EFL learners at different levels of proficiency in this context. In that sense, we believe that YESCOLE includes invaluable spoken data to inspire studies that might reveal significant instructional implications contributing to the field of ELT. The paper describes in detail the steps taken into account to build YESCOLE as summarized below:

- Checking the proficiency level of the learners of English,
- selecting the prompts to elicit oral data,
- planning the data collection timings and preparing the setting for oral recordings,
- transcribing the oral recordings,
- identifying the utterances as units of analysis,
- annotating the corpus, and
- conducting automatic corpus analyses.

These steps can be guiding and inspirational for future researchers who aim to compile a developmental/longitudinal spoken learner corpus as well. We truly hope that the spoken corpus compilation becomes a more common practice in years to come to be able to provide a corpus-based evidence to the language development of EFL learners all over the world.

REFERENCES

Ahangari, Saeideh and Morteza Abdi. 2011. The effect of pre-task planning on the accuracy and complexity of Iranian EFL learners' oral performance. *Procedia – Social and Behavioral Sciences* 29: 1950–1959.

Anthony, Laurence. 2017. *AntFileConverter* (version 1.2.1). Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/software

Anthony, Lawrence. 2019. *AntConc* (version 3.5.8). Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/software

Asik, Asuman and Pasa Tevfik Cephe. 2013. Discourse markers and spoken English: Nonnative use in the Turkish EFL setting. *English Language Teaching* 6/12: 144–155.

Belz, Julie A. and Nina Vyatkina. 2008. The pedagogical mediation of a developmental learner corpus for classroom-based language instruction. *Language Learning & Technology* 12/3: 33–52.

Boers, Frank, June Eyckmans, Jenny Kappel, Helene Stengers and Murielle Demecheleer. 2006. Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research* 10/3: 245–261.

Bigi, Brigitte. 2015. SPPAS – Multi-lingual approaches to the automatic annotation of speech. *The Phonetician – International Society of Phonetic Sciences* 111: 54–69.

Council of Europe. 2005. *Reference Supplement to the Preliminary Version of the Manual for Relating Examinations to the Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. DGIV/EDU/LANG 2005, 13. Strasbourg: Language Policy Division.

De Cock, Sylvie. 2010. Spoken learner corpora and EFL teaching. In Mari Carmen Campoy-Cubillo, Begona Bellés-Fortuño and M. Luisa Gea-Valor eds. *Corpus-based Approaches to English Language Teaching.* London: Continuum, 123–137.

De Jong, Nivja H., Margarita P. Steinel, Arjen Florijn, Rob Schoonen and Jan H. Hulstijn. 2013. Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics* 34/5: 893–916.

De Moraes, Helmara Febeliana Real. 2018. Use of corpora in teaching speaking. In John I. Liontas ed. *The TESOL Encyclopedia of English Language Teaching*. New York: John Wiley and Sons, 1–6.

Demirel, Elif Tokdemir and Koray Şahin. 2015. The use of spoken learner corpora to detect problems with lexical accuracy. *HUMANITAS-Uluslararası Sosyal Bilimler Dergisi* 3/5: 73–83.

Demirel, Elif Tokdemir and Semin Kazazoğlu 2015. The comparison of collocation use by Turkish and Asian learners of English: The case of TCSE corpus and ICNALE corpus. *Procedia – Social and Behavioral Sciences* 174: 2278–2284.

Du Bois, John W. 1991. Transcription design principles for spoken discourse research. *Pragmatics* 1/1: 71–106.

Dulay, Heidi C., Marina K. Burt and Stephen D. Krashen. 1982. *Language Two*. Oxford: Oxford University Press.

Ebrahimi, Alice and Esmail Faghih. 2017. Integrating corpus linguistics into online language teacher education programs. *ReCALL: The Journal of EUROCALL* 29/1: 120–135.

Ellis, Rod and Fangyuan Yuan. 2004. The effects of planning on fluency, complexity, and accuracy in second language narrative writing. *Studies in Second Language Acquisition* 26/1: 59–84.

Farr, Fiona. 2008. Evaluating the use of corpus-based instruction in a language teacher education context: Perspectives from the users. *Language Awareness* 17/1: 25–43.

Gilquin, Gaëtanelle. 2015. From design to collection of learner corpora. In Sylviane Granger, Gaëtanelle Gilquin and Fanny Meunier eds. *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 9–34.

Gilquin, Gaëtanelle, Sylviane Granger and Magali Paquot. 2007. Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes* 6/4: 319–335.

Granger, Sylviane. 2004. Computer learner corpus research: Current status and future prospects. In Ulla Connor and Thomas A. Upton eds. *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi 123–145.

Granger, Sylviane, Estelle Dagneaux, Fanny Meunier and Magali Paquot eds. 2009. *International Corpus of Learner English*. Louvain-la-Neuve: Presses universitaires de Louvain

Hedayati, Hora Fatemeh and S. Susan Marandi. 2014. Iranian EFL teachers' perceptions of the difficulties of implementing CALL. *ReCALL* 26/3: 298–314.

Hobbs, James. 2005. Interactive lexical phrases in pair interview tasks. In Corony Edwards and Jane Willis eds. *Teachers Exploring Tasks in English Language Teaching*. London: Palgrave Macmillan, 143–156.

Huang Lan Fen. 2011. *Discourse Markers in Spoken English: A Corpus Study of Native Speakers and Chinese Non-native Speakers.* Birmingham: University of Birmingham dissertation.

Johns, Tim. 1991. From printout to handout: Grammar and vocabulary teaching in the context of data-driven learning. In Tim Johns and Philip King eds. *Classroom Concordancing*. *English Language Research Journal*, 4, Birmingham: University of Birmingham. 1–16.

Khan, Sarah. 2011. *Strategies and Spoken Production of Three Oral Communication Tasks: A Study of High and Low Proficiency EFL Learners*. Barcelona: Universitat Autònoma de Barcelona dissertation.

Kilimci, Abdurrahman. 2014. LINDSEI-TR: A new spoken corpus of advanced learners of English. *International Journal of Social Sciences and Education* 4/2: 401–410.

Kormos, Judit, and Zoltán Dörnyei. 2004. The interaction of linguistic and motivational variables in second language task performance. *Zeitschrift für Interkulturellen Fremdsprachenunterricht* 9/2. https://ojs.tujournals.ulb.tu-darmstadt.de/index.php/zif/article/view/482/458 (20 May, 2021.)

MacWhinney, Brian. 2000. *The CHILDES Project: Tools for Analyzing Talk* (third edition). Mahwah, NJ: Lawrence Erlbaum Associates.

Masrom, Umi Kalsom, Nik Aloesnita Nik Mohd Alwi and Nor Shidrah Mat Daud. 2015. The role of task complexity and task motivation in language production. *GEMA Online Journal of Language Studies* 15/2: 33–49.

McEnery, Tony and Costas Gabrielatos. 2006. English corpus linguistics. In Bas Aarts, April MS McMahon and Lars Hinrichs eds. *The Handbook of English Linguistics*. Oxford: Blackwell, 33–71.

Meunier, Fanny. 2016. Introduction to the LONGDALE Project. In Erik Castello, Katherine Ackerley and Francesca Coccetta eds. *Studies in Learner Corpus Linguistics. Research and Applications for Foreign Language Teaching and Assessment*. Berlin: Peter Lang, 123–126.

Ortega, Lourdes. 1999. Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition* 21/1: 109–148.

Oxford Quick Placement Test (Version 1). 2001. Oxford University in collaboration with University of Cambridge, Local examinations Syndicate, Oxford: Oxford University Press.

Rea Rizzo, Camino. 2010. Getting on with corpus compilation: From theory to practice. *ESP World* 9:1–23.

Selinker, Larry. 1972. Interlanguage. *IRAL-International Review of Applied Linguistics in Language Teaching* 10: 209–232.

Skehan, Peter. 1998. *A Cognitive Approach to Language Learning*. Oxford: Oxford University Press.

Stolcke, Andreas, Elizabeth Shriberg, Rebecca Bates, Mari Ostendorf, Dilek Hakkani, Madelaine Plauche, Gökhan Tur and Yu Lu. 1998. Automatic detection of sentence boundaries and disfluencies based on recognized words. In the *Fifth International Conference on Spoken Language Processing*. Sydney, Australia (November 30-December 4, 1998). https://www.isca-speech.org/archive/archive_papers/icslp_1998/i98_0059.pdf (20 February, 2021.)

Thewissen, Jennifer. 2013. Capturing L2 accuracy developmental patterns: Insights from an error-tagged EFL learner corpus. *The Modern Language Journal* 97/1: 77–101.

Ting, Su-Hie, Mahanita Mahadhir and Siew-Lee Chang. 2010. Grammatical errors in spoken English of university students in oral communication course. *GEMA Online Journal of Language Studies* 10/1: 53–70.

Tono, Yukio and María Belén Díez-Bedmar. 2014. Focus on learner writing at the beginning and intermediate stages: The ICCI corpus. *International Journal of Corpus Linguistics* 19/2: 163–177.

Traum, David R. and Peter A. Heeman. 1997. Utterance units in spoken dialogue. In Elisabeth Maier, Marion Mast and Susann LuperFoy eds. *Dialogue Processing in Spoken Language Systems*. Heidelberg: Springer, 125–140.

Worm, Karsten L.1998. A model for robust processing of spontaneous speech by integrating viable fragments. In Association for Computational Linguistics eds. *COLING 1998 Volume 2: The 17th International Conference on Computational Linguistics*, 1403–1407. https://www.aclweb.org/anthology/P98-2229/ (12 November, 2020.)

Yaman, Sibel, Li Deng, Dong Yu, Ye-Yi Wang and Alex Acero. 2008. An integrative and discriminative technique for spoken utterance classification. *IEEE Transactions on Audio, Speech, and Language Processing* 16/6: 1207–1214.

Yıldız, Mustafa. 2016. Contrastive analysis of Turkish and English in Turkish EFL learners' spoken discourse. *International Journal of English Studies* 16/1: 57–74.

Yuan, Fangyuan and Rod Ellis. 2003. The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 monologic oral production. *Applied linguistics* 24/1: 1–27.

Corresponding author
Ece Genç-Yöntem
Yeditepe University
Kayışdağı BLVD, 326A
Ataşehir, PO Box 34755
İstanbul, Turkey
e-mail: ece.genc@yeditepe.edu.tr

received: November 2020
accepted: May 2021
published online: 2021