

COMPILING, ANNOTATING, AND ANALYZING SPOKEN CORPORA

Eric Friginal, Georgia State University

Corpus-based analyses of spoken discourse have provided directions for empirical investigations of linguistic features in various types of formal and informal conversations. Work in this area has incorporated corpus-based methods in analyzing a range of academic and professional registers, and various recent innovations are making it possible, albeit slowly, to successfully merge corpus approaches with those employed in studies of segmental and suprasegmental pronunciation. This paper presents highlights of a workshop and a demonstration overviewing the process of designing, compiling, and annotating spoken corpora for participants at the Pronunciation in Second Language Learning and Teaching (PSLLT) Conference 2019. The theme of PSLLT 2019 focused on the many important contributions of corpus linguistics (CL) to the field of pronunciation teaching and learning, including future directions in quantitative and CL-informed analyses. Domains discussed include English-based, cross-cultural workplace spoken interactions in settings such as outsourced call centers (business telephone transactions), pilot-air traffic controller communications (Aviation English radio-telephony), and office interactions with workers who use augmentative and alternative communication (AAC) devices. A model of an iterative research cycle with corpora, which combines computational approaches to data extraction and analyses, and a progression of stages involving quantitative and qualitative, interpretive techniques is discussed. Available tools that could be used in compiling and annotating spoken corpora across various settings and conditions are presented.

INTRODUCTION

I developed this presentation as a (modified) workshop and a demonstration overviewing the process of designing, compiling, and annotating spoken corpora for participants at the Pronunciation in Second Language Learning and Teaching (PSLLT) Conference, September 2019 at Northern Arizona University (NAU), Flagstaff, AZ. The primary theme of the conference focused on the many important contributions of corpus linguistics (CL) to the field of pronunciation teaching and learning, with my mentor, Dr. Douglas Biber, as the plenary speaker, discussing theoretical and practical implications of CL. In particular, I wanted to focus on English-based, cross-cultural workplace spoken interactions in settings such as outsourced call centers (business telephone transactions), pilot-air traffic controller communications (Aviation English radio-telephony), and office interactions with workers who use augmentative and alternative communication (AAC) devices. Although these domains often involve professionals outside traditional classroom settings, innovative pronunciation teaching approaches, especially for speakers communicating in English with English-first language (L1) co-workers or customers, are essential for their required training programs.

In my workshop/demo, I provided a summary of corpus-based approaches to text transcription and annotation, a brief overview of advancements in Multimodal Annotation (e.g., Gu, 2008; the ITACorp Project at Penn State University), and discussed current and future approaches and

directions. Outlined below are the topics of my presentation, with historical perspectives and core issues in the collection and analysis of spoken corpora and my two focal topics for the demonstration: (1) working with specialized, professional spoken corpora, and (2) collecting a corpus representing spoken English in the Philippines.

HISTORICAL PERSPECTIVES AND CORE ISSUES

CL is a methodological approach to the study of language structure, patterns, and use. Exploring corpora has become a popular approach in the quantitative analysis of the linguistic characteristics of written and spoken discourse, resulting in the development of more accurate teaching materials, frequency-based dictionaries, and ESL textbooks, especially for university-level learners of English (Friginal, 2018). Corpora are, in a sense, datasets of systematically collected, naturally-occurring language stored and processed in computer platforms (Friginal & Hardy, 2014). Primarily, corpora are (1) authentic, (2) relatively large, (3) electronic, and (4) conform to specific principled criteria (Bowker & Pearson, 2002; Friginal, Dye, & Nolen, 2019). There are corpora containing a variety of spoken registers large enough to allow a systematic analysis of relevant, target linguistic (especially lexico-syntactic) patterns, and hopefully, specific features that may also capture segmental and suprasegmental characteristics of speech.

Many corpus-based analyses of spoken discourse have provided directions for empirical investigations of linguistic features in various types of formal and informal conversations. Work in this area has incorporated corpus-based methods in analyzing television talk shows, job interviews, and professional interactions. Staples (2015), for example, analyzed the spoken discourse characteristics of patient-provider interactions in healthcare, and much earlier, Rayson, Leech, and Hodges (1997) conducted a corpus-based analysis of speech that is differentiated socially and contextually. Several studies have utilized corpora in describing the lexis and grammar of business interactions. Among these studies is McCarthy and Handford's (2004) work on defining the structure of spoken business English using the Cambridge and Nottingham Corpus of Business English (CANBEC). They explored the different dimensions of business talk in relation to everyday casual conversation. One strength of corpus-based methods is that the quantitative collection and analysis of language allows for linguistic features in use that would otherwise remain hidden or undetected by speaker's perceptions to be found and disclosed. Macro analyses to address groups of people, various demographics, registers, or situational contexts can be conducted to produce a range of numerical data for interpretation and potential application in practical contexts (Friginal & Hardy, 2014).

It is clear that corpus-based methods are still limited when it comes to studying the sociophonetic features of speech. Pronunciation, including such features as intonation, rhythm, pitch, volume, and stress) of words and discourse is complex and difficult to easily program or capture through algorithms. However, there are advancements in the use of computational tools, dictation and transcription software, qualitative coding programs, and automated sentiment analyzers (e.g., those utilized in customer service and social media platforms) that may serve as models for a robust collection of a new generation of specialized spoken corpora especially developed for pronunciation teaching and learning. The annotation of spoken corpora for prosody, for example, the Hong Kong Corpus of Spoken English (HKCSE) (Cheng, Greaves, & Warren, 2008) and more

detailed contextual transcriptions and annotations of spoken texts suggest promising prospects for capturing some phonetic features of speech in orthographic transcripts.

Although not necessarily considered corpora in the traditional sense, available databases of speech that are designed to be analyzed phonetically, phonologically, or acoustically point to a possible framework for developing a phonetically-annotated corpus. For example, the Speech Accent Archive (<http://accent.gmu.edu/>) (Weinberger, 2018), currently with over 2,000 speech samples, is an online database of speakers from around the world reading aloud a short paragraph in English. The audio samples are then transcribed phonetically using the International Phonetic Alphabet (IPA), resulting in a “corpus” of IPA-transcribed texts. By using crowdsourcing techniques, various speakers are also able to submit their own speech patterns (and “accents”) digitally. All of these speakers are asked to read aloud a single paragraph:

Please call Stella. Ask her to bring these things with her from the store: Six spoons of fresh snow peas, five thick slabs of blue cheese, and maybe a snack for her brother Bob. We also need a small plastic snake and a big toy frog for the kids. She can scoop these things into three red bags, and we will go meet her Wednesday at the train station.

This paragraph was designed to elicit many of the possible sounds and sound combinations occurring in English. Although the sample is read and not naturally-occurring, the Speech Accent Archive is an example of what might be a possible methodology in phonetically transcribing a corpus. Every entry in the archive is thus tagged for a speaker’s birthplace, native language, other language known, age, age when first learning English, method of English learning (in school or not), length of time having lived in an English-speaking country (and which country, if that is the case). All of these variables are also searchable on the website. That makes it easy for a teacher, phonetician, speech pathologist, or anyone interested in accents to search for a group of speakers to explore phonetic and phonological processes. Another useful feature of the Speech Accent Archive is that its website allows users to search for audio and transcripts by categories of phonetic characteristics as they differ from General American English (GAE). Phonetic generalizations for the samples can be searched by vowel, consonantal, and syllabic differences from the GAE (Friginal & Hardy, 2014).

The pioneering TIMIT Corpus from 1993 is also a corpus of read speech designed to provide speech data for acoustic-phonetic studies and for the development and evaluation of automatic speech recognition systems. TIMIT contains broadband recordings of 630 speakers of eight major dialects of American English, each reading ten phonetically rich sentences. The TIMIT corpus includes time-aligned orthographic, phonetic and word transcriptions as well as a 16-bit, 16kHz speech waveform file for each utterance (Garofolo et al., 1993). Also related to phonetic analyses of corpora, the C-ORAL-ROM project (Cresti & Moneglia, 2005) was developed to acoustically analyze spoken texts of Italian, French, Spanish, and Portuguese (no English samples yet analyzed using this model). For American English, Clopper and Pisoni (2006) have developed the Nationwide Speech Project corpus which contains nearly 60 hours of recorded speech from 60 informants, 5 males and 5 females from 6 dialect regions in the U.S.: New England, the North, the Mid Atlantic, the South, and the West. What is now becoming increasingly common is the multi-

modal annotation of spoken interactions (see, e.g., Gu, 2002, 2008). Together with enhanced prosodic and acoustic mark-ups of spoken corpora, multi-modal transcripts linking video recordings to non-linguistic features that play a crucial role in communication, such as facial expressions, hand gestures and body position are highlighted and can be automatically extracted. Studies like these indicate that the strengths of corpus analysis can be extended to include aspects of communication and related variables beyond the analysis of the lexico-grammatical fabric of spoken and written texts (Biber, Reppen, & Friginal, 2010; Friginal & Hardy, 2014).

Finally, in the domains of learner (ESL or L2) classrooms, there have been several projects developed by corpus linguists working primarily with learner written and spoken language, e.g., scholars affiliated with the International Corpus of Learner English (ICLE) from the Center for English Corpus Linguistics at the Université Catholique de Louvain. I worked with Joseph Lee, Brittany Polat, and Audrey Roberson (*Exploring Spoken English Learner Language Using Corpora: Learner Talk*, Palgrave Macmillan, 2017) on a corpus-based study of spoken learner language produced by university-level ESL students in the classroom. We utilized contemporary second language acquisition theories as a guide and employed corpus analysis tools and methods to analyze a variety of learner corpora to offer new insights into the nature and characteristics of the spoken language of college ESL learners. In the following sections below, I summarize my two focal discussion topics at PSLLT 2019 and provide a list of recommended resources for recording and annotating spoken corpora for research and teaching purposes.

FOCAL DISCUSSION AND DEMONSTRATION 1: WORKING WITH SPECIALIZED, PROFESSIONAL SPOKEN CORPORA

Figure 1 shows my (proposed) model of an iterative research cycle with corpora which combines computational approaches to data extraction and analyses, and a progression of stages involving quantitative and qualitative, interpretive techniques. The model may support pronunciation-specific studies in conjunction with the use of additional tools such as Praat or ELAN as part of the expansion of research design to create sub-corpora and conduct contextual analysis of spoken English.

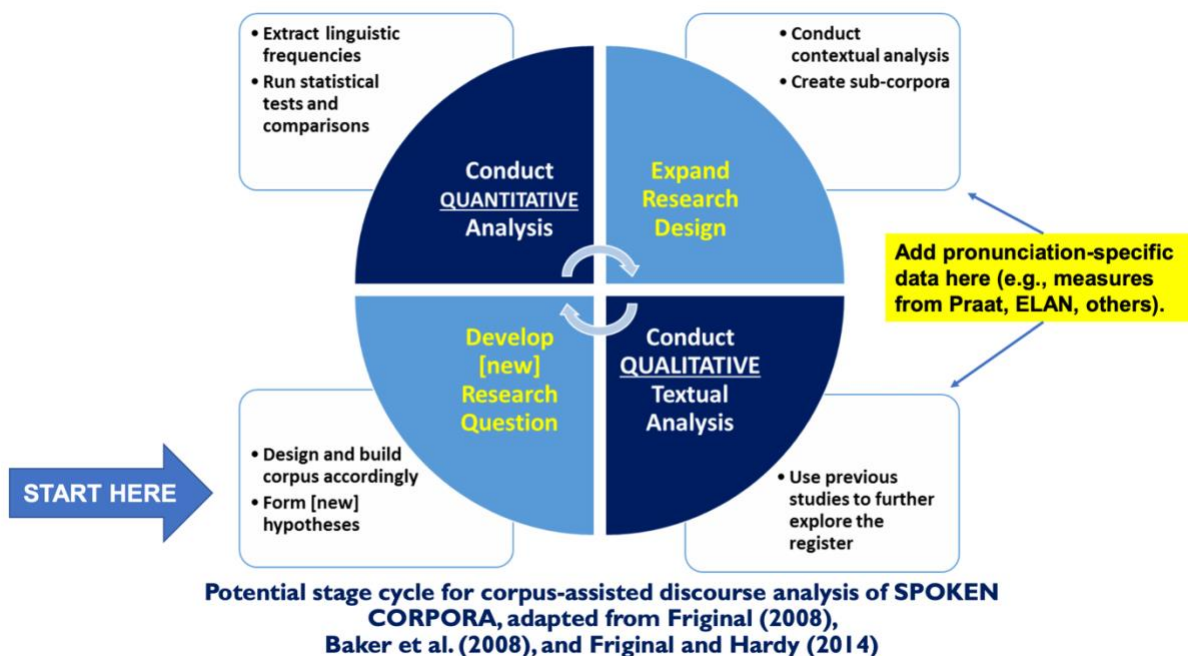


Figure 1. Proposed model for corpus-assisted analysis of spoken corpora

I use this model in establishing an overarching, corpus-based research question, followed by very specific sub-questions that are intended to analyze micro-linguistic features from the list of communicative domains and spoken registers that I have been studying over the past several years:

- Outsourced Call Center Industry
- Global Aviation
- International Maritime Industry
- Multicultural Workplaces in the U.S.
- Hotel and Customer Service Industry
- Augmentative/Alternative Communication (AAC) in the Workplace
- Spoken English in U.S. Academia [ITAs, Foreign-Born Professors-Students Interactions]

Below are short descriptions of selected corpora I have developed (or co-designed, in the case of ANAWC, with Lucy Pickering and Carrie Bruce) of professional interactions, often involving multi-lingual speakers.

ANAWC: The AAC and Non-AAC Workplace Corpus

The Augmentative and Alternative Communication (AAC) and Non- Augmentative and Alternative Communication (Non-AAC) Workplace corpus was collected in 2009 as a collaborative project between Georgia State University and Georgia Institute of Technology (Pickering & Bruce, 2009). The ANAWC was conceived as a specialized corpus focused on the workplace experiences of AAC device users in comparison to their non- AAC using counterparts

in similar working environments. It comprises over 200 hours of spoken interaction (approximately one million words) involving eight focal participants and more than 100 interlocutors in seven different workplace locations. Focal participants comprised four AAC users and four non-AAC users who were provided with an audio-recorder and asked to record their workplace interactions. AAC devices are used by people with complex communication needs who have some form of dysarthria (difficult or unclear articulation) or who are no longer able to speak due to developmental or acquired disorders such as cerebral palsy or motor neurone disease (also known as Amyotrophic Lateral Sclerosis or ALS). The devices are usually portable speech-generating technologies housed in laptops, tablets or smartphones that enable the user to create messages by selecting pictures, letters, words, or sentences and can be accessed using a variety of methods such as keyboarding, eye gaze, or switch input. A detailed discussion of ANAWC's corpus design, composition, and sample exploratory analysis can be found in Pickering et al. (2019).

Aviation Corpus: Corpus of Pilot and Air-Traffic Controller Communication (CORPAC)

My recent book, co-authored with Elizabeth Mathews and Jennifer Roberts of Embry-Riddle Aeronautical University, *English in Global Aviation: Context, Research, and Pedagogy* (Bloomsbury, 2019) explores major issues involved in the use of English in the global aviation industry, turning research into practice in the field of English for Specific Purposes (ESP), specifically Aviation English. With an impetus on evidence-based practice, we discussed the critical role of English in aviation in a variety of contexts, including the national and global policies impacting training and language assessment for pilots, air-traffic controllers, ground staff, and students. The CORPAC, collected in collaboration with Aline Pacheco and Joao Cavallet from the Pontifical Catholic University of Rio Grande do Sul (PUCRS), Brazil features original, authentic audio communication which sampled actual language used by pilots and air-traffic controllers intended for materials and resources production. A majority of texts from CORPAC were obtained from YouTube channels <VASAviation> and LiveATC and recorded training/flight sessions provided by various institutions and airlines. CORPAC is a monitor corpus, currently and relatively small but increasing in size, with 100,000+ words across 150+ routine and emergency situations; coded and minimally annotated, as shown by the short transcript below.

-ATC-Pilot Transcripts -

ACA759 -Tower, good evening. Air Canada seven-five-nine with you uh-On the Bridge Visual two-eight right.

SFO TWR -Air Canada seven-five-nine, San Francisco Tower. Runway two-eight right, cleared to land. Wind is two-seven-zero at eight.

ACA759 -Cleared to land on two-eight right. Air Canada seven-five-nine. (...) And uh-Tower, just wanna confirm - it's Air Canada seven-five-nine -we see some lights on the runway there, across the runway. Can you confirm we're cleared to land?

SFO TWR -Air Canada seven-five-nine, confirmed. Cleared to land, runway two-eight right. There is no one on two-eight right but you.

ACA759 -OK. Air Canada seven-five-nine.

UAL1 -Where's this guy going? (...) He's on the taxiway!

SFO TWR -Air Canada, go around.

ACA759 -In the go-around. Air Canada seven-five-nine.

SFO TWR -Seven-five-nine, it looks like you were lined up for Charlie there. Uh-Fly heading two-eight-zero, climb and maintain three thousand.

ACA759 -Heading two-eight-zero, three thousand. Air Canada seven-five-nine.

UAL1 -Uh-United one. Air Canada flew directly over us.

SFO TWR -Yeah. I saw that, guys. (...) Seven-five-nine, contact NorCal one-three-five point one. Will catch you in a couple minutes.
 ACA759 -Thirty-five decimal one. Air Canada seven-five-nine.
 SFO TWR -United 1, we're gonna get you going here in just-.
 UAL1 -We're ready!
 -End of transcript –

Call Center Corpus

The influx of outsourced call centers from the U.S. to the Philippines, India, and other countries since the late 1990s has generated employment opportunities for English-speaking professionals who are able to communicate in English and provide telephone-based customer services to American callers. The Philippines has become one of the major centers for U.S.-based outsourcing because of its tradition of English education, affinity to the American culture, and overall cheap labor market (Friginal & Cullom, 2014). The \$2 billion-a-year call center industry in the Philippines employs more than 150,000 individuals, and the Philippine government continues to invite U.S. companies to relocate their business process operations into the country's major cities by providing tax incentives, improving technology architecture, and focusing on the marketability of its human resources (Friginal & Friginal, forthcoming).

My current Call Center Corpus was collected in the Philippines over a 10 year period from 2006 to 2016 and provided by a U.S.-owned call center company, which sponsored the corpus collection and transcription. The transactions were retrieved following the list of audio files cued in the database of recorded calls for a particular work shift. The calls that qualified in the corpus ranged from five to 25 minutes in duration. Convenience sampling of audio files was done to ensure a comparable number of files per transaction type (e.g., troubleshooting vs. customer service) and achieve a balanced number of male and female call-takers (or “agents”) and callers. The calls were transcribed into machine readable text files by trained transcriptionists following conventions used in the collection of the service encounter corpus of T2K-SWAL (TOEFL 2000 Spoken and Written Academic Language; see Biber, 2006, for a description of this corpus). Personal information about the callers, if any (e.g., names, addresses, phone numbers, credit card or social security numbers, etc.) was consistently and scrupulously replaced by different proper nouns or a series of numbers in the transcripts. No attempt was made to transcribe phonetically, but some comments about pronunciation, whenever they resulted in misunderstanding were added in the texts. The transcribed texts were manually checked for format and accuracy.

The corpus design and collection of the Call Center corpus focused on the following goals and themes (Friginal, 2009):

Overarching (ambitious) Goal: To compile a representative corpus of outsourced and U.S.-based call center interactions

Agent locations: Primarily Manila, Philippines (I have also collected transactions from Bangalore, India; San Jose, Costa Rica; Florida and Georgia in the U.S.)

Callers: All U.S.-based callers (it is also possible to also “code” caller demographics)

Communicative Tasks –Inbound: Troubleshooting, customer service/inquiry, product purchase or reservation

Communicative Tasks –Outbound: Telemarketing, surveys

Task Difficulty: Neutral –Difficult –Problematic (intense, irate caller)

Agent Performance Ratings: Low-mid-high (based on a monthly quality assurance assessment report)

Agent Demographics: Gender, age, length of service or experience, level of education, degree, school graduated from (in the Philippines), others

Auto-Transcription: Auto-Speech-to-Text (internal to the sponsoring call center company)

Manual Transcription: Full verbatim transcription from the auto-generated transcripts

Other considerations:

SOME TRANSCRIPTION CONVENTIONS: (added through Nvivo) Capturing holds and pauses (length in seconds/minutes); annotating interruptions, overlaps, repeats; coding dysfluencies

RUNNING SELECTED TEXTS in PRAAT <<http://www.fon.hum.uva.nl/praat/>>

POS-TAGGING: Biber Tagger and others (e.g., LIWC, Coh-Metrix)

EVALUATION SCORES PER AGENT, PER CALL: Quality assurance, language and performance scores [Language: includes an assessment of segmental and suprasegmental pronunciation. See Friginal (2013) for a detailed discussion of assessment and evaluation and testing instrument used.]

LEGAL and INSTITUTIONAL REVIEW BOARD CHALLENGES and CONSIDERATIONS (important issues for corpus collection not discussed in PSLLT 2019).

Below is a sample transcript and annotation of the Call Center Corpus segmenting the utterances of agents and callers.

Agent Utterances	Annotation
01: Thank you for calling [company name] global services this is [agent name] how may I help you? 02: ok [caller name] uh I'll I'll be putting you on break for a minute while I check on that store number. Would that be ok? 03: [hold 25 seconds] and [caller name] may I ask what store is this 7555? 04: a branch? 05: ok uh [caller name] uh do you happen to have any ticket number uh previous ticket number for this store so I can check on the entry 06: uh I only need the previous ticket number so that I can verify if this is under [unclear] because it might be uh [inter] 07: ok alright I'll put you on hold again, I'll be checking this information from the customer store. 08: ok please hold [hold 2 minutes 30 seconds] ok sorry for the long wait I'm still on the process of checking some information here and I'll be putting you on hold again 09: please hold [hold 2:38 seconds] ok [caller name]	[Notations here included references to the agent's task performance scores; some L2 pronunciation issues, when potentially influencing communication may be provided; other performance-related concerns were added here during manual transcription and annotation.]

<p>10: what I'll be doing is I'll be quoting you a ticket and before that I'll investigate or check some information for this store number 75555 and as soon as I have some information or status I'll be calling you back for for more. Would that be ok? 11: ok 12: call us back ok 13: bye 14: ok bye</p>	
--	--

Caller Utterances	Annotation
<p>01: hello? 02: [caller name] 03: [caller spelling last name] 04: excuse me? [*] 05: no you with [company name]? 06: right ok good 07: yes garbage disposal 08: uh I have a unit e505 09: and it's only a couple of years old and I'm having trouble with it leaking from underneath 10: like it leaks through where the reset button is and what not 11: yeah the serial number is [sh pause] where will that be at? 12: uh ok vh 73516257 13: it's been leaking almost from day one. Sometimes it goes away sometimes I mean I can see that the reset button is about completely rusted at the top of the screws there's something that's inside there's also another little panel on it I'm not sure what it's for but it got rusted by itself 14: uh there's no rust or water from there 15: there's no spout leaking from the outside of the unit either that will lead me to believe that it drips from the side from somewhere and then leaks you know what I mean and it settles underneath 16: yeah underneath the part of the unit there's something seal inside the unit that's not, I mean it's very very slight because it's done this probably, it's really done this from day one now I'm not sure if there's a defect from this unit, because I actually switched the unit from where I first got it. 17: but that was such a hassle for this thing, I just put a little light paste on it it went away it doesn't leak constantly that's the weirdest thing about it but now I have to clean out the hole underneath the sink [lng pause] 18: uh probably not 19: yeah</p>	<p>[Annotations for caller utterances may include markers of customer satisfaction or frustration, clarification sequences, nature of inquiry or questions, and various references to sentiment and semantic markers.]</p>

FOCAL DISCUSSION AND DEMONSTRATION 2: COLLECTING A CORPUS REPRESENTATING SPOKEN ENGLISH IN THE PHILIPPINES (SPOKEN PHILIPPINE ENGLISH OR SPEC)

During my presentation, I also shared my experiences in leading a project that aimed to collect a corpus of spoken professional English in the Philippines. The Spoken Philippine English or SPEC is an on-going project, focusing on the guide questions and considerations below:

1. Define and operationalize “naturally-occurring spoken Philippine English.”
2. What to collect? (registers, contexts, speech events)
3. Who are target speakers?
4. What are the linguistic characteristics of Filipino speakers of Philippine English?
5. What levels of spoken English abilities should be represented?
6. DEVELOP an INCLUSION and EXCLUSION criteria.
7. What models are available? What general or monitor English corpora (e.g., BNC, ANC, COCA, others) could we follow?

Models or templates that were considered by the research team included the International Corpus of English or ICE and the Hong Kong Corpus of Spoken English with their texts/registers shown below.

International corpus of English (ICE) Philippines

Spoken Texts (300 texts, 2,000-word samples) in ICE

Dialogues (180 texts): Spontaneous conversations (90), Telephone conversations (10), Class lessons (20), Broadcast discussions (20), Broadcast interviews (10), Political debates (10), Legal cross-examinations (10), Business transactions (10).

Monologues (120 texts): Spontaneous commentaries (20), Unscripted speeches (30), Demonstrations (10), Legal presentations (10), Broadcast news (20), Broadcast talks (20), Scripted speeches (10).

Hong Kong Corpus of Spoken English

Sub-Registers of the HKCSE: Academic Sub-Corpus, Business Sub-Corpus, Conversation Sub-Corpus, Public Sub-Corpus



**RC Research Centre for
PCE Professional Communication in English**

THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

Department of
ENGLISH
英文學系

The Hong Kong Corpus of Spoken English

The Hong Kong Corpus of Spoken English is comprised of four sub-corpora (academic, business, conversation and public).

- 1. Composition of academic sub-corpus**
 - Student presentation and Q&A
 - Lecture
 - Seminar and tutorial
 - Writing Assistance Programme (WAP) consultation
 - Workshop for academic and research staff and postgraduate research students
- 2. Composition of business sub-corpus**
 - Job and placement interviews
 - Presentations and Q&A sessions
 - Meetings
 - Informal office talk
 - Announcements and Q&A sessions
 - Presentations
 - Service encounters
 - Conference calls/ video conferencing
 - Workplace telephone talk
- 3. Composition of conversation sub-corpus**
 - Conversation collected in a wide range of social settings
- 4. Composition of public sub-corpus**
 - Speeches
 - Interviews
 - Speeches and Q&A sessions
 - Press briefings and Q&A sessions
 - Discussion forums
 - Press briefings
 - Radio announcements

Figure 2. Composition of the HKCSE (obtained from <http://rcpce.engl.polyu.edu.hk/HKCSE/default.htm>)

The corpus design or (intended) composition of SPEC is shown in Table 1, with most of academic and media registers already collected. Business and Government texts are being compiled, with some sub-registers requiring legal and institutional reviews and permissions.

Table 1


Composition of Spoken Philippine English (Various Levels: ACADEMIC – Public/Private Universities in Manila (2); Public/Private Provincial Universities)

ACADEMIC	BUSINESS	GOVERN- MENT	MEDIA	SPE MODEL
Student Presentations	Job Interviews	Congressional Hearings	TV/RADIO interviews	Recorded interviews with 50+ NNES
Lectures	Office Meetings	Presidential Speeches	Newscast	models in teaching
Student Monologues, Extemporaneous Speeches	On-the-Job: Call Centers, Pilots, Hotel Staff, Othes	Legal Proceedings	TV English-Based Programs	ESL pronunciation
Full Classroom Recordings	Tele-conferences	Various Provincial or City-Based Events	TV/Radio Announcements	50+ speakers grouped according to
Student Interviews	Presentations	Ceremonies	Sports Broadcast	a few criteria

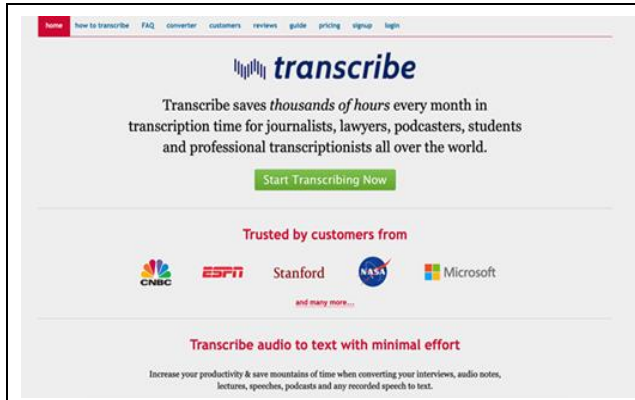
RECOMMENDED RESOURCES

Below are available tools that could be used in compiling and annotating spoken corpora across various settings and conditions.

Smart Phone Apps for Audio Recording and Transcription

	<p>Smart phones and tablets, especially iPhone, iPad, and Android devices allow for easy access to various applications (or <i>apps</i>) that record high-quality audio in the form of MP3, MP4, voice memos, and other formats. Many of these recorded files could be saved and downloaded into PCs and laptops or cloud storage services. After audio recording, supplemental apps for transcription (e.g., Speech to Text or Voice to Text shown here) may be used in building your spoken corpus. Most of these automated transcriptions may require manual checking for accuracy (e.g., of spelling and some content). These apps may range from \$10 to \$75.</p>
---	---

Recording and Transcription

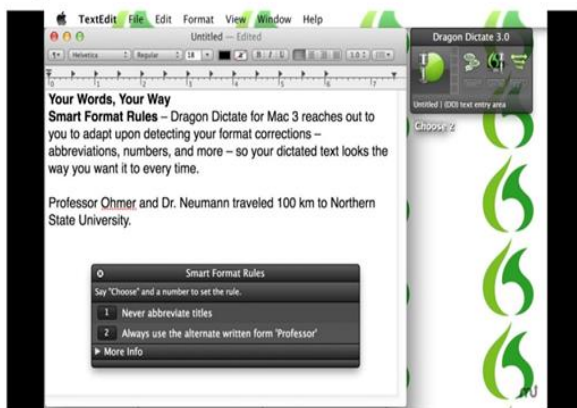


Transcribe is an app that allows users to convert audio/video to text formats, supporting automatic transcription, dictation and “self-transcription.” From the makers of the app: “Transcribe makes manual transcription faster and less painful with dictation and a tightly integrated editor and player. We still offer this workflow (called self-transcription) if your audio has background noise or has frequent interruptions.”



RecUp is a universal app, running on iPhone/iPod Touch and iPad, recording high-quality MP3 audio files that allow users to save bandwidth and interface with a Dropbox storage. The app is accessible with VoiceOver for the visually impaired.

<http://www.irradiatedsoftware.com/recup/>

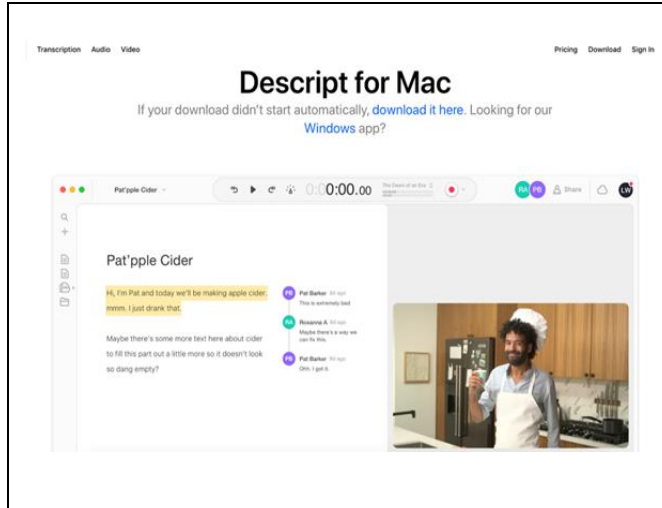


Dragon Speech Recognition software.

From the developers: Dragon is fast, accurate speech recognition, dictation and transcription software. Dragon by Nuance is the world’s leading speech recognition solution with over two decades of continuous development to meet the needs of the most businesses and individual researchers.

<https://shop.nuance.com>

Audio (and video) Transcription Tool and Text Editor

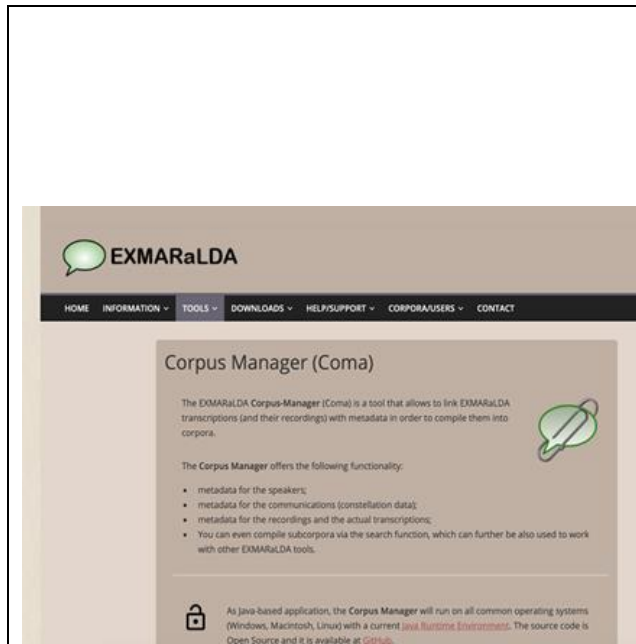


Descript is primarily a podcasting application which could be used in corpus collection for recording and transcription.

From the developers: Descript is an industry leading transcription service which partners with the most accurate transcription providers to make sure users receive accurate transcripts and multitrack recording, dynamically generating a single combined transcript for multiple audio/video files.

<https://www.descript.com>

Managing and Annotating Spoken Corpora



From the developers: “EXMARaLDA is a system for working with oral corpora on a computer. It consists of a transcription and annotation tool (Partitur-Editor), a tool for managing corpora (Corpus-Manager) and a query and analysis tool (EXAKT).

EXMARaLDA’s features include, for instance:

- time-aligned transcription of digital audio or video
- flexible annotation for freely choosable categories
- systematic documentation of a corpus through metadata
- flexible output of transcription data in various formats
- computer-assisted querying of transcription and metadata
- interoperable as it works XML based data formats that allow for data exchange with other tools (like Praat, ELAN, Transcriber etc.) and enable a flexible processing and sustainable usage of the data.

<https://exmaralda.org/en/>

Windows	Mac	Unix
CLANWin	CLAN	UnixCLAN
V 27-Aug-2019 11:00	V 27-Aug-2019 11:00	V 27-Aug-2019 11:00

For Windows:
CLANWin works with Windows 7, 8.x, and 10. Windows installation involves clicking on the installation file and following the directions given by InstallShield. If you have an older version of CLAN on your machine, InstallShield will overwrite it with the newer version. By default, CLAN uses the Arial Unicode MS font, if it is available. If not, it will use the Arial font.

For Macintosh:
CLAN is for Mac OS X users from 10.6 and up.
By default, CLAN uses the Arial Unicode MS font.

Unix Installation:
For Unix users, we are distributing the source code for CLAN. This code is distributed under the terms of the [GNU General Public License](#). UnixCLAN only provides the analysis commands of CLAN in the Unix environment. We have not yet constructed a Unix version of the CLAN editor.

A Brief Description:
The CLAN Programs are downloaded, installed, and used as a single application. Functionally, however, CLAN has two parts. The first part is the CLAN editor which can be used to edit files in either CHAT or CA (Conversation Analysis) format. The editor also provides a wide range of additional functions, such as audio and video playback, linkage to audio and video, fonts for Roman and non-Roman orthographies, data validation, adding codes to files, and shipping data to other programs. The second part of CLAN is the set of data analysis programs. These programs are run from a separate window called the Commands window. The results of the analytic programs are sent to the CLAN Output window.

TalkBank and CLAN
From the developers: The CLAN Programs are downloaded, installed, and used as a single application in two parts. The first part is the CLAN editor which can be used to edit files in either CHAT or CA (Conversation Analysis) format. The editor also provides a wide range of additional functions, such as audio and video playback, linkage to audio and video, fonts for Roman and non-Roman orthographies, data validation, adding codes to files, and shipping data to other programs. The second part of CLAN is the set of data analysis programs.

<http://dali.talkbank.org/clan/>

What is SALT?
Systematic Analysis of Language Transcripts (SALT) is software that standardizes the process of eliciting, transcribing, and analyzing language samples. It includes a transcription editor, **standard reports**, and **reference databases** for comparison with typical peers. [View brochures](#) [What's new?](#)

- Software for Windows and Mac
- Elicitation Materials
- Transcription Services
- Free Online Training
- Resources for Instructors
- New Zealand and Australia

[View case study examples](#)
New Performance Report makes report writing easier

Systematic Analysis of Language Transcripts (SALT)
From the developers: SALT is a software that standardizes the process of eliciting, transcribing, and analyzing language samples. It includes a transcription editor, standard reports, and reference databases for comparison with typical peers.

<https://www.saltsoftware.com>

PRAALINE
A sweeter way of analysing speech corpora!

[FEATURES](#) [DOWNLOAD](#)

Praaline is a system for managing, annotating, visualising and analysing spoken language corpora. It is a sweeter way of doing corpus linguistics and speech analysis!

PRAALINE is a system for managing, annotating, visualizing and analyzing spoken language corpora. From the developer: It is a sweeter way of doing corpus linguistics and speech analysis!

<https://www.praaline.org>

ABOUT THE AUTHOR

Eric Friginal is professor of applied linguistics at the Department of Applied Linguistics and ESL and Director of International Programs at the College of Arts and Sciences, Georgia State University. He is the founding co-editor-in-chief of Applied Corpus Linguistics (ACORP) Journal published by Elsevier (with Paul Thompson, University of Birmingham).

REFERENCES

- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., Reppen, R., & Friginal, E. (2010). Research in corpus linguistics. In R.B. Kaplan (Ed.), *The Oxford handbook of applied linguistics* (2nd ed., pp. 548–570). Oxford: Oxford University Press.
- Bowker, L., & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora*. New York: Routledge.
- Cheng, W., Greaves, C., & Warren, M. (2008). *A corpus-driven study of discourse intonation*. Amsterdam: John Benjamins.
- Clopper, C.G. & Pisoni, D.B. (2006). The Nationwide Speech Project: A new corpus of American English dialects. *Speech Communication*, 48(6), 633–644.
- Cresti, E., & Moneglia, M. (2005). *C-ORAL-ROM: Integrated reference corpora for spoken Romance languages*. Amsterdam: John Benjamins.
- Friginal, E. (2009). *The language of outsourced call centers: A corpus-based study of cross-cultural interaction*. Amsterdam: John Benjamins.
- Friginal, E. (2013). Assessment of oral performance in outsourced call centers. *English for Specific Purposes*, 32(2), 25-35.
- Friginal, E. (2018). *Corpus linguistics for English teachers*. New York: Routledge.
- Friginal, E. & Cullom, M. (2014). Saying ‘no’ in outsourced call center communications. *Asian Englishes Journal*, 4, 23-37.
- Friginal, E., Dye, P., and Nolen, M. (2019). Corpus linguistics in TESOL: Doing what works. *TESOL AL Forum*. Available at <http://newsmanager.commpartners.com/tesolalis/issues/2019-08-26/2.html>
- Friginal, E. & Friginal, R. (forthcoming). Philippine English in outsourced call centers: A corpus-based comparison. In A. Borlongan (Ed.), *Philippine English: Development, structure, and sociology of English in the Philippines*. Singapore: Routledge Asia-Pacific.
- Friginal, E. & Hardy, J.A. (2014). *Corpus-based sociolinguistics: A guide for students*. New York: Routledge.
- Garofolo, J., Lamel, L., Fisher, W., Fiscus, J., Pallet, D., Dahlgren, L., & Zue, V. (1993). *TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1*. Web Download. Philadelphia: Linguistic Data Consortium.
- Gu, Y. (2002). Towards an understanding of workplace discourse: A pilot study for compiling a spoken Chinese corpus of situated discourse. In C.N. Candlin (Ed.), *Research and practice in professional discourse* (pp. 137–186). Hong Kong: City University of Hong Kong Press.

- Gu, Y. (2008). Segmenting and annotating a multimodal corpus (with special reference to SCCSD). Keynote speech at the 3rd International Corpus Linguistics Conference, University of Birmingham, UK.
- McCarthy, M. & Handford, M. (2004). Invisible to us: A preliminary corpus-based study of spoken business English. In U. Connor, & T.A. Upton (Eds.), *Discourse in the professions: Perspectives from corpus linguistics* (pp. 167–201). Amsterdam: John Benjamins.
- Pickering, L. & Bruce, C. (2009). *AAC and Non-AAC Workplace Corpus (ANAWC)*. Atlanta, GA: Georgia State University.
- Pickering, L., Di Ferrante, L., Bruce, C., Friginal, E., Pearson, P., & Bouchard, J. (2019). An introduction to the ANAWC: The AAC and Non-AAC Workplace Corpus. *International Journal of Corpus Linguistics*, 24(2), 229-244.
- Rayson, P., Leech, G., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: Some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2, 133–150.
- Staples, S. (2015). *The Discourse of nurse-patient interactions: Contrasting the communicative styles of U.S. and international nurses*. Amsterdam: John Benjamins.
- Weinberger, S.H. (2013). *Speech Accent Archive*. Fairfax, VA: George Mason University. Available from <http://accent.gmu.edu>