The Learner Corpora of Spoken English: What Has Been Done and What Should Be Done?

Soyeon Yoon†
Incheon National University

ABSTRACT
The number of spoken learner corpora is smaller than that of written corpora; however, demand for spoken corpora has continuously increased. This study investigates the current state of spoken English-language learner corpora both in the world and in Korea, with a focus on their size, speakers, and genres. Based on this survey, the study discusses factors to consider when building and publishing a spoken learner corpus, especially with respect to the issues of conversation genre, proficiency level, and annotation.

Keywords: spoken corpus, learner corpus, conversation, transcription, annotation

1. Introduction

Ever since Brown Corpus was constructed in 1967, various types and sizes of corpora have been constructed and widely used as quantitative and qualitative evidence not only for language research but also for foreign/second language education. In language education, the corpora of native language have served as reference that the learners' language patterns can be compared to or as a model that the learners pursue. On the other hand, learner corpora collect second or foreign language data that the learners use in either written or spoken forms. They have been constructed to examine learners' linguistic knowledge and how their usage patterns differ from those of native speakers or differ depending on the learners' age, gender, first language, and fluency.

According to Center for English Corpus Linguistics[1] (CECL) at Universite Catholique de Louvain, there are 177 learner corpora in the world currently available. Only 40 of them contain spoken data while most are written corpora. Since

1) https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

* This research was supported by the Incheon National University Research Grant in 2018. I would like to thank three anonymous reviewers of this journal for their helpful and insightful comments.
† Corresponding author: syyoon@inu.ac.kr

spoken corpora[2] require more time, effort, and financial cost for recording and transcription, most of the learner corpora so far have focused on written data. Nevertheless, in order to obtain a more complete picture of learners' language, we need to build more spoken corpora that encompass various genres and speakers.

As speaking skills are considered significant not only in high-stakes situations, such as interviews, oral examinations, and presentations, but also in casual interactions, more attention has been paid to speaking ability than before. However, due to the lack of systematic and large-sized spoken learner corpora, the number of corpus study using learners' spoken data still remains small (Friginal et al., 2017, p. 3), and some researchers even make their own small-sized corpora to analyze learners' speech (see Chapter 3).

Recognizing the need for systematic spoken learner corpora, this study surveys learner corpora of spoken English that have been constructed so far, focusing on the ones publicly available. The current study provides those who plan to employ a spoken learner corpus in their analyses with the information of the currently available spoken learner corpora. Also, by proposing what to consider when constructing the spoken data, it contributes to those who plan to build a large-sized spoken learner corpus or their own corpus. Chapter 2 introduces some corpora that were constructed outside Korea and world-widely used. Chapter 3 deals with the learner corpora of English that were constructed in Korea and shows how researchers in Korea made their own spoken learner corpora for their studies. In Chapter 4, I discuss how we can improve learner corpora so they can better represent the learners' linguistic knowledge and use. Chapter 5 concludes the study. In the Appendix is the list of corpora introduced in this study along with their websites.

2. Learner Corpora of Spoken English in the World

2.1. Overview

As of February 2020, CECL (Feb 05, 2020) at Universite Catholique de Louvain lists 177 learner corpora in the world. The list may not be complete (although aiming for a complete list), at least it shows current status of learner corpora in general.

2) *Spoken corpora* in this study refer to not only the corpora providing both speech recording format and their transcribed text format, but also the ones providing either format.

The list includes the corpora of 24 target languages, among which are Arabic, Chinese, English, French, Finnish, German, Italian, Korean, Persian, Portuguese, Spanish, etc. More than half of the learner corpora (105 out of 177, 59.3%) are targeting English, and among the English corpora, only 35 (33.3%) are either pure spoken corpora or the ones that contain both spoken and written data.

The size of the spoken corpora varied from 35,000 to 2,000,000 words. The tasks that were given to the speakers for the spoken data collection varied. Some recorded sentence reading (e.g., Corpus of Writing, Pronunciation, Reading, and Listening by Learners of English as a Foreign Language, Kotani et al., 2016), national spoken English test (the College Learners' Spoken English Corpus, Yang and Wei, 2005), role plays (the Barcelona English Language Corpus, Muñoz, 2006), story description (the Santiago University Learner of English Corpus), and interviews (The Louvain International Database of Spoken English Interlanguage, Gilquin et al., 2010). Some corpora provide transcription of a phonetic level (the LeaP Corpus : Learning Prosody in a Foreign Language, Slavianova, 2007), while some corpora focus on word level transcription (ICNALE: the International Corpus Network of Asian Learners of English, Ishikawa, 2013). The first language of the learners also varied, but according to the list, only three contained Korean speakers' data. However, only ICNALE are publicly available, and the others are either unavailable (The Neungyule Interlanguage Corpus of Korean Learners of English) or under development (The EFL Teacher Corpus[3]).

The CECL list suggests that the spoken corpora are still rarer compared to the written corpora. Moreover, the speech was recorded in a constrained and controlled situation, such as interviews, tests, and presentations. It is true that these tasks are the activity that the learners are exposed to in EFL and ESL institutional settings, but the corpora may not represent the learners' spoken English usage in real situation that they may encounter outside the classrooms.

In the next section, I will look more closely how some of the learner corpora of spoken English in the world have constructed.

2.2. Selected learner corpora of spoken English in the world

Louvain International Database of Spoken English Interlanguage (LINDSEI) (Gilquin et al., 2010) is an 800,000-word corpus of learner interviews, which was

3) The corpus is marked as 'under development' in the CECL list, but its brief description can be found in Kwon & Lee (2014).

launched in 1995. There were 554 non-native speakers (NNS) of English of 11 different first languages, including Chinese, Dutch, French, German, Japanese, but not Korean (about 50 speakers per L1). With a native speaker (NS) of English interviewer, each NNS participant had an interview, which may be applicable to the real world interview context. During the interview, the speakers selected one of the three topics and expressed their thought about it, followed by an interview of question-answer regarding the speakers' answer to the topic. Then the speakers described a strip of picture. In general, the running time of each recording was 15 minutes. The speakers' demographic information, such as age, gender, L1, and English learning experience, was collected. This corpus became the source for various linguistic phenomena from grammar, discourse markers, and word collocation. For example, Aijmer (2011) used the Swedish component of LINDSEI to study the discourse marker *well* and compared it with the Louvain Corpus of Native English Conversation (LOCNEC), which was compiled within the same framework of LINDSEI (De Cock, 2007).[4] De Cock (2004) used French component of LINDSEI to study collocation and compared it with LOCNEC.

We can see that the structure of LINDSEI is consistent through the interviews: similar running time, the same number of learners of each L1, and the same tasks and topics. Thus, the structure of LINDSEI serves as a model when researchers build their own corpus for the purpose of their own study. For example, to examine the learners' use of discourse marker *so*, Buysee (2012) followed the model of LINDSEI and LOCNEC. He built a corpus of spoken English produced by 40 Belgian native speakers of Dutch and 20 NS of English.

Overall, LINDSEI is a large sized corpus of consistently structured data with various L1 learners. Therefore, the corpus can be reliably used for comparison not only between different L1 speakers but also with native speakers' data, LOCNEC, which has the same structure.

While LINDSEI collected spoken data from various L1 learners, the International Corpus Network of Asian Learners of English (ICNALE) collected both written and spoken data of the learners in Asia, including China, Hong Kong, Indonesia, Japan, Korea, Pakistan, the Philippines, Singapore, Malaysia, Taiwan, and Thailand. It also carries the data collected from NSs of English. As of August 2019, 1,450,000 words of written essays (including edited essays), and 2,100,000 words of spoken data were collected. The spoken data were collected in two forms: One is a one-minute

4) Although LOCNEC contains conversation in its name, note that it is a collection of informal interviews between native speakers of English, not spontaneous conversation (De Cock, 2007, p.219).

monologue about ICNALE common topics, and the other is a dialogue composed of 30-40-minute oral interviews that include picture descriptions and role plays, along with 3-4-minute follow-up L1 reflections. The speakers talked about the ICNALE common topics for both the monologue and dialogue. The number of the monologue recordings was 4,400, and 425 for the dialogues.

The corpus aimed to be available for contrastive analysis among the learners' English from different language backgrounds. Therefore, it is strictly and well constrained in terms of the topics (two topics: smoking in the restaurant and college students' part-time job), the interview structure, and running time. In addition, the corpus provides the speakers' proficiency level in English. The speakers were required to take a standard L2 vocabulary size test (Nation & Beglar, 2007) and to present English proficiency test such as TOEIC and TOEFL if available. Then, the scores were carefully matched with Common European Framework of Reference for Languages (CEFR). With this corpus, we can compare the learners' English in different genres (monologues vs. dialogues) and in different proficiency levels. Moreover, the speakers' individual information includes learning motivations, experience in English as well as basic information such as gender, age, academic majors, and year of studying English. With various information about the speakers, we can research on how learners' different background affects English learning in relation to their actual utterances.

However, even though the tasks of LINDSEI and ICNALE are meant to collect spontaneous speech, the corpora are still not free from the criticism that these spoken data do not represent utterances of a naturally occurring conversation which the learners may be exposed to in real world. Actually, the same criticism is also applicable to most of the learner corpora listed in CECL. In most spoken learner corpora, the interlocutors had not known each other until the day of the interviews, and the tasks were far from a natural setting, such as talking about picture strips, or talking alone about one topic, which may make the speakers nervous and uncomfortable to respond naturally. Also, the interview contents were limited to only some given topics, and thus, the vocabulary and expressions are relatively limited. Nevertheless, it can be claimed that collecting data through these tasks may be practical, especially in an EFL context like South Korea where majority of the speakers are NSs of Korean: The learners' speech in a naturally occurring event may be difficult to obtain because the speakers are hardly exposed to the occasion where they have to use English.

Nonetheless, we can still observe naturally produced speech of NNS of English

in some corpora. For example, the Michigan Corpus of Academic Spoken English (MICASE, Simpson et al., 2002) collected spoken data naturally occurring in various authentic academic setting, such as advising, dissertation defenses, interviews, labs, lectures, study groups, student presentation, etc. There are 152 transcriptions, totaling 1,848,364 words. The corpus has been used for examining academic English in reality, but the corpus did not target to collect learners' English. Therefore, only 12% of the corpus was NNS speech (Friginal et al., 2017).

More NNS speech can be observed in the corpora that collected the spoken English used as a Lingua Franca, i.e., English as a language of communication among speakers of different L1s. Vienna-Oxford International Corpus of English (VOICE, 2013), which started recording from 2001, is one of the examples. The corpus collected spoken data from naturally occurring authentic face-to-face interaction (interviews, press conferences, working group discussions). Over a million words (120 hours of transcribed speech) produced by 1250 experienced speakers from 49 L1 backgrounds were transcribed.

The Hong Kong Corpus of Spoken English (HKCSE) (Cheng et al., 2005) collected speech produced by NSs and NNSs in naturally occurring speech events, i.e., academic discourse, business, public discourse, and conversation, each of which comprises a sub-corpus. Each consists of 50 hours of naturally occurring talk (2million words in total), and 53% of the corpus is prosodically transcribed. Among the sub-corpora, the Hong Kong English Corpus of Conversational English (HKCCE) collected 130 conversations in English (average 23 minutes per conversation, 50 hours and about 500,000 words in total). Note that conversation in this corpus is different from interviews, and is defined as follows: 'A speech event outside of an institutionalized setting involving at least two participants who share responsibility for the progress and outcome of an impromptu and unmarked verbal encounter consisting of more than a ritualized exchange' (Cheng & Warren, 1999, p. 8). Among 341 participants, only half of the were NNS (NNS = 48%, NS = 52%), but the corpus is useful to study Hong Kong English accent, because it contains prosodic transcription. The strong point of the corpus is that the data were recorded in a naturally occurring situation where the participants are engaged in real world. Therefore, we can observe how the NNSs communicate in English (including prosodic patterns) with NSs or other foreign interlocutors in real life.

Both VOICE and HKCSE made it possible to examine NNS speech in natural setting. However, the speakers were mostly advanced speakers whose English is proficient enough to competently communicate with NSs. Therefore, the lower level

learners' natural speech still lacks.

In summary, more than half of learner corpora targeted English corpora, and one third of the English learner corpora contain spoken data. Most of the spoken data were collected from speaking tasks such as reading, monologues (opinion about a given topic, English proficiency test, and picture description), or dialogues (mostly, interviews with NSs about a given topic). NNS speech in a natural setting can be found in the corpora of English as Lingua Franca setting, but they lack the data from learners of lower level proficiency. This suggests that the naturally occurring spoken data of lower level learners may still be added to the spoken learner corpora.

3. Learner Corpora of Spoken English in Korea

3.1. Learner corpora of English in Korea

There are only limited number of learner corpora of English in Korea that are publicly available now. Unfortunately, none of these contain spoken data. Yonsei English Learner Corpus (YELC, Rhee & Jung, 2012) is a written corpus that collected writing compositions produced by prospective students of Yonsei University at Yonsei English Placement Test. 3,286 people took the test in 2011, in which the students wrote about casual topics in 100 words (e.g., What was your favorite extracurricular activity in high school? What made you join the activity?) and about academic topics in 300 words (e.g., Why should people receive a college education? State your opinion). In total, 1,085,828 words from 6,572 texts were collected. The data are available at the archive of English Informatics Laboratory upon request.

The Gachon Learner Corpus (Carlstrom & Price, 2012-2014) is another written corpus publicly available. Its final version was collected from 2012 through 2014 at Gachon University. It consists of over 2.5 million words from 25,073 individual texts produced by 2500 participants. The participants were asked to write about one topic out of 20, using 100-150 words.

It has been reported in CECL that Neugyule Interlanguage Corpus of Korean Learners of English (NICKLE) contained student essays (890,000 words) and transcriptions (100,000 words) of student interviews and oral speech tests. However, the project was suspended, and the corpus is not publicly available.[5)]

We can access to YELC and Gachon corpus relatively easily and study the written

English of Korean learners. Unfortunately, there is no learner corpora of spoken English focusing on Korean learners and publicly available.

3.2. Learner corpora of spoken English in Korea

Many scholars have constructed their own learner corpora of spoken English for their own studies although they are not open to public. In this section, I present some of these corpora focusing on the tasks that the speakers are given.

First, face-to-face interview is the most realistic solution to obtain relatively natural spoken English. For example, M. Kim (2009) compared the errors and collocations in written and spoken English. The study revealed that learners' written English showed higher lexical density and more diverse word types, but the error types and the errors in collocations were not very different from the spoken English. For this study, M. Kim collected from Korean L1 speakers 134 writing (30,555 words in total) and 134 transcripts (33,140 words). The participants wrote an essay of 250-350 words by selecting one topic out of three and had an interview with an NS for 10 minutes. The corpus has been used by some other researchers. For example, by using the same corpus, Back (2011) analyzed errors in preposition. Shin et al. (2018) also employed this corpus and compared frequency and lexical variety of single words and multi-word units with NS corpora BNC Spoken Sampler and BNC Written, each of which contains one million words.

Second, R.-E. Kim & Rhee (2019) investigated Korean learners' pronunciation of individual sounds, specifically, /l/ and /r/ distinction. For this study, they used a part of Genie SpeeCor. This corpus aims to collect the learners' English pronunciations, so the speakers (elementary and middle school students and adults) repeated sentences or short paragraphs after they looked at or listened to them. The learners were classified into 5 levels according to the pronunciation test. The corpus is distinguished from most learner corpora in that it specializes in phonetic and phonological characteristics of the learners.

Other special learner corpora are those of teacher talk in EFL class rooms. Kwon & E.-J. Lee (2014) compiled 62 hours of recordings from EFL classes in a university in Seoul. They collected 123,122 words from four NNS EFL teachers and 124,275

5) In addition to these corpora, Han & Lee (2009) developed a model for preposition error detection based on Chungdahm English Learner Corpus. According to this study, as of November 2008, the corpus contains 130,754,000 words from 861,481essays produced by 11-16-year-old students. A part of the corpus is error-annotated. However, the corpus is not open to public.

words from five NS EFL teachers. Also, J. Lee (2019) analyzed the use of DM *so* in teacher talk. Six teachers who were highly proficient English speakers participated in the project. Each taught an EFL class of 30-35 students in English in a secondary public school for 45 minutes. The lessons were video-recorded and transcribed by a low degree of elaboration to access to the linguistic features that she was interested in, which is *so*. These corpora are different from typical learner corpora introduced in Section 2 and 3 because the teachers were highly proficient speakers of English, and their talk may have been quasi-planned according to the lesson plans. However, they were still NNSs of English, and thus, we can observe NNSs' speech of a specific genre in an institutionalized setting, which is teacher talk.

Y. Lee (2012) collected learners' speech through the task called diapix. In this task the two speakers (one Korean and one NS of English or Chinese) sit in the sound booth with a shoulder-high barrier in-between. Each has the identical picture with only a few exceptions. They have to find out the different parts through a 8-10-minute conversation. Lee claims that the spoken data obtained through diapix are spontaneous, and two speakers are in equal status, which means that the balance of utterances between the speakers can be maintained (p. 9), unlike interviews in which one speaker's utterance dominates the conversation. The study found out that NNSs used more *like* and *kind of* as a discourse marker, which was different from the result of Fuller (2003) and Müller (2005) that NNSs used discourse markers less frequently than NSs. The author claims that the learners used these markers as an approximation to achieve the goal of finding different pictures in a limited time. For this reason, I suspect that the excessive use of the approximation discourse markers may not be the learners' linguistic characteristics, but may be caused by the diapix method itself. This suggests the possible limitation of the task.

There is a corpus that collected authentic and natural conversation where English was used as a lingua franca (Chung et al., 2016). The corpus contained 150 conversations collected at the Chinese University of Hong Kong, each of which lasted about 5 minutes (about 750 minutes in total). For each conversation, four speakers from different L1 backgrounds (Cantonese, Mandarin, Korean, Japanese, Tagalog and others, 100 speakers in total) had a conversation in English on the given topics. The corpus was used to examine English segmental pronunciation produced by Koreans and other Asians (Chung et al., 2016), and to examine the discourse marker *well* (Yun & Kim, J.-R., 2018). This corpus is distinguished from the corpora discussed so far, in that it aimed to collect English used as a tool to

communicate with people of different L1 backgrounds, which is a plausible situation where EFL learners may encounter in real life. Moreover, the speech genre was conversation in which the contribution of each speaker is expected to be equal, unlike interviews or monologues. However, the corpus is not publicly available, and is questionable in terms of representativeness: The number of Koreans was only 20, all of whom are quite proficient enough to take the college courses in English.

Although not available to public yet, Incheon National University Multi-language Learner Corpus (INU-MULC, Yoon et al., 2020) is being developed so it can encompass college students' written short essays, monologues, and conversations. The corpus aims to collect English, Chinese, Japanese, and French data produced by Korean college students and Korean data produced by foreign students in Korea. As its first step, the English part is under development. In each conversation, there are two to four speakers, who are friends or at least had talked before. Each conversation may or may not include a foreigner (non-native speaker of Korean), and lasted for 25 minutes without any given specific topic. As of February 2020, 318 speakers participated in 106 recordings with 310,000 words in total. Also, each participant recorded a monologue (39,000 words) and wrote an essay of 300 words (31,000 words in total) about a topic that he or she chose from a set of prompts. The corpus will be open to public so researchers can use it to compare the language of different genres (essays vs. monologues vs. conversations), to conduct discourse analysis of learner language, and to examine grammar and vocabulary acquisition according to different proficiency levels, and so on. The corpus is expected to be available from 2021.

In general, it seems that the learner corpora of spoken English constructed in Korea vary in their speech genres (repeating sentences, interviews, teacher talk, and conversations) and the type of the tasks (diapix and talking about a given topic). However, none of these are open to public and the process of data collection still has to be described in more detail: Some corpora do not report total number of transcribed words, how the recordings were transcribed, how the speakers' proficiency level was judged, etc. In the next section, I will discuss some factors to consider when constructing a spoken corpus.

4. Challenges of Learner Corpora of Spoken English

Although not focusing on learner corpora, S.-S. Kim (2018) discussed some

challenges in corpus linguistics in general. He reviewed what factors have to be considered when we construct a corpus and what researchers have to be careful of when using the corpus data, from six perspectives, which are representativeness, spoken corpus, annotation, software tool, analysis, and language teaching. Among these, the former three are relevant to corpus design while the latter three are relevant to using and applying the corpora. In this chapter, I will specifically discuss former three; the necessity of spoken corpora, especially conversation corpora, representativeness specifically in relation to proficiency levels, and annotation.

4.1. Spoken corpora of conversation

It is well known that constructing a spoken corpus is costly (Love et al., 2017; S.-S. Kim, 2018), and thus, the number of spoken learner corpora is much smaller than that of written corpora. In addition, the genres are limited to interviews and monologues in general. These speech genres reflect EFL context where English is spoken mostly in institutional settings: speaking tests in the form of an interview with an English teacher, speaking tests in the form of a monologue, role plays with a colleague in a classroom, etc.

However, demand for conversation corpora have been recognized in L1 research area. Santa Barbara corpus of spoken American English (SBCSAE, Du Bois et al., 2000-2005) contains 249,000 words of casual conversation in natural settings. Its transcription is highly elaborated, even with the pause length and the number of pulses in laughter. British National Corpus 2014 (BNC2014, Love et al., 2017) focused on collecting naturally occurring conversations. The corpus contains 11.5 million words of orthographically transcribed conversations among L1 speakers of British English from across the UK. The appearance of L1 corpora of conversation implies that the spoken data collected from speeches and TV shows (as in the case of the Corpus of Contemporary American English) are insufficient to fully observe native speakers' spoken language. Also, learner corpora try to elicit natural utterances from the speakers in authentic settings, which led some corpora, like VOICE and HKCSE, to collect the utterances in real life.

Conversation is different from other institutionalized speech. The number of participants in one conversation may vary, not just limited to one or two like monologues and interviews. Also, casual conversation is characterized of 'equal distribution of speaker rights' (Cheng & Warren, 1999, p. 8), which is different from interviews where at least one of the speakers controls the rights of speaking.

Therefore, the interaction in a conversation is more dynamic than a one-on-one interview. Moreover, we can observe the strategies that the participants use to manage turn takings (e.g., use of discourse markers, response tokens, and pauses), and how the speakers use their grammar and vocabulary to deal with this spontaneous and dynamic interaction.

As we have seen in Section 3, there are no learner corpora of spoken English publicly available in Korea. Moreover, the ones that some researchers build for their own research recorded interviews or monologue type speeches. However, these types of speech are not everything that the learners use in English. As conversation is a 'pre-eminent form of language' (Swales, 1990, p.43), English learners in Korea are exposed to a situation where they are involved in 'conversation' in English more frequently than ever. For example, college students have to discuss class assignments or chat with foreigners as the number of foreign students on campus increases. Also, more employees have a conversation with their foreign clients, not just in a conference room, but in a restaurant or in a tour site. As the occasions where Korean speakers use English for conversation increases even in an EFL context, we need a learner corpus of conversation to study how the learners interact in the conversation.

Even after recording spoken data of conversation, transcribing the recordings makes it difficult to complete the data as a corpus. The utterances in a conversation are irregular and unplanned, and include non-verbal features (S.-S. Kim, 2018, p.51). Therefore, its consistent and thorough transcription requires a lot of time and cost because each transcriber's judgments (Andersen, 2010) on spellings of contractions (e.g., *'cause, coz, cuz*), response tokens (e.g., *hmm, ah*), pauses, intonations, and dealing with visual or situational information (Creer & Thompson, 2004) may vary.

To achieve consistency in transcription and to save cost in time and effort, BNC 2014 adopted "standard orthographic" transcription (Leech et al. 2001, as cited in Love et al., 2017, p. 333), with linguistic details such as false start and hesitation normalized or disregarded. However, such linguistic details may show important characteristics of learners, which may never be discovered in written corpora. Therefore, learner corpora introduced in Section 2 still transcribe these linguistic details, but limit the selections of spellings that the transcribers can choose from. It is also necessary to cross-check the transcription, as BNC 2014 did.

INU-MULC deals with conversation transcription in the following way. It basically adopts standard orthographic transcription like BNC 2014, but to present conversational details it employs Discourse Transcription (DT, Du Bois, 2006) by

simplifying it. For example, DT is designed to describe elaborate details such as marking the number of pulses in laughter with the corresponding number of symbol '@,' or marking the length of the pause in number. However, this fine-grained level of delicacy delays and complicates the transcription. Instead, INU-MULC records the laughter with <LAUGH> regardless of the number of pulses and the pause less than 1 second is marked as <.>. Also, the spellings for the response tokens were restricted. The overlapping parts are surrounded by square brackets. The example is presented in (1). Note that all speakers' names are pseudonyms.

(1) An example of INU-MULC transcription
MIKE; Dormitory is so far from here.
JUDY; Yes.
DAN; Ah? It's run [away].
BEN; [<LAUGH>]
JUDY; [<LAUGH>]
DAN; It's two minutes' walk.
MIKE; <.> We couldn't go there earlier early time.
DAN; <.> Really?

Once the recordings were transcribed as in (1), they underwent three checking processes: In the first checking stage, a pair of transcribers checked the partner's transcriptions, in the second stage, editors checked the transcriptions again, and in the final stage, native speakers checked them focusing on the consistency. This multi-step checking may minimize errors and inconsistency, but it requires a lot of time, effort, and financial expense.

Despite the cost of collecting the data and transcribing the recordings, learner corpora of spoken English, especially conversation, are essential to understand learners' language that has not been much explored.

4.2. Representativeness and proficiency levels

It is hard to judge how large a corpus must be in order for it to represent a certain genre or group of speakers. Depending on the aim, genre, and the details of transcription and annotation, the size of the spoken corpus may vary. For example, HKCCE and SBCSAE, which are conversation corpora with fine-grained transcription, contain less than a million words (500,000 and 249,000 words, respectively). VOICE

and ICNALE that recorded genres other than conversation are the size of over a million words.

Representativeness can be discussed in terms of speakers and texts. First, a corpus must represent the speakers' language in the target community. Most learner corpora target a specific group of learners rather than comprehensive population. For example, many corpora focus on adult learners, especially university students, with similar number of gender. Depending on the purpose of the corpus, the sampled learners and their information collected for the corpus vary. Usually, when collecting the recordings, researchers consider learners' demographic information such as age, gender, proficiency level, and experience of English (length of English education, length of stay in English-speaking countries, hours spent on English learning in a week, etc.). Sometimes, the academic fields that the college students major in are considered, especially, if the corpus is interested in academic English (e.g. MICASE, LINDSEI). If the corpus collects the speakers of different L1 backgrounds, the number of each L1 speakers is also considered (e.g., ICNALE, LINDSEI, HKCSE, VOICE). However, in Korea where most learners' L1 background is homogeneous, the corpora may sample only Korean L1 speakers.

Among these speakers' characteristics some are controlled from the sampling procedure so the number of sampled speakers in each category represents the population. For example, the researchers can collect the speakers considering the ratio of gender, L1 background, age group, and academic majors. On the other hand, some characteristics are collected, but not controlled. For example, learners' experience of English is collected but usually they are not controlled from the beginning, and this information is treated as only supplementary. ICNALE is unique in that it even surveyed each learner's learning motivation and exposure to English quite in detail.

In a learner corpus, the speakers' proficiency level in English is important. Instead of collecting the learners of all different levels, some corpus focus only on the advanced learners who can use English as a daily language with no problem (e.g. HKCSE, VOICE, MICASE). However, if the corpus aims to collect the learners of different levels, the proficiency judgment is an important factor. There are well-known English proficiency tests, such as TOEIC, TOEFL, and IELTS. However, not all learners have the test scores, and the scores themselves are not straightforward to interpret and to apply to the estimation of the learners' linguistic ability (Ito et al., 2005). For example, what does TOEIC score 850 mean? What can the learner of this score do with English? Also, some researchers (Andrade,

Table 1. Six levels of CEFR (spoken interaction and production skill)

Proficient User	C2	Can express him/herself spontaneously, very fluently and precisely, differentiating finer shades of meaning even in more complex situations.
	C1	Can use language flexibly and effectively for social, academic and professional purposes.
Independent User	B2	Can interact with a degree of fluency and spontaneity that makes regular interaction with native speakers quite possible without strain for either party.
	B1	Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans.
Basic User	A2	Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters.
	A1	Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help.

2014; Runnels, 2016) claim that these scores do not necessarily correlate with their ability or progress in language skills. Instead, Common European Framework of Reference (CEFR, Council of Europe, 2001) has been proposed as an alternative measurement to indicate proficiency levels. CEFR describes what the learner can do in three language skills (reception (reading and listening), writing, and spoken interaction/production) across six levels as in Table 1, which presents only spoken interaction skill. (For more information, see the CEFR website.[6]) As shown in Table 1, the description of each level is transparent enough to understand and apply to language education and assessment of the learners.

For the learners' proficiency, ICNALE and YELC provide CEFR level. The participants in ICNALE took a standard L2 vocabulary size test (VST) covering the top 5,000 words (Nation & Beglar, 2007), and presented the high-stake English proficiency tests scores such as TOEFL, TOEIC, IELTS, etc. Then, the scores were related with CEFR. Most students fell in either A2, B1 low, B1 high, or B2 and higher. YELC is a written corpus, but according to the report (Rhee & Jung, 2014), the participants were classified mostly into A2, B1, and B1+. In the case of INU-MULC, the evaluators listened to the recordings of the monologues and judged the level based on the criteria in Table 1. As a result, most speakers fell in A2 and B1 (37% and 40%, respectively).

For the consistent CEFR assessment of the spoken learner corpora, it is the best to have a few evaluators (either NSs or qualified NNSs) listen to the recordings and determine the levels. In order to avoid subjective and inconsistent judgements,

6) https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

it is essential to train the evaluators sufficiently before the actual assessment so that all of them share the same criteria. Also, they must cross-check the others' assessment to achieve consensus. However, most studies in 3.2 assume the proficiency levels based on the proficiency test scores like TOEIC that the participants submit, probably due to the cost of this process. Even with this assumption, the scores must be carefully checked and their levels must be classified based on the guidelines which compare the scores of each tests (See ICNALE website).

4.3. Annotation

Depending on its aim, a corpus may have additional information attached to the text itself, such as demographic information of the speakers, part of speech, pragmatic information, and type of errors and their corrections, especially in the case of learner corpora.

Some L1 spoken corpora contain demographic data of a speaker, such as age, gender, regional dialect, socio-economic class, nationality, birth place, education, etc. (e.g., BNC 2014). A learner corpus may carry additional information such as the speakers' L1 (in learner corpora of multiple L1s), proficiency level (if different), experience / education of the target language, and learning motivation (e.g., ICNALE). Also, the transcribed text can annotate its additional information, like part of speech (POS) of individual words, non-verbal information of the recordings (e.g., overlap, situational information, and transcriber's comments), and speaker attribution. It is possible that the speakers' demographic information can be saved in a separate electronic file, and the text is transcribed with carefully designed transcription conventions. Researchers can use this type of raw corpus.

However, some corpora like BNC 2014 converted this information into XML format so the corpus can be used in the web-based software[7] or is machine-readable. Although it is L1 corpus, BNC 2014 is an example for annotated spoken learner corpora. First, the recordings were transcribed with human-friendly conventions as in (2). Then, the transcriptions were converted into XML format as in (3).

7) BNC 2014 is available publicly via Lancaster University's CQPweb server at https://cqpweb.lancs.ac.uk/

(2) Transcription (pre-XML conversion)[8)]

<0211> I haven't met you

<0216> oh hi

(3) Transcription (post-XML conversion)

<u n="1" who="S0211">I haven't met you</u>

<u n="2" who="S0216">oh hi</u>

In this example (3), <u> indicates the beginning of an utterance, followed by the attribute-value format of line number and speaker ID. After the actual utterance (i.e., *I haven't met you*), </u> indicates the end of the utterance. Now, let us suppose that the speaker S0211 is a male, and this information is also annotated. When a researcher wants to find a certain linguistic feature from male's utterances only, then S0211's utterances will be searched.

If each word is POS tagged and lemmatized, the application of the corpus is much more convenient. For example, Yoon (2020) used a part of English INU-MULC to compared the use of the discourse marker *like* in the conversations where the speakers are all Koreans with the conversations where there is at least one foreigner. In the selected six conversations (21,000 words) from the raw corpus, the 56 verb usages had to be manually sorted out from all 299 instances of *like*, which could have been excluded from the beginning if the corpus were POS tagged. While many large-sized L1 corpora are annotated with at least POS, learner corpora, especially spoken corpora are rarely POS-tagged. Nevertheless, if the size is large enough, the annotations of POS and lemma will be useful.

There are softwares for POS tagging, such as Coh-Metrix, CLAWS, LIWC, Sketch Engine, and Wmatrix.[9)] Stanford POS Tagger is open to public for free although the tag set of the POS is limited to 58 (including punctuations). This tagger employs the Penn Treebank tag set (Marcus et al., 1993). For the L1 written text, the accuracy is claimed to be almost 97%. Even for the learners' spoken utterances, most words are correctly tagged, but still they require manual checking because some response tokens like *uh, ah*, and *mm* are inconsistently tagged, and omission and repetition cause incorrect tagging. (4) shows some lines of an excerpt from the learner corpus INU-MULC. For the purpose of presentation, transcription symbols are omitted. (5) is the text tagged by Stanford POS Tagger.

8) The examples are from the manual of BNC2014 (The British National Corpus 2014, 2018, p.61).
9) A brief introduction of each tagger can be found in Friginal et al. (2017, pp. 21-23).

(4) raw text

a. Ah actually I'm not really from Chungju.

b. Oh then mm. Have you been Central Park? I guess you have.

(5) POS-tagged text

a. Ah_NN actually_RB I_PRP 'm_VBP not_RB really_RB from_IN Chungju_NNP ._.

b. Oh_UH then_RB mm_NN ._. Have_VBP you_PRP been_VBN Central_JJ park_NN ?_. I_PRP guess_VBP you_PRP have_VB ._.

In (5a) and (5b), both *ah* and *oh* are interjections, which are supposed to be tagged with UH. However, in (5a), *ah* is NN (singular noun) and *oh* in (5b) is correctly tagged. Also, *have* in *Have you been Central Park?* is tagged with VBP (verb, non-3rd person singular present) while it is VB (verb, base form) in *I guess you have* in (5b).

The transcription symbols and non-verbal information in the transcription must be cleared before or after the automatic tagging. In addition, the incorrect and inconsistent tagging must be checked manually in order for more accurate tagging. However, once again, this process costs time and expense. Moreover, if one wants to build a learner corpus with pragmatic information (reference, speech act, etc.) and error information (incorrect forms, their correct forms, and error categorization, S.-S. Kim 2018), it will cost even more for accuracy checking.

5. Conclusion

As more studies exploit empirical data from corpora to apply the results to language education, the demand for spoken learner corpora also increases. However, building and publishing a spoken corpus require much cost in all stages: designing a setting that can elicit spontaneous and natural utterances in accordance with the aim of the corpus, recording utterances, evaluating proficiency levels, transcribing recordings, and annotating the transcriptions. Therefore, spoken learner corpora is rarer than written corpora as was seen in Section 2. Although many researchers in Korea make their own learner corpus of spoken English, they don't make it available to public. In order to investigate learners' spoken language with larger sized systematic corpora, we need to develop corpora of quality high enough to publish.

The current study discusses the factors to consider when building a learner corpus

of spoken English, focusing on genres, transcription, proficiency level, and annotation. Besides these factors, however, research ethics involving humans participants is another important issue if the corpus is to be published. The data collection and transcription must be performed under the approval of an institutional review board (IRB), so the participants' voluntary consent to provide their speech data is ensured, and the participants' personal information and their rights are fully protected. When a carefully designed learner corpus of spoken English is constructed and published under proper process, more comprehensive understanding of learners' language will be possible.

References

Aijmer, K. (2011). Well I'm not sure I think… the use of well by non-native speakers. *International Journal of Corpus Linguistics, 16*(2), 231-254.

Andersen, G. (2010). How to use corpus linguistics in sociolinguistics. In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp.547-562). London: Routledge.

Andrade, M. (2014). TOEIC scores: How many points are enough to show progress? *Sophia University Junior College Division Faculty Journal*, 35, 15-23.

Back, J. (2011). Preposition errors in writing and speaking by Korean EFL learners: A corpus-based approach. *Studies in British and American Language and Literature, 99*, 227-247.

Buysee, L. (2012). So as a multifunctional discourse marker in native and learner speech. *Journal of Pragmatics 44*. 1764-1782.

Carlstrom, B. & Price, N. (2012-2014). The Gachon Learner Corpus. Available online at http://koreanlearnercorpusblog.blogspot.kr/p/corpus.html

Centre for English Corpus Linguistics (Feb 05, 2020). Learner corpora around the world. Louvain-la-Neuve: Université catholique de Louvain. Retrieved from https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html

Cheng, W., Greaves, C., & Warren, M. (2005). The creation of a prosodically transcribed intercultural corpus: The Hong Kong Corpus of Spoken English (prosodic), *ICAME Journal 29*, 47-68.

Cheng, W. & Warren, M. (1999). Facilitating a description of intercultural conversations: the Hong Kong Corpus of Conversational English. *ICAME Journal 23*, 5-20.

Chung, H., Kim, Y.-K., & Lee, S.-K. (2016). A study on the features of English as a lingua franca in Asian contexts: Segmental features. *Language and Linguistics, 71*, 237-266. http://dx.doi.org/10.20865/20167111.

Council of Europe. (2001). *The common European framework of reference for languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Creer, S., & Thompson, P. (2004, May). *Processing spoken language data: The BASE experience*. Paper presented at the LREC 2004 International Conference - Workshop on compiling and processing spoken language corpora. Lisbon, Portugal.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgian Journal of English Language and Literatures, 2*, 225-246.

De Cock, S. (2007). Routinized building blocks in native speaker and learner speech: Clausal sequences in the spotlight. In M. C. Campoy & M. J. Luzon (Eds.), *Spoken corpora in applied linguistics* (pp. 217-234). Bern: Peter Lang.

Du Bois, J. W. (2006). Transcription Symbols by Delicacy. Retrieved from http://www.linguistics.ucsb.edu/projects/transcription/representing

Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., Englebretson, R., & Martey, N. (2000-2005). Santa Barbara Corpus of Spoken American English, Parts 1-4. Philadelphia: Linguistic Data Consortium.

Friginal, E., Lee, J. J., Polat, B., & Roberson, A. (2017). Corpora of spoken academic discourse and learner talk: A survey. In E. Friginal, J. J. Lee, B. Polat, & A. Roberson (Eds.), *Exploring spoken English learner language using corpora* (pp. 35-63). Palgrave Macmillan.

Fuller, Janet M., 2003. Discourse marker use across speech contexts: a comparison of native and non-native speaker performance. *Multilingua 22*, 185--208.

Gilquin, G., De Cock, S., & Granger S. (Eds.). (2010). The Louvain international database of spoken English interlanguage, handbook and CD-ROM. Nouvain-la-Neuve: Presses Universitaires de Louvain.

Han, N. R. & Lee, S. H. (2009). Developing a model for English preposition errors using a learner corpus. *Linguistics, 53*. 163-185.

Ito, T., Kawaguchi, K., & Ohta, R. (2005). A Study of the relationship between TOEIC scores and functional job performance: Self-assessment of foreign language proficiency. *TOEIC Research Report*, 1-40.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 1* (pp. 91-118). Kobe, Japan: Kobe University.

Kim, M. (2009). *An error analysis of a learner corpus of written and spoken English produced by Korean university students* (Doctoral dissertation). Korea University, South Korea.

Kim, R.-E. & Rhee, S.-C. (2019). A study on English liquids in the rated L2 English speech corpus of Korean learners. *Korean Journal of English Language and Linguistics, 19*(1), 53-75.

Kim, S.-S. (2018). Challenges and prospects of corpus linguistics. *English Language Teaching, 30*(1), 47-71. http://dx.doi.org/10.17936/pkelt.2018.30.1.3

Kotani, K., Yoshimi, T., Nanjo, H., & Isahara, H. (2016). A corpus of writing, pronunciation, reading, and listening by learners of English as a foreign language. *English Language Teaching, 9*(9), 139-155.

Kwon, Y. E. & Lee, E. J. (2014). Lexical bundles in the Korean EFL teacher talk corpus: A comparison between non-native and native English teachers. *The Journal of Asia TEFL,*

11(3), 73–103.

Love, R. Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics 22*(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Lee, J. (2019). Functional spectrum of a discourse marker so in Korean EFL teacher talk. *Korean Journal of English Language and Linguistics, 19*(3), 371-406.

Lee, Y. (2018). A study on the use of discourse markers by non-native learners of English in spontaneous communication. *Korean Journal of Communication Studies Volume, 20*(4), 5-28.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: Based on the British National Corpus*. Harlow: Pearson Education Limited.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*(2), 313–330.

Müller, S. (2005). *Discourse markers in native and non-native English discourse*. Amsterdam; Philadelphia: John Benjamins.

Muñoz, C. (Ed.). (2006). *Age and the rate of foreign language learning*. Clevedon: Multilingual Matters.

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9-13.

Rhee, S.-C. & Jung, C. K. (2012, March). Yonsei English Learner Corpus (YELC). In *Proceedings of the First Yonsei English Corpus Symposium* (pp. 26-36). Seoul, Korea.

Rhee, S.-C., & Jung, C. K. (2014). Compilation of the Yonsei English Learner Corpus (YELC) 2011 and its use for understanding current usage of English by Korean pre-university students. *Journal of the Korea Contents Association, 14*(11), 1019-1029.

Runnels, J. (2016). Self-assessment accuracy: Correlations between Japanese English learners' self-assessment on the CEFR-Japan's can do statements and scores on the TOEIC. *Taiwan Journal of TESOL, 13*(1), 105-137.

Simpson, R. C., Briggs, S. L., Ovens, J., & Swales, J. M. (2002). The Michigan Corpus of Academic Spoken English. Ann Arbor, MI: The Regents of the University of Michigan.

Shin, D., Chon, Y., Lee, S., & Park, M. (2018). A comparison of single word and multi-word unit profiles in spoken and written corpora of Korean learners and English native speakers. *Journal of the Korea English Education Society, 17*(2), 93-112.

Slavianova, E. (2007). *The LeaP Corpus: Generating a relational database for linguistic query support* (Research Activity). Institute of Computer Science, University of Freiburg.

Swales, J. (1990). *Genre analysis*. Cambridge: Cambridge University Press.

The British National Corpus 2014 (2018). User manual and reference guide (version 1.1) Retrieved from http://corpora.lancs.ac.uk/bnc2014/doc/BNC2014manual.pdf

VOICE. (2013). The Vienna-Oxford International Corpus of English (version 2.0 Online). Director: Barbara Seidlhofer; Researchers: Angelika Breiteneder, Theresa Klimpfinger, Stefan Majewski, Ruth Osimk-Teasdale, Marie-Luise Pitzl, Michael Radeka. Retrived from http://voice.univie.ac.at

Yang, H., & Wei, N. (2005). *Construction and data analysis of a Chinese learner spoken English corpus*. Shanhai Foreign Languse Eduacation Press.

Yoon, S. (2020, October). *Casual Conversations of Same-L1-Group and Foreigner-Including-Group: A Case study of Korean EFL Learner Corpus*. Abstract accepted and paper to be presented at the 5th Annual CLIC Conference. Houston, Texas.

Yoon, S., Park, S., Kim, J.-T., Yoo, H., & Jung, C. K. (2020, June). *Incheon National University Multi-language Learner Corpus (INU-MULC): Its design and application*. Abstract accepted and paper to be presented at Asia Pacific Corpus Linguistics Conference 2020. Seoul, South Korea.

Yun, S. & Kim, J.-R. (2018). A study on the discourse marker well in conversation between Asian speakers of English. *Korean Journal of Teacher Education, 34*(30), 89-106.

Appendix

<Corpora introduced in this study>

Barcelona English Language Corpus (BELC)
https://slabank.talkbank.org/access/English/BELC.html

British National Corpus 2014 (BNC2014)
http://corpora.lancs.ac.uk/bnc2014/

Brown Corpus
https://www1.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html

Corpus of Contemporary American English (COCA)
https://www.english-corpora.org/coca/

Gachon Learner Corpus.
http://koreanlearnercorpusblog.blogspot.kr/p/corpus.html

Hong Kong Corpus of Spoken English (HKCSE)
http://rcpce.engl.polyu.edu.hk/HKCSE/default.htm

International Corpus Network of Asian Learners of English (ICNALE)
http://language.sakura.ne.jp/icnale/

LeaP Corpus : Learning Prosody in a Foreign Language
http://wwwhomes.uni-bielefeld.de/gibbon/Docs/LeapCorpus_Manual.pdf

Louvain International Database of Spoken English Interlanguage (LINDSEI)
https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html

Michigan Corpus of Academic Spoken English (MICASE)
https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple

Santiago University Learner of English Corpus (SULEC)
http://sulec.cesga.es/filologia/index.jsp
Santa Barbara corpus of spoken American English (SBCSAE)
https://www.linguistics.ucsb.edu/research/santa-barbara-corpus
Vienna-Oxford International Corpus of English (VOICE) (version 2.0 Online)
http://voice.univie.ac.at
Yonsei English Learner Corpus (YELC)
https://web.yonsei.ac.kr/yonseicorpuslab/

Soyeon Yoon
Associate Professor
Department of English Language and Literature
Incheon National University
119 Academy-ro, Yeonsu-gu, Incheon 22012, Korea
E-mail: syyoon@inu.ac.kr

Received: February 28, 2020
Revised version received: March 26, 2020
Accepted: March 26, 2020