8
SPOKEN CORPORA

Amanda Huensch and Shelley Staples

1 Introduction/Definitions

In second language acquisition research on L2 speaking,[1] there is growing interest in the use of spoken corpora to understand language development. Corpora (the plural of corpus) are generally defined as large collections of speech (or writing) that are balanced and representative of a particular discourse domain (Biber et al., 1998; McEnery & Hardie, 2012). Most corpus researchers consider corpora to be collections of naturally occurring oral and/or written texts. However, learner corpus researchers often also include in their definitions collections of texts containing elicited material, such as classroom or assessment tasks. We propose a definition of corpora incorporating a cline of spoken data that is more controlled and more naturally occurring, with items like reading of words or sentences at one end of the spectrum; followed by picture description tasks, or narrative recount tasks; followed by speaking performance assessments (e.g., oral interviews or monologues); followed by open-ended classroom tasks (e.g., introducing oneself or describing a trip); followed by conversation and other spoken domains outside of the classroom. As with the surge of written corpora starting in the 1980s, when computing capabilities improved, one of the reasons why spoken corpora are growing in popularity is the increase of digital tools that can make building, analyzing, and sharing spoken corpora easier.

2 Historical Perspectives

Early "corpora" (which were not referred to as such) consisted of individual words, focusing primarily on the study of spoken utterances. These datasets were rightly criticized as unrepresentative of speech as a whole and were particularly attacked for their focus on empiricism versus rationalism by Chomsky (1962). However, interestingly, even after the Chomskian revolution of the 1950s, phoneticians continued to work with naturally observed data, as did second language acquisition researchers (see McEnery & Wilson, 2001). With the advent of machine-readable capabilities, modern corpora were built in the 1950s and 1960s, particularly in English. Most of these comprised writing by L1 speakers, and were not intended for the study of second language acquisition. In fact, of the earliest modern (computerized) corpora, only one made a significant contribution to understanding spoken English, the London-Lund corpus. Started in 1975 and completed in the early 1980s, for

DOI: 10.4324/9781003022497-11

years it was the only corpus of spontaneous spoken English with prosodic annotation. In the 1990s, an explosion of learner and other corpora usable for SLA research took place. Two major databases for accessing these corpora are the Université Catholique de Louvain's CECL[2] list of learner corpora around the world and TalkBank's SLABank.[3] Major spoken corpora used for study of SLA include the ESF[4] family of corpora, LANGSNAP[5] (as well as FLLOC[6] and SPLLOC[7]), LINDSEI,[8] and TLC.[9] Many other corpora exist (see Appendix 8A for spoken corpora and Appendix 8B for selected existing corpora) but few languages are represented (mostly English, French, or Spanish). In addition, there are limitations on the types of analyses that can be conducted. Many corpora (including LINDSEI) do not provide researchers with sound files but are limited to transcripts. For those that do include sound files (e.g., FLLOC and SPLLOC), no phonological annotation is provided, and thus analysis of features such as prosody would be very time consuming.

3 Critical Issues and Topics

The Potential of Learner Corpora for Spoken SLA enquiry

Many scholars have argued for the potential benefits of bringing together the fields of corpus linguistics and SLA. Although not specifically focused on speaking, the general arguments for the benefits of more collaboration in these fields applies here. For example, scholars have argued that to best understand SLA, multiple types of data are essential: corpus data, experimental data, and information about individual differences (MacWhinney, 2017; Meunier & Littre, 2013). Using both experimental and corpus data has the potential to avoid the disadvantages of each (Gilquin & Gries, 2009). For instance, the use of corpora can be advantageous because, if large enough, corpora can provide SLA scholars with rich data sets from many learners to document the paths of second language learning (Granger, 2009; McEnery et al., 2019). As Myles (2015) reminds us, however, for corpora to be useful in this way, they must contain enough examples of the target feature in question to be analyzed, and must include the full array of contexts where that feature would normally occur to avoid misinterpretation of the findings (p. 314).

As development over time is a critical variable in SLA research (Ortega & Iberri-Shea, 2005), a subset of corpora particularly beneficial for SLA enquiry are longitudinal. These types of corpora track the same language learners across multiple data collection points. Nevertheless, longitudinal corpora require collaborative efforts/teamwork as data collection is particularly work and time intensive, and researchers must keep in mind questions of planning, research design, and participant attrition (Tracy-Ventura & Huensch, 2018). It is important to point out that it is not necessary for corpora to be large, longitudinal, general in scope, or naturally occurring to benefit the study of spoken SLA. Many smaller, more specialized corpora exist that maintain balance and representativeness (see Main Research Methods part) for their domain of enquiry. A final advantage of using corpora for SLA enquiry is that they can often be easily shared (McEnery et al., 2019; Myles, 2015) which increases the impact of the data because it allows the possibility for more researchers to use the data, for new questions to be asked and answered with the data, and for replication studies to occur. Researchers have to start with the mindset of sharing from the beginning, however, to ensure ethical use of corpus data.

Research Questions Using Corpora to Investigate Spoken SLA

Despite the fact that written corpora outnumber spoken corpora, there are still many research questions being asked and answered using spoken corpora. These range from the acquisition of grammatical and lexical features (e.g., Crossley et al., 2015), pragmatic features (e.g., Fernández, 2013), utterance fluency (e.g., Huensch & Tracy-Ventura, 2017), phonological features (e.g., Götz, 2013), complexity/accuracy/fluency (CAF) analyses (e.g., Vercellotti, 2017), and more. A 2019 special issue in the *International Journal of Learner Corpus Research* (*IJLCR*) highlights some of the possibilities of using oral corpora to explore spoken SLA using the TLC[9], a large (4.2 million words) corpus collected from 2012–2018 which includes monologic and interactive speech from the Graded Examinations in Spoken English assessment developed by Trinity College London.

One area of spoken SLA research fairly well-represented by the use of corpora is oral fluency development (Huensch, 2020). In the *IJLCR* special issue, Götz (2019) used a subset of the TLC to investigate utterance fluency, specifically the relationship between filled pause frequency and variables such as proficiency level, country of origin, and age of acquisition. Using regression modelling to predict filled pause frequency, Götz demonstrated that the factor with the strongest explanatory power was country of origin, which is a loose proxy for L1 background. With evidence that filled pause usage is particularly linked to L1 influence, Götz calls into question the practice of high-stakes assessment such as the Common European Frame of Reference explicitly mentioning this feature in rubrics designed to test all learners on the same scale. Many other studies have used spoken corpora to explore oral fluency, such as the PAROLE[10] corpus which includes speech from learners of English and French as well as NS control groups and has been used to compare utterance fluency characteristics among NSs and learners at different proficiency levels (Hilton, 2014). The WiSP[11] corpus, including English and Turkish L1 learners of L2 Dutch, has also been used to explore multiple research questions regarding L2 fluency, including investigations of L1–L2 fluency relations (e.g., De Jong et al., 2015).

Another area of research using corpora to investigate SLA pertains to the development of constructions, or form-meaning pairings ranging from morphemes to words to idiomatic expressions to syntactic frames (Ellis et al., 2016). Verb constructions are the focus of Gilquin (2019) and Römer and Garner (2019) in the *IJLCR* special issue. Römer and Garner examined the development of verb argument constructions (e.g., V *about* n, V *for* n) across proficiency levels (low intermediate to high advanced). One benefit of using corpora for such an analysis is the ability to compare results to a large reference corpus, in this case the British National Corpus. Römer and Garner discovered that learners at advanced proficiency levels evidenced similar distributions to the British National Corpus in both the number and distribution of verbs in the constructions and were also able to demonstrate how lower-level learners differed in terms of the types of verbs used in the constructions.

A host of other studies have focused on lexico-grammatical patterns of learner speech across proficiency levels (e.g., Biber et al., 2016; Staples et al., 2017). A fairly consistent finding across research contexts is that task type strongly influences the use of features associated with informational elaboration (e.g., use of nouns and noun modifiers, longer words, passive voice, and relative clauses), more often associated with writing than speech. While mode clearly plays a major role in determining learners' use of these features, tasks requiring more informational content (e.g., an oral interview focused on students' professional experience or an integrated speaking task) lead to greater production of these features. In addition, speakers at higher proficiency levels use more of these features within informationally driven tasks.

The final two studies in the *IJLCR* special issue used the TLC to investigate pragmatic development in the use of backchannels (Castello & Gesuato, 2019) and stance adverbs (Pérez-Paredes & Díez-Bedmar, 2019). Pérez-Paredes & Díez-Bedmar explored the impact of task (monologic vs. dialogic) and proficiency level on the use of adverbs such as *really*, *actually*, and *obviously* to display stance. Using both quantitative and qualitative analyses, the researchers provided evidence that task type differentially impacted adverb usage: *actually* was more task-independent compared to *really*.

Pragmatics of spoken language development has been the subject of several studies of L2 spoken discourse outside of assessment contexts (Fernández & Yuldashev, 2011; Friginal et al., 2017; Gilquin, 2008; Polat, 2011). Friginal et al. (2017) explore how hedges (e.g., *think, sort of*) and boosters (e.g., *so*) along with first person pronouns and modal verbs are used by learners in EAP classroom discourse. Their results show that learners used *think* overwhelmingly as a hedging device, and did not use modals for this purpose as much as their teacher interlocutors. Modal verbs were also used more frequently by L2 learners in collaborative tasks when compared to non-collaborative tasks. Possibility, ability, and permission modals (e.g., *can*, *could*) were particularly frequent, reflecting learners' negotiation of meaning during collaborative tasks (e.g., *can you explain…*).

Current Gaps in the Literature

While the development and use of spoken corpora for SLA research is on the rise with several research questions being explored, there are notable gaps in the literature. The first relate to a dearth of two types of learner corpora: phonological and longitudinal. Phonological corpora include both audio (or video) data and time-aligned annotations of some phonological feature (Gut & Voormann, 2014). Some explanations for the limited research using L2 phonological corpora are (1) because relatively few phonological corpora exist, (2) some accessible spoken corpora lack sound files for researchers to make their own time-aligned annotations, and (3) creating time-aligned annotations of phonological features requires specialized phonological knowledge and much time. Some examples of L1 phonological corpora include IViE,[12] PFC,[13] and the child phonology component of the TalkBank, PhonBank.[14] While some L2 phonological corpora exist (e.g., L2-Arctic[15] and LeaP,[16] discussed later) they are certainly in the minority of spoken corpora. Similarly, (dense) longitudinal corpora are rare despite being argued to be particularly critical for SLA research (Granger, 2009; MacWhinney, 2017; Ortega & Iberri-Shea, 2005). As with phonological corpora, the limited number of longitudinal corpora is most likely because they are particularly time-consuming and expensive to compile and annotate.

Another gap pertains to limited investigations using corpora in addition to other types of research methods. Specifically, there have been multiple calls to conduct more work combining corpus research methods with experimental methods and for those using corpora for SLA research to make greater connections to SLA theory (McEnery et al., 2019; Myles, 2015). McEnery et al. (2019) argued that "the key fault line between SLA research and [learner corpus research] LCR" is that "SLA research has been largely theory-driven…test[ing] theory through psycholinguistic and other (quasi)experimental methods" while "by contrast, learner corpus researchers have been more exploratory and pre-theoretical in their approach to learner language" (p. 83). Meunier and Littre (2013) provide an example of this type of approach, albeit with written production. They investigated the development of tense and aspect in French-speaking learners of English using evidence from a longitudinal learner corpus of argumentative essays. Features identified as problematic for the learners based on continued errors in use from the corpus (e.g., the present progressive) were then used to create stimuli for multiple experimental tasks whose purpose was to tease apart which specific functions continued to be problematic.

4 Current Contributions and Research

The four corpora described here were selected to demonstrate breadth and variation with regard to the L1s/L2s represented, accessibility to the data/materials, tasks used for data collection, and research questions asked.

LANGSNAP

As described earlier, corpora have been used to investigate many research questions in spoken second language acquisition. The Languages and Social Networks Abroad Project (LANGSNAP, Mitchell et al., 2017) is a good example of the benefits of publicly shared longitudinal corpora and how a corpus can be designed and used to answer a wide range of research questions. The LANGSNAP corpus contains data from UK university students who were L2 learners of French or Spanish and required to spend their third year of a four-year degree programme living in a French- or Spanish-speaking country. From 2011 to 2013, 56 participants completed a picture-based narration and a semi-structured interview at each of six data collection points before, three times during, and two times after returning home from their 9-month sojourn abroad. Participants also completed an argumentative writing task. The audio files and transcriptions (in CHAT format, discussed later) are available for download on TalkBank. The oral data have been used to explore spoken language development of modality (McManus & Mitchell, 2015), CAF (McManus et al., 2020), L1–L2 fluency relationships (Huensch & Tracy-Ventura, 2017), and identity (Mitchell et al., 2020). Because it is a publicly available corpus, it has also been used by other research groups. For example, Gudmestad et al. (2019) explored the development of grammatical gender marking in L2 Spanish from a variationist SLA perspective and, using a multifactorial analysis, demonstrated how multiple linguistic (e.g., noun gender, noun frequency) and extralinguistic (e.g., task) factors contribute to different components of stability and variability in the gender marking of advanced L2 speakers. Data are still being added to this “productive” corpus. In 2016 and 2019, 33 and 31, respectively, of the original 56 speakers participated in two additional rounds of data collection, bringing the total project to 8 years and allowing new research questions examining factors that impact foreign language attrition/development/maintenance (Huensch et al., 2019).

LINDSEI

The LINDSEI corpus (Gilquin et al., 2010) has been used for an impressive number of research studies (see https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei-bibliography.html). LINDSEI was designed as a spoken counterpart to written argumentative essays provided in the *International Corpus of Learner English* (ICLE) corpus (Granger, 1998). The LINDSEI corpus consists of interviews with university-level English as a Foreign Language learners following a set structure in three parts. Each interview begins with a warm-up comprising a monologic speaking task on a given topic followed by an informal dialogic interview about speakers’ lives at university. To finish, speakers completed a picture description task. The corpus (transcripts only) is available for purchase and currently includes interviews with 554 participants. Two main strengths of the LINDSEI corpus are the variety of L1s represented (11 different backgrounds) and its parallel L1 English corpus, the *Louvain Corpus of Native English Conversation* (LOCNEC, De Cock, 2004). This design allows for both cross-linguistic and L1–L2 comparisons. Several studies have focused on discourse markers and other “small words” in LINDSEI (Buysse, 2012; Guilquin, 2008). For instance, Buysse (2012) explored spoken usage of the discourse marker *so* in the LINDSEI Dutch L1 subcorpus (n = 40 interviews) between learners majoring in English Linguistics versus those majoring in Commercial Sciences and also compared the learners to the L1 English LOCNEC corpus. Results indicated that both groups of learners and L1s evidenced the use of *so* in a variety of different functions, but that learners (from both majors) tended to overuse *so* in comparison to the L1 reference corpus. Similarly, Götz (2013) is a book-length treatment exploring native and non-native speaker utterance fluency using the German L1 subcorpus of LINDSEI. One finding from her analysis of the patterns of use of discourse markers was that learners often underused them and used a limited variety in comparison to native speakers. Rosen (2016) used the French L1 subcorpus of LINDSEI to explore the constructs of error and innovation by comparing the LINDSEI corpus data to a variety of English influenced by Norman French, Jersey English. Rosen's analysis brings together SLA research and research on indigenized varieties of English and asserts that "the difference between the notions of (not yet conventionalized) innovations on the one hand and errors on the other seems to be terminological and attitudinal – a matter of perspective and norm-orientation rather than a linguistic difference" (p. 304). Data collection for the LINDSEI corpus involves multiple researchers across several international institutions following a protocol to ensure that data collected are suitable for comparison. Additional subcorpora are continually being added to the LINDSEI corpus.

LeaP

The LeaP corpus (Gut, 2012) is one of the few L2 phonological corpora. The corpus consists of spoken data from L2 learners of German and English collected between 2001 and 2003. The project examined the acquisition of prosodic features (e.g., intonation, stress) and the potential impact of factors such as proficiency, formal instruction, and individual differences variables such as motivation and musicality. Over 12 hours of speech was collected from learners and native speakers completing tasks comprising both read and spontaneous speech. The reading tasks included a list of nonsense words and a narrative passage. The spontaneous speech tasks included a re-telling of the narrative passage and an informal interview. Time-aligned annotations were completed in Praat (Boersma & Weenink, 2020) and included segmentation of words, syllables, phonemes, tone, and pitch. The corpus (including sound files, texgrids, xml files, and manual) is freely available for download. One potential limitation is that the tools developed for its analysis are not publicly available and likely require basic knowledge of programming in *Perl* language (Edalatishams, 2017). Investigations of both the development of phonological features and oral language fluency have been published. For example, Gut (2017) used a subset of learners from the LeaP corpus to conduct a mixed-methods analysis of the effects of learning context on phonological development in different tasks over time. Contexts included study abroad, study abroad with participation in a phonology course, and at-home learners who participated in a phonology course. Phonological variables included vowel reduction, intonation, and fluency (articulation rate and mean length of run). The quantitative results showed no clear advantage for one of the contexts over another (although there were trends indicating benefits for the groups who received explicit teaching). Additionally, the qualitative analysis revealed a large amount of individual variability across learners in all contexts, and indicated that making gains in a phonological feature typically resulted in doing so across multiple tasks in the corpus.

CCOT

The Corpus of Collaborative Oral Tasks[17] (CCOT; Crawford, 2021) was created at Northern Arizona University between 2009 and 2012. The tasks in the corpus were given to students as part of their achievement tests during their study in an Intensive English Programme, from one to three times. There are 24 tasks, with at least ten learner performances of each task for a total of 775 files. There are 600 speakers from three proficiency levels. The most common tasks are problem solving (e.g., where learners decide which patient to treat or create an advertisement together). Both the audio files and the transcriptions are available by contacting the creator, William Crawford. An edited volume (Crawford, 2021) includes research on lexico-grammar, pronunciation, and other types of speech analysis. For example, Staples (2021) investigates lexico-grammatical features (e.g., nouns, conditional clauses, *that* complement clauses), interactional features (turn length, backchannels, questions), fluency (speech rate, length of pauses), and pitch range across task types (informational and argumentative). Not surprisingly, nouns and other informational features were used more in the informational task while conditional clauses were more common in the argumentative task. These findings align with numerous studies supporting the use of these lexico-grammatical features for these particular purposes. However, perhaps more interesting are the findings for interactional features, fluency, and pitch range. Backchannels were used more frequently in informational tasks, perhaps reflecting the listener's uptake of information provided by the speaker. Speech rate was faster and number of pauses was lower for the argumentative tasks, likely reflecting the less dense use of informational content in the argumentative tasks. Pitch range was also higher in the argumentative task, perhaps due to the need to stress syllables at higher pitch to make points more salient and arguments appear stronger. These findings have important implications for the understanding of interactional variables, fluency, and pronunciation across tasks.

5 Main Research Methods

Corpus Building and Research Design

Methods for corpus design rely heavily on the definition of corpus used by the researcher. This part assumes the definition used by corpus linguists: spoken corpora consist of speech samples from a naturally occurring discourse domain. From this perspective, corpus developers generally work to ensure two characteristics of corpora: balance and representativeness. Balance refers to providing appropriate numbers of texts associated with subdomains within the research domain one is investigating. For example, researchers in SLA often work to balance the data across task type, or across speaker L1 groups, among other variables. Depending on the research questions to be answered by the corpus, the type of balance required will change. In addition, corpus developers work to ensure that the sample included in their corpus is representative of the domain they are trying to represent. So, a corpus that consists of words read aloud cannot represent conversational discourse. However, a corpus of spoken assignments can represent what learners are doing in a classroom context. Thus, it is important to consider the research questions when evaluating the representativeness of the corpus for a given project. In addition to evaluating extra-linguistic characteristics of the corpus (e.g., L1 background of the speakers or task types), representativeness can also be evaluated linguistically. Depending on the type of language data a researcher is investigating, a corpus may be more or less representative of that language feature. For example, if a researcher is investigating syllable stress, as long as multi-syllabic words are represented in the corpus, the corpus can be used for that linguistic feature; the corpus size may be small as long as there are multi-syllabic words at a high enough rate. However, if a researcher is investigating particular idioms, it will be harder to find a representative corpus, as some idioms occur quite infrequently and thus are not well represented in all spoken corpora. This is a reason to have a large corpus. In general, features like pausing (both filled and unfilled), stress patterns, vowel or consonant sounds, or grammatical features that are common to speech can be well represented in many types of corpora. However, to investigate intonation and rhythm, more naturally occurring speech is needed. To investigate vocabulary, larger corpora are needed.

Corpus methods typically take three different approaches to research design (coined Type A, Type B, and Type C by Biber & Jones, 2009). In type A studies, researchers investigate a linguistic feature to determine how that feature varies based on the linguistic environment. For example, one might examine copular verbs in Spanish and Portuguese learner corpora to see how they vary depending on type of complement. Logistic regression can then identify whether the patterns vary across L1 background or learner level (Picoral, 2020).

Type B studies take as their unit of observation an individual text. Linguistic features are examined within each text, but the output is the frequency of occurrence of that feature in each text. Thus, the focus is not on the behaviour of a linguistic feature in a linguistic environment, but rather how frequent that feature is used across L1 backgrounds, learner levels, and/or text types. Types of statistical methods used with these corpora include the ANOVA family, to investigate differences across subgroups (e.g., by proficiency level or L1 background, for example) or from the correlation/regression family, to determine relationships between a continuous operationalization of proficiency (e.g., scores on a proficiency test) and linguistic features.

Type C studies are similar to type B, but they investigate the frequencies of an entire subcorpus rather than getting the frequency for each individual text within that subcorpus. In this case, the use of inferential statistics is more limited, and it is commonplace for researchers to report frequency data. Reporting range along with normed frequencies is advisable, to help researchers determine whether the phenomena are spread throughout speakers in a subcorpus or are used by only one or two speakers.

Most corpora are sampled from one period of time and thus are typically cross-sectional. Corpus compilers ideally balance the corpus across score or proficiency levels, and also typically try to balance across L1 backgrounds. Such corpora provide valuable information about linguistic and other features that characterize performance at different levels. However, more recently, there has been a call for more quantitative longitudinal studies. Longitudinal corpora provide an ideal dataset for examining spoken development. One of the choices researchers must make is whether they prioritize the similarity of task across time periods (e.g., the same task is administered to learners at two or more points in time) or whether they want to prioritize the type of tasks suitable for learners at different developmental stages. The former has the obvious advantage of being more controlled, while the latter has the advantage of more ecological validity. Researchers are exploring these two options in corpus data, and it is clear that new methods are needed to address different types of longitudinal datasets.

For corpora consisting of read words or sentences, balance and representativeness are not important considerations. As discussed earlier, such corpora have the advantage that they can be designed to have the control of a psycholinguistic experimental setting with shared prompts and lab-quality recordings but have the disadvantage of not representing a spoken discourse domain. Methods for these types of corpora are similar to those for psycholinguistic data, discussed in Nagle et al., this volume.

Digital Tools

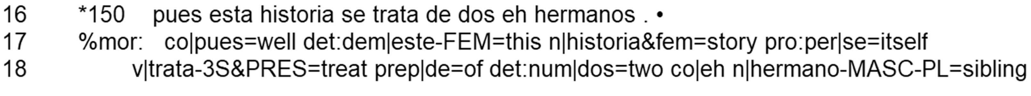
A variety of digital tools exist to transcribe, annotate (including tagging and segmentation), and analyze spoken corpora. This part describes these processes and some of the most commonly used tools to complete them.

Transcription is the process of representing oral language in some form of written script, such as orthographic transcription (e.g., following typical spelling conventions) or phonemic or phonetic transcription (e.g., using IPA symbols and diacritics). While digital tools can assist in (semi-) automating other processes with spoken corpora, manual transcription is often a necessary and time-consuming first step. For instance, Brezina et al. (2019) reported that it took 5 years and nearly 3,500 hours to transcribe the TLC (see footnote 8). Spoken corpora can be transcribed in text editors (e.g., Microsoft Notepad++, Mac TextEdit) or software programs specialized for linguistic analysis such as the freely available CLAN or ELAN. CLAN (Computerized Language Analysis, MacWhinney, 2000) is a software program developed for the TalkBank system. CLAN is designed to work in conjunction with the CHAT (Codes for Human Analysis of Transcripts) transcription and coding format, a set of standardized conventions for creating computerized transcripts of speech. ELAN (EUDICO Linguistic Annotator, Wittenburg et al., 2006) is another software program that allows for transcription and analysis of audio and video. A useful feature of ELAN is its organization around tiers, which can be hierarchically structured.

Annotation is the process of providing additional linguistic information to the transcription. One of the most common forms of annotation across both written and spoken corpora is known as tagging. This is the process of marking up words in the corpus with part-of-speech information based on the word and its context. For example, in Figure 8.1, lines 17 and 18 represent the part-of-speech (POS) tagged words from the orthographically transcribed Spanish utterance in line 16. As shown in Figure 8.1, POS tagging in this case provides information about word class, tense, gender, number, etc. Typically, the tagging process is automatic, although some follow-up disambiguation might be necessary depending on the accuracy of the tagger. At a minimum, it is important to include accuracy checking as one of the steps when using automatic annotators.

Once transcribed and potentially POS annotated, concordancing tools can be used for analyses related to the frequency and distribution of words in a corpus. These often involve extracting not only key words or phrases, but also the words occurring before and after them [known as Key Word In Context (KWIC) analyses]. These tools are available as stand-alone (e.g., AntConc) or web-based (e.g., SketchEngine) applications and have been used not only for linguistic research, but also as pedagogical tools. For example, SketchEngine provides access to 500+ corpora in over 90 languages, but researchers can also upload their own corpora for analysis.

Many other forms of annotation are possible (for an overview, see Leech, 2005). Regarding annotation specific to spoken corpora, for example, prosodic annotation could be used to indicate information about intonation, stress, and pausing. Additionally, symbols may be added to transcripts for features such as filled pauses (e.g., *uh*, *um*), repeated or reformulated words or phrases. Segmentation is a common form of annotation in spoken corpora and can be used at multiple levels. For instance, segmentation might be used to separate speech from silence, to indicate discourse units such as turns, to separate phonemes or syllables within a word, etc. In addition to ELAN, Praat is a commonly used digital tool for segmentation and annotation of speech. Annotations in Praat are created in TextGrid files, which can have multiple tiers as shown in Figure 8.2. Praat has a built-in feature to automatically segment silence from speech called *Annotate To TextGrid (silences…)*. Its accuracy varies depending on the sound quality of the file, so manual post-checking is recommended. Given Praat's wide usage, many other digital corpus tools (e.g., CLAN, ELAN,



```
16    *150   pues esta historia se trata de dos eh hermanos . •
17    %mor:  co|pues=well det:dem|este-FEM=this n|historia&fem=story pro:per|se=itself
18           v|trata-3S&PRES=treat prep|de=of det:num|dos=two co|eh n|hermano-MASC-PL=sibling
```

Figure 8.1 Example POS-tagged utterance in the CLAN software program

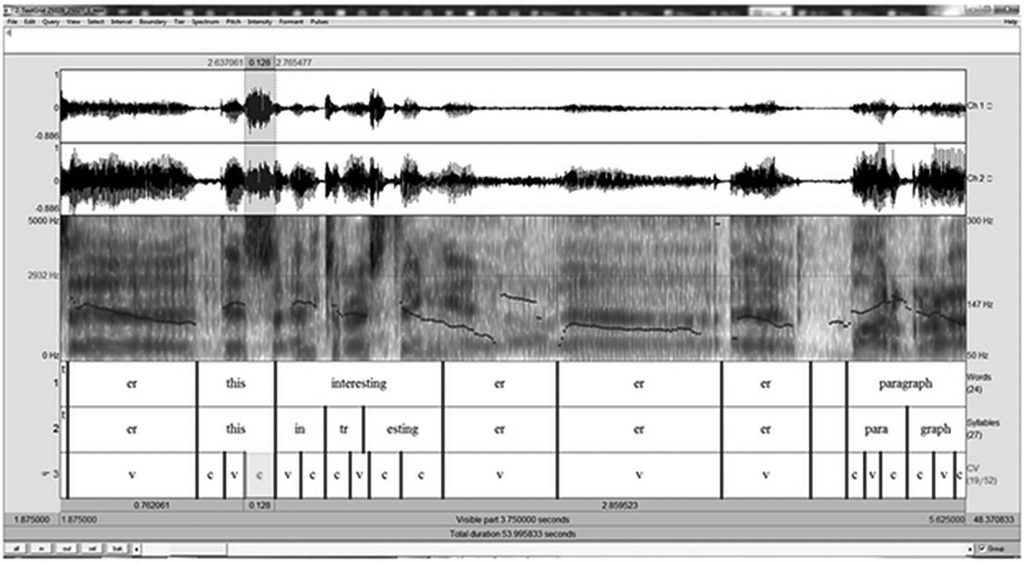


Figure 8.2 Praat TextGrid with annotation including a word tier (1), syllable tier (2), and a consonant/vowel tier (3). Reprinted from Ghanem et al. (2020) with permission

Phon) can read and/or write Praat TextGrid files. Praat has also been used for annotation of phonological corpora, which are particularly time- and effort-intensive to annotate. For example, the LeaP corpus involved approximately 1,000 events annotated per minute (Gut, 2012) whose reliability varied as a factor of what was being annotated. One of the most important benefits of annotation, however, is that once completed, it allows for automatic analysis of the corpus.

Ghanem et al. (2020) provide an overview and evaluation of five commonly used digital tools for spoken discourse, and recommendations for combining them efficiently across different stages of data preparation and analysis for pronunciation corpora. They provide documentation and evaluation of the five digitals tools on their website.[18]

6 Recommendations for Practice

Increasing the trend of data-sharing is an important and efficient way to move forward. A simple step is to encourage researchers collecting oral data to include permissions in IRB/ Human Ethics consent forms for sharing data. TalkBank provides a template and another example can be found on the Ghanem et al. (2020) website (see footnote 18). Once approval has been received, multiple options exist for sites to share sound files and other data, such as researchers' institutional websites (if available), SLABank (if CHAT transcripts are included), Instruments for Research into Second Languages (IRIS, Marsden et al., 2016), or the Open Science Framework (OSF, osf.io). For those data including annotation, it is useful to share protocols describing annotation procedures and decisions. Documenting and providing access to these facilitates other researchers' use of the data. Finally, in choosing a data format, researchers should consider using programs designed with interoperability in mind (e.g., CLAN, ELAN, Praat) and/or plain text. Beyond data-sharing efforts, encouraging project collaboration across multiple institutions and researchers is another possibility for building corpus resources that has been successful in the past (e.g., the LINDSEI project).

Two major database providers, CECL and TalkBank, provide suggestions and/or example guidelines for creating new corpora and embarking on multi-institutional collaborations.

Another recommendation is to increase training opportunities and materials for researchers who currently work with or would like to work with spoken corpora. Many programs are available for transcription, annotation, and analysis of spoken corpora – so many that those new to the field might have difficulty deciding where to start. More established programs (e.g., Praat, AntConc) usually have detailed documentation user guides on the web. Providing additional training opportunities during pre-conference workshops or conference presentations at venues such as the American Association of Corpus Linguistics (AACL), the Pronunciation in Second Language Learning and Teaching (PSLLT) conference, or the Second Language Research Forum (SLRF) or as part of a summer workshop session is a great way to increase skills and encourage wider use of spoken corpora for SLA.

7 Future Directions

Ultimately, while many potential benefits of using spoken corpora for SLA research exist, the field is in its early stages. We have indicated areas of research that represent important next steps. Thus far, we have highlighted the need for more longitudinal and phonological corpora and research as well as examinations of spoken SLA combining experimental and corpus-based methods. For projects such as these, collaborative efforts that bring together researchers from multiple institutions and methodological expertise are likely to be most successful. Such efforts require careful planning as well as consistency in data collection and preparation.

One future direction that deserves mention is the need for spoken corpora representing less commonly taught languages (LCTLs) as well as greater L1–L2 pairings in general. Not surprisingly, most corpora represent languages such as English, French, or Spanish. One project working to collect data from two LCTLs, Russian and Portuguese, is the Multilingual Corpus of Assignments – Writing and Speech (MACAWS).[19]

Notes

[1] To clarify, by "second language acquisition of speaking," we mean L2 acquisition (in or outside instructional contexts) in the spoken mode, including investigation of phonology, syntax, and pragmatics.

[2] Université Catholique de Louvain's Centre for English Corpus Linguistics (CECL); https://uclouvain.be/en/research-institutes/ilc/cecl/corpora.html

[3] SLABank; MacWhinney (2020), https://slabank.talkbank.org/

[4] European Science Foundation Second Language (ESF); (Perdue, 1993), https://slabank.talkbank.org/access/Multiple/ESF/

[5] Languages and Social Networks Abroad Project (LANGSNAP); (Mitchell et al., 2017), http://langsnap.soton.ac.uk/, https://scholarcommons.usf.edu/langsnap/

[6] French Learner Language Oral Corpora (FLLOC); http://www.flloc.soton.ac.uk/

[7] Spanish Learner Language Oral Corpora (SPLLOC); http://www.splloc.soton.ac.uk

[8] Louvain International Database of Spoken English Interlanguage (LINDSEI); (Gilquin et al., 2010), https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html

[9] Trinity-Lancaster Corpus (TLC); (Brezina et al., 2019), http://cass.lancs.ac.uk/trinity-lancaster-corpus/

[10] Parallèle Oral en Langue Etrangère 'Parallel Oral Foreign Language' (Parole); (Hilton, 2009), https://slabank.talkbank.org/access/English/PAROLE.html

[11] What is Speaking Proficiency (WiSP); (De Jong et al., 2015).

[12] Intonational Variation in English (IViE); (Grabe et al., 2001), http://www.phon.ox.ac.uk/files/apps/IViE/

13 Phonologie du Français Contemporain (PFC); (Durand et al., 2002), https://www.projet-pfc.net/
14 PhonBank; Rose and MacWhinney (2014), https://phonbank.talkbank.org/
15 L2-ARCTIC; Zhao et al. (2018), https://psi.engr.tamu.edu/l2-arctic-corpus/
16 Learning Prosody in a Foreign Language (LeaP); (Gut, 2012); https://sourceforge.net/projects/leapcorpus/
17 Corpus of Collaborative Oral Tasks (CCOT); Crawford (under contract)
18 Digital Tools Used with Pronunciation Corpora; Ghanem et al. (2020), https://sites.google.com/view/psllt2019/home
19 Multilingual Corpus of Assignments – Writing and Speech (MACAWS), http://macaws.corporaproject.org

Further Reading

Biber, D., & Reppen, R. (Eds). (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge: Cambridge University Press.

This handbook covers major research areas within corpus linguistics, with an emphasis on English but with relevance to research of other languages as well. The book contains helpful chapters on common research areas, such as keyword and collocational analysis, as well as introductions to both spoken corpus and learner corpus research.

Granger, S., Gilquin, G., & Meunier, F. (Eds.) (2015). *The Cambridge handbook of learner corpus research*. Cambridge: Cambridge University Press.

This handbook provides a comprehensive guide to the rapidly-developing field of learner corpus research. The volume contains 27 chapters divided among parts devoted to corpus design and methodology, learner language analysis, and intersections between learner corpus research and SLA, language teaching, and natural language processing.

Tracy-Ventura, N., & Paquot, M. (2020). *The Routledge handbook of SLA and corpora*. New York: Routledge.

This handbook begins with introductory chapters on corpus linguistics, LCR, SLA, and the intersections of SLA and LCR. The remainder of the handbook is comprised of three parts (a) aspects of corpus design, annotation, and analysis, (b) the role of corpora in SLA theory and practice, and (c) SLA constructs (e.g., input, interaction, accuracy) and corpora. The handbook ends with a chapter on future directions of the use of corpora in SLA.

References

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, D., Gray, B., & Staples, S. (2016). Predicting patterns of grammatical complexity across textual task types and proficiency levels. *Applied Linguistics*, *37*, 639–668.

Biber, D., & Jones, J. K. (2009). Quantitative methods in corpus linguistics. In *Corpus linguistics: An international handbook* (pp. 1286–1304). Berlin: De Gruyter Mouton.

Boersma, P., & Weenink, D. (2020). Praat: Doing phonetics by computer (Version 6.1.09) [Computer program]. Retrieved from http://www.praat.org/

Brezina, V., Gablasova, D., & McEnery, T. (2019). Corpus-based approaches to spoken L2 production: Evidence from the Trinity Lancaster Corpus. *International Journal of Learner Corpus Research*, *5*, 119–125.

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

Castello, E., & Gesuato, S. (2019). Holding up one's end of the conversation in spoken English: Lexical backchannels in L2 examination discourse. *International Journal of Learner Corpus Research*, *5*, 231–252.

Chomsky, N. (1962). Paper given at the University of Texas 1958. In *3rd Texas conference on problems of linguistic analysis in English*. Austin, TX: University of Texas.

Crawford, W. (2021). *Multiple perspectives on learner interaction: The corpus of collaborative oral tasks*. New York: DeGruyter.

Crossley, S. A., Salsbury, T., & Mcnamara, D. S. (2015). Assessing lexical proficiency using analytic ratings: A case for collocation accuracy. *Applied Linguistics*, *36*, 570–590.

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

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, *36*, 223–243.

Durand, J., Laks, B., & Lyche, C. (2002). La phonologie du français contemporain: usages, variétés et structure. In C. Pusch & W. Raible (Eds.), *Romanistische Korpuslinguistik- Korpora und gesprochene Sprache/Romance corpus linguistics – Corpora and spoken language* (pp. 93–106). Tübingen: Gunter Narr Verlag.

Edalatishams, I. (2017). LeaP corpus (review). In M. O'Brien & J. Levis (Eds.), *Proceedings of the 8th pronunciation in second language learning and teaching conference*, ISSN 2380-9566, Calgary, AB, August 2016 (pp. 236–240). Ames, IA: Iowa State University.

Ellis, N. C., Römer, U. & O'Donnell, M. B. (2016). *Usage-based approaches to language acquisition and processing: cognitive and corpus investigations of construction grammar*. Language Learning Monograph Series. Hoboken, NJ: Wiley-Blackwell.

Fernández, J. (2013). A corpus-based study of vague language use by learners of Spanish in a study abroad context. In C. Kinginger (Ed.), *Social and cultural aspects of language learning in study abroad* (pp. 299–332). Philadelphia: John Benjamins.

Fernández, J., & Yuldashev, A. (2011). Variation in the use of general extenders and stuff in instant messaging interactions. *Journal of Pragmatics*, *43*, 2610–2626.

Friginal, E., Lee, J. J., Polat, B., & Roberson, A. (2017). *Exploring spoken English learner language using corpora: Learner talk*. New York: Springer.

Gablasova, D., Brezina, V., & McEnery, T. (2019). The Trinity Lancaster corpus: Development, description and application. *International Journal of Learner Corpus Research*, *5*, 126–158.

Gilquin, G. (2008). Hesitation markers among EFL learners: Pragmatic deficiency or difference? In J. Romero-Trillo (Ed.), *Pragmatics and corpus linguistics: A mutualistic entente* (pp. 119–149). Berlin: Mouton de Gruyter.

Gilquin, G. (2019). Light verb constructions in spoken L2 English: An exploratory cross-sectional study. *International Journal of Learner Corpus Research*, *5*, 181–206.

Ghanem, R., Edalatishams, I., Huensch, A., Puga, K., & Staples, S. (2020). The effectiveness of computer programs in the transcription and analysis of spoken discourse: towards a protocol for pronunciation corpora. In O. Kang, S. Staples, K. Yaw, & K. Hirschi (Eds.), *Proceedings of the 11th pronunciation in second language learning and teaching conference* (pp. 97–114). Ames, IA: Iowa State University.

Gilquin, G., De Cock, S., & Granger, S. (2010). The Louvain international database of spoken English interlanguage. *Handbook and CD-ROM. Louvain*. Belgium: Presses universitaires de Louvain.

Gilquin, G., & Gries, S. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*, 1–26.

Götz, S. (2019). Filled pauses across proficiency levels, L1s and learning context variables: A multivariate exploration of the Trinity Lancaster Corpus Sample. *International Journal of Learner Corpus Research*, *5*, 159–180.

Götz, S. (2013). *Fluency in native and nonnative English speech*. Philadelphia, PA: John Benjamins.

Grabe, E., Post, B. & Nolan, F. (2001). *The IViE corpus.* Department of Linguistics, University of Cambridge. http://www.phon.ox.ac.uk/old_IViE.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. *Corpora and Language Teaching*, *33*, 13–32.

Granger, S. (1998). The computerized learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3–18). London: Longman.

Gudmestad, A., Edmonds, A., & Metzger, T. (2019). Using variationism and learner corpus research to investigate grammatical gender marking in additional language Spanish. *Language Learning*, *69*, 911–942.

Gut, U. (2017). Phonological development in different learning contexts. *International Journal of Learner Corpus Research*, *3*, 196–222.

Gut, U. (2012). 'The LeaP corpus. A multilingual corpus of spoken learner German and learner English. In Th. Schmidt, & K. Wörner, K. (Eds.), *Multilingual corpora and multilingual corpus analysis* (pp. 3–23). Amsterdam: John Benjamins.

Gut, U., & Voormann, H. (2014). Corpus design. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 13–26). Oxford: Oxford University Press.

Hilton, H. E. (2014). Oral fluency and spoken proficiency: Ideas for testing and research. In P. Leclercq, A. Edmonds, & H. Hilton (Eds.), *Measuring L2 proficiency: Perspectives from SLA* (pp. 27–53). Bristol, UK: Multilingual Matters.

Hilton, H. (2009). Annotation and analyses of temporal aspects of spoken fluency. *CALICO Journal*, *26*, 644–661.

Huensch, A. (2020). Fluency. In N. Tracy-Ventura, & M. Paquot (Eds.), *The Routledge handbook of SLA and corpora*. New York: Routledge.

Huensch, A., & Staples, S. (2018). Towards a protocol for a multilingual corpus for pronunciation researchers. *Pronunciation in Second Language Learning and Teaching*Ames, Iowa. https://apling.engl.iastate.edu/conferences/pronunciation-in-second-language-learning-and-teaching-conference/psllt-archive/

Huensch, A., & Tracy-Ventura, N. (2017). Understanding second language fluency behavior: The effects of individual differences in first language fluency, cross-linguistic differences, and proficiency over time. *Applied Psycholinguistics*, *38*, 755–785.

Huensch, A., Tracy-Ventura, N., Bridges, J., & Cuesta-Media, J. (2019). Variables affecting the maintenance of L2 proficiency and fluency four years post-study abroad. *Study Abroad Research in Second Language Acquisition and International Education*, *4*, 96–125.

Leech, G. (2005). Adding linguistic annotation. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 17–29). Oxford: Oxbow Books, http://users.ox.ac.uk/~martinw/dlc/index.htm

MacWhinney, B. (2020). TalkBank and SLA. In N. Tracy-Ventura, & M. Paquot (Eds.), *The Routledge handbook of SLA and corpora*. New York: Routledge.

MacWhinney, B. (2017). A shared platform for studying second language acquisition. *Language Learning*, *67*, 254–275.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd edn). Mahwah, NJ: Lawrence Erlbaum.

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS repository of instruments for research into second languages* (pp. 1–21). New York: Routledge.

McEnery, T., Brezina, V., Gablasova, D., & Banerjee, J. (2019). Corpus linguistics, learner corpora, and SLA: Employing technology to analyze language use. *Annual Review of Applied Linguistics*, *39*, 74–92.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.

McEnery, T., & Wilson, A. (2001). *Corpus linguistics* (2nd edn). Edinburgh: Edinburgh University Press.

McManus, K., Mitchell, R., & Tracy-Ventura, N. (2020). Longitudinal study of advanced learners' linguistic development before, during, and after study abroad. *Applied Linguistics*. doi:10.1093/applin/amaa003.

McManus, K. & Mitchell, R. F. (2015). Subjunctive use and development in L2 French: A longitudinal study. *Language, Interaction and Acquisition*, *6*(1), 42–73.

Meunier, F., & Littre, D. (2013). Tracking learners' progress: Adopting a dual 'corpus cum experimental data' approach. *The Modern Language Journal*, *97*, 61–76.

Mitchell, R., Tracy-Ventura, N., & Huensch, A. (2020). After study abroad: The long-term evolution of multilingual identity among anglophone languages graduates. *Modern Language Journal.*,*104*(2), 327-344

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *The Anglophone student abroad: Identity, social relationships and language learning*. New York: Routledge.

Myles, F. (2015). Second language acquisition theory and learner corpus research. In A. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge handbook of learner corpus research* (pp. 309–331). Cambridge: Cambridge University Press.

Myles, F. (2008). Investigating learner language development with electronic longitudinal corpora: Theoretical and methodological issues. In L. Ortega, & H. Byrnes (Eds.), *The longitudinal study of advanced L2 capacities* (pp. 58–72). New York: Routledge.

Ortega, L., & Iberri-Shea, G. (2005). Longitudinal research in second language acquisition: Recent trends and future directions. *Annual Review of Applied Linguistics*, *25*, 26–45.

Perdue, C. (Ed.) (1993). *Adult language acquisition. Vol 1: field methods*. Cambridge: Cambridge University Press.

Pérez-Paredes, P., & Díez-Bedmar, M. B. (2019). Certainty adverbs in spoken learner language: The role of tasks and proficiency. *International Journal of Learner Corpus Research*, *5*, 253–279.

Picoral, A. (2020). L3 Portuguese by Spanish-English bilinguals: Copula construction use and acquisition in corpus data (Publication No. 27957666). [Doctoral dissertation, University of Arizona]. ProQuest Dissertations Publishing.

Polat, B. (2011). Investigating acquisition of discourse markers through a developmental learner corpus. *Journal of Pragmatics*, *43*, 3745–3756.

Rose, Y., & MacWhinney, B. (2014). The PhonBank project: Data and software-assisted methods for the study of phonology and phonological development. In J. Durand, U. Gut, & G. Kristoffersen (Eds.), *The Oxford handbook of corpus phonology* (pp. 380–401). Oxford: Oxford University Press.

Rosen, A. (2016). The fate of linguistic innovations: Jersey English and French learner English compared. *International Journal of Learner Corpus Research*, *2*, 302–322.

Römer, U., & Garner, J. R. (2019). The development of verb constructions in spoken learner English: Tracing effects of usage and proficiency. *International Journal of Learner Corpus Research*, *5*, 207–230.

Staples, S. (2021). Exploring the impact of situational characteristics on the linguistic features of spoken oral assessment tasks. In W. Crawford (Ed.), *Multiple perspectives on learner interaction: The corpus of collaborative oral tasks* (pp. 123–144) Berlin: DeGruyter.

Staples, S., LaFlair, G., & Egbert, J. (2017). A multi-dimensional comparison of oral proficiency interviews to conversation, academic and professional spoken registers. *Modern Language Journal*, *101*, 194–213.

Tracy-Ventura, N., & Huensch, A. (2018). The potential of publicly shared longitudinal learner corpora in SLA research. In A. Gudmestad & A. Edmonds (Eds.), *Critical reflections on data in second language acquisition* (pp. 149–170). Philadelphia/Amsterdam: John Benjamins.

Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics*, *38*, 90–111.

Wittenburg, P., Brugman, H., Russel, A., Klassmann, A., Sloetjes, H. (2006). ELAN: A professional framework for multimodality research. In *Proceedings of LREC 2006, Fifth international conference on language resources and evaluation*. https://tla.mpi.nl/tools/tla-tools/elan/

Zhao, G., Sonsaat, S., Silpachai, A., Lucic, I., Chukharev-Hudilainen, E., Levis, J. M., & Gutierrez-Osuna, R. (2018). L2 ARCTIC: A non-native English speech corpus. *Proceedings of interspeech (Hyderabad, India)*.

Appendix 8A List of Spoken Corpora (modified from Huensch & Staples, 2018)

L2 Corpora and Datasets

1. BeMaTaC (Berlin Map Task Corpus) https://hu-berlin.de/bematac
2. DiapixFL https://datashare.is.ed.ac.uk/handle/10283/346
3. EuroCoAT (European Corpus of Academic Talk) http://www.eurocoat.es/web_sections_1/the_corpus_eurocoat_the_european_corpus_of_academic_talk_12
4. FLLOC (French Learner Language Oral Corpora) http://www.flloc.soton.ac.uk/
5. The Hong Kong Bilingual Child Language Corpus http://www.cuhk.edu.hk/lin/home/bilingual.htm
6. Hong Kong Corpus of Spoken English http://rcpce.engl.polyu.edu.hk/HKCSE/
7. IDEA (International Dialects of English Archive) http://www.dialectsarchive.com
8. IJAS (International Corpus of Japanese as a Second Language) https://chunagon.ninjal.ac.jp/static/ijas/about.html
9. Japanese polite speech by native speakers and non-native speakers https://www120.secure.griffith.edu.au/research/items/b11042e5–7588-4d0d-b1ea-dad2320716cc/1/
10. Japanese learners' conversations (contains OPI interviews with transcriptions) https://nknet.ninjal.ac.jp/nknet/ndata/opi/
11. L2-ARCTIC https://psi.engr.tamu.edu/l2-arctic-corpus/
12. L2 Mandarin Chinese by non-native speakers https://www120.secure.griffith.edu.au/research/items/9a3e0b74-20f8–4229-baf1-d9ec84d300da/1/

13. LeaP (Learning Prosody in a Foreign Language) https://benjamins.com/#catalog/books/hsm.14.03gut/details
14. LANGSNAP (Languages and Social Networks Abroad Project) http://langsnap.soton.ac.uk/; http://scholarcommons.usf.edu/langsnap/
15. LINDSEI (Louvain International Database of Spoken English Interlanguage) https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html
16. MICASE (Michigan Corpus of Academic Spoken English) https://quod.lib.umich.edu/m/micase/
17. NBTale (Norwegian database) https://www.nb.no/sprakbanken/show?serial=sbr-31&lang=en
18. NIM (Spanish, English and Catalan) https://psico.fcep.urv.cat/utilitats/nim/eng/about.php
19. PRESEEA http://preseea.linguas.net/
20. Speech Accent Archive http://accent.gmu.edu
21. Spin TX (Spanish in Texas) http://spanishintexas.org/https://www.coerll.utexas.edu/spintx/home
22. SPLLOC (Spanish Learner Language Oral Corpora) http://www.splloc.soton.ac.uk/
23. Trinity Lancaster Corpus (http://cass.lancs.ac.uk/trinity-lancaster-corpus/)
24. Wildcat corpus http://groups.linguistics.northwestern.edu/speech_comm_group/wildcat/
25. VOICE (Vienna-Oxford International Corpus of English) https://www.univie.ac.at/voice/

L1 Phonology Corpora

1. IViE (Intonational Variation in English) http://www.phon.ox.ac.uk/files/apps/IViE/
2. NoTa-Oslo (Norwegian Spoken Language Corpus) http://www.tekstlab.uio.no/nota/oslo/english.html
3. PFC Programme (Phonologie du Français Contemporain: usages, variétés et structure) https://www.projet-pfc.net/
4. PhonBank https://phonbank.talkbank.org/
5. TAUS (Spoken Language Investigation in Oslo) http://www.tekstlab.uio.no/nota/taus/english.html

Other Widely Used Spoken Corpora

1. ANC (American National Corpus) http://www.anc.org/
2. BASE (British Academic Spoken English Corpus) https://warwick.ac.uk/fac/soc/al/research/collections/base/
3. BNC (British National Corpus Audio Edition) http://www.phon.ox.ac.uk/AudioBNC
4. BYU Corpora https://corpus.byu.edu/
5. COCA (Corpus of Contemporary American English) https://corpus.byu.edu/COCA/
6. C-Oral-Rom (Benjamins) https://benjamins.com/#catalog/books/scl.15/main
7. ICE (International Corpus of English) http://ice-corpora.net/ice/index.html
8. Santa Barbara http://www.linguistics.ucsb.edu/research/santa-barbara-corpus

Appendix 8B Summary of Select Existing Corpora (modified from Huensch & Staples, 2018)

Corpus	*Language/Proficiency*	*Size*	*Data/Annotations*	*Strengths*	*Weaknesses*	*Task*	*Access*
BeMaTaC	German L2 (n = 10); [advanced prof = C1/C2]German NS (n = 24)	18,123	Transcripts (EXMARaLDA) Sound files (wav, mp3) Video files (mov, webm)	Rich metadata Dialogic	Limited PRON annotations Single task	Info-gap	Free; download ANNIS
CCOT	English L2 (n = 600);three proficiency levels [TOEFL 32–69]	268,324; 775 files	TranscriptsSound files (wav)	Dialogic; multiple L1s represented; multiple tasks represented	No PRON annotations Variable sound quality	24 different tasks	Free; contact creator (William.Crawford@nau.edu)
FLLOC	Collection of 8 corpora(n = 491 participants, aged 5–23) [varying proficiency]	40,00 files>3 million words	Transcripts (CHAT/CLAN) Sound files (wav, mp3) POS-tagged/MOR	Large; Comparisons w/ SPLLOC	Limited PRON annotations Variable sound quality	Elicitation tasks; Narratives; Interview	Free; download
HKCSE (Prosodic)	L1 Hong Kong Chinese, L2 English (n=643,286)[advanced proficiency] L1 English (227,894) L1 Other (29,064)	900,214 words; 311 recordings	Transcripts (txt) Brazil annotations (searchable through iConc interface)	Large; Wide range of naturally occurring tasks; Annotation using Brazil's system	Highly proficient L2 speakers No sound files	Business (e.g., job interview; presentations, service encounters); Academic (student presentations, lectures); Public (speeches, interviews)	$125CD-ROM
LANGSNAP	French L2 (n=29) Spanish L2 (n=27) [advanced proficiency] French NS (n=10) Spanish NS (n=10)	742,203 words; 1,238 files	Transcripts (CHAT/CLAN)Sound files (wav, mp3)POS-tagged/MOR	Longitudinal; Controlled and free tasks; Multiple L2s	Limited PRON annotations Variable sound quality	Interview; Story retell; Essay	Free; download
LeaP	L1 German, L2 English (n=176) L1 English, L2 German (n=183) [advanced and intermediate? proficiency] L1 English (n=8) L1 German (n=10)	73,941 words; 12 hours	Transcripts (XML-based TASX format) Syllables, segments, pitch accents and boundary tones, intonation contours, part of speech, lemmas (Praat) Sound files (wav)	Detailed segmental/suprasegmental annotation; Controlled and free tasks	Low inter-annotator agreement; Relatively small	Free speech in an interview; Reading a story; Retelling a story; Reading nonsense word list	Free; download

LINDSEI	~50 files each of L2 English from the following L1s: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish, Swedish (Intermediate to Advanced [10 files per L1 CEFR evaluated])	> 1 million words; 792,000 of learner language 554 interviews 130 hours	Transcripts (XML) Fluency annotations: filled and unfilled pauses	Multiple L1s; Controlled and free tasks	Limited PRON annotations No sound files Proficiency information not clear (only 5 interviews from each group were rated)	Interview; Informal chat; Picture description	€211.75CD-ROM
Speech Accent Archive	L1 and L2 English; ~350 L1s [proficiency not provided]	2,642 samples 69 words per sample	Phonetic transcription	Extremely broad range of L1s; Comparability across samples	Read speech No proficiency information	One passage of read speech	Free
SPLLOC	2 corpora, each having: L2 Spanish (*n*=60) [varying proficiency] L1 Spanish (*n*=15)	575 files	Transcripts (CHAT/ CLAN)Sound files (wav, mp3) POS-tagged/MOR	Comparisons w/ FLLOC	Limited PRON annotations Variable sound quality	Elicitation tasks; Narratives; Interview	Free; download