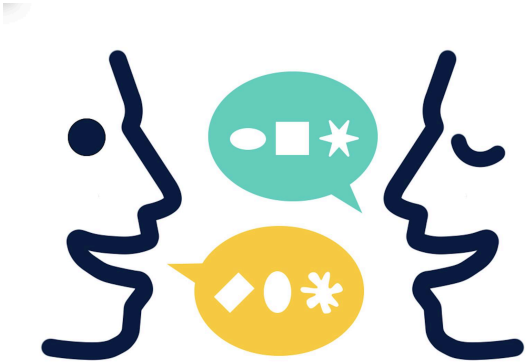

The Corpus of Mobile Speakers (CorMS)

User Guide

Editor: Jennifer Nycz

June 25, 2024

GEORGETOWN UNIVERSITY
College of Arts & Sciences
Department of Linguistics

Table of Contents

1. Introduction

This manual provides an overview of the data currently included in the Corpora of Mobile Speakers (CorMS),details about how the data in each corpus was collected and processed, and information about how to use the data.

The founding datasets CorMS were collected through the the *Second Dialect Acquisition and Stylistic Variation in Mobile Speakers* project (supported by NSF-BCS award #1651108). The goal of this project was to create a database of speech produced by mobile speakers of North American English conversing freely with an interviewer. Interviews with 40 native speakers from Toronto, Canada who moved as adults to New York City, United States, and 30 native speakers from New York City who moved to Toronto were conducted in 2018-2020. Of these, 31 participants in New York City and 28 participants in Toronto gave consent to have their data shared with other researchers. The audio files of these interviews along with their time-aligned orthographic transcripts in ELAN and Praat formats are now available through CoRMS.

We hope to expand CorMS to include recordings and transcripts representing speakers (and signers!) of other languages and dialects with different mobility histories. If you are a researcher interested in linking your own mobile speaker data to CorMs, please contact Dr Jen Nycz.

Section 2 of this manual describes aspects of the data collection along with characteristics of the speakers in each corpus. Sections 3 and 4 describe the conventions followed during the transcription process and redaction of identifying information, respectively. Section 5 gives information about the files formats used in the corpus and the terms of use.

Why Assemble Corpora of Mobile Speakers?

When someone moves to a region where a different language is spoken (from New York to Nice, for example, or from Berlin to Bangkok), the influence is often clear: that person learns a new language, or at least enough of the new language to get by. A vast literature on Second Language Acquisition (SLA) examines how people take up new languages as well as the linguistic and social details of this process (Ortega 2009).

What about when someone moves to a place where they speak the same *language*, but a different ***dialect*** of that language – from New York to Toronto, or from Montreal to Paris? A much smaller body of research on Second Dialect Acquisition (SDA) suggests that people do adopt some (though not all) features of that new dialect, depending on a complex set of development, linguistic, and social-attitudinal factors (see Siegel 2010 and Nycz 2015 for further discussion). A significant obstacle to the study of SDA is the relative lack of large, accessible corpora of speech data from mobile speakers who share a movement history: any researcher interested in

this topic must first undergo the time- and resource-intensive process of recruiting their own participants, conducting their own data collection and processing that data to some extent before any analysis can take place. Those who have done the fieldwork and amassed the recordings may not be able to carry out every possible analysis of that data, limited by time, interest, or their own expertise; students or early-career researchers with an SDA research question but (as yet) no data of their own may be stymied before they can even start.

The long-term goal of CorMS is to advance the study of SDA by hosting a repository of high-quality, shareable corpora of data collected from people who have had contact with other dialects as a result of mobility. Scholars who share their recordings and transcripts will increase the (citable!) impact of their fieldwork labor, while those who do not have the time or resources to devote to extensive fieldwork (graduate students with a semester to write a qualifying paper, for example) can still examine questions about SDA.

Nycz, Jennifer. 2015. Second Dialect Acquisition: A Sociophonetic Perspective. *Language and Linguistics Compass* 9/11: 469–482.

Ortega, Lourdes. 2009. *Understanding Second Language Acquisition*. Routledge.

Siegel, Jeff. 2010. *Second Dialect Acquisition*. Cambridge: Cambridge University Press.

Acknowledgements

Many thanks to Abby Killam for assistance in drafting this user guide, and to Tyler Kendall and Amir Zeldes for guidance and advice on corpus administration.

2. The Corpora

The Torontonians-in-NYC Corpus (TO-in-NYC)

Author: Jennifer Nycz
Release date: June 2024 (v.2024.1)
Interviewer: Susan Cohen
Transcribers: Natalie Bazata, Amelia Becker, Nicole Rybak
Redaction assistance: Matthew Dearstyne, Ping Hei Yeung

Participants

TO-in-NYC consists of 31[1] primary speakers (18 women, 13 men) across 34 audio files, collected as part of the *Second Dialect Acquisition and Stylistic Variation in Mobile Speakers* project (NSF-BCS-1651108). The speakers were recorded between February and December 2018. All participants had been living in New York City for at least 4 years at time of interview, though time in that city as well as the age at which speakers moved there varied across the sample; the mean age at time of interview is 40.5 (sd=13.6), the mean age at which speakers moved to New York City is 24.5 years (sd=5.1), and the mean years in New York City is 16.0 (sd=12.2). Metadata for TO-in-NYC includes broad information about speakers' demographic backgrounds (Table 1). No restrictions on social class or other characteristics were placed, other than those implied by the inclusion of participants who are able to move internationally. The speakers reported no speech or hearing disorders.

Metadata details

Speaker Code: each participant is assigned a unique code which denotes their corpus, gender, age at interview, and age moved to New York City, and ends with an two-letter code assigned by the fieldworker (typically the first two letters of the participant's chosen pseudonym).
Corpus: indicates that the speaker's data is part of the TO-in-NYC corpus
Gender: each participant was asked to report their gender early in the interview while filling out a short demographic questionnaire; all participants identified as either male/man (M) or female/woman (F).
Ethnicity: participants were also asked to describe their ethnicity on the questionnaire; this column reflects the wording that each person used
Age at Interview: the age of the participant when they took part in the interview
Date of Interview: when data collection took place
Age Moved to NYC: the age of the participant when they moved to New York City
Years in NYC: the number of years the participant has been living in New York City

[1] Data was collected from 40 speakers total; only 31 primary participants gave unambiguous consent for their data to be shared with other researchers.

Friend present: "Yes" if a local friend of the participant was present for the interview, "No" if the interview was one-on-one. If the participant's friend did not also give clear consent to have their data shared, their contributions were redacted from the audio and transcript.

Table 1. Speakers in the TO-in-NYC corpus.

Speaker Code	Corpus	Gender	Ethnicity	Age at Interview	Date of Interview	Age Moved to NYC	Years in NYC	Friend present
TO-in-NYC_F_26_18_NA	TO-in-NYC	F	white	26	9/4/2018	18	8	Yes
TO-in-NYC_F_27_21_NA	TO-in-NYC	F	white	27	5/14/2018	21	6	Yes
TO-in-NYC_F_28_20_AL	TO-in-NYC	F	white	28	5/24/2018	20	8	No
TO-in-NYC_F_29_17_VA	TO-in-NYC	F	Caucasian	29	12/11/2018	17	12	No
TO-in-NYC_F_29_19_SH	TO-in-NYC	F	Jewish	29	10/30/2018	19	10	No
TO-in-NYC_F_30_23_NE	TO-in-NYC	F	Mediterranean	30	5/22/2018	23	7	No
TO-in-NYC_F_31_27_AL	TO-in-NYC	F	white	31	5/10/2018	27	4	No
TO-in-NYC_F_33_24_RA	TO-in-NYC	F	Caucasian	33	5/1/2018	24	9	No
TO-in-NYC_F_35_24_SA	TO-in-NYC	F	Jewish	35	3/26/2018	24	11	Yes
TO-in-NYC_F_36_26_AT	TO-in-NYC	F	Canadian	36	7/5/2018	26	10	No
TO-in-NYC_F_39_26_EL	TO-in-NYC	F	Black (Afro-Carribean)	39	3/14/2018	26	13	Yes
TO-in-NYC_F_39_27_KA	TO-in-NYC	F	Greek	39	5/25/2018	27	12	No
TO-in-NYC_F_43_39_RA	TO-in-NYC	F	white	43	4/2/2018	39	4	No
TO-in-NYC_F_46_26_CD	TO-in-NYC	F	Croatian	46	6/15/2018	26	20	No
TO-in-NYC_F_57_30_DO	TO-in-NYC	F	Caucasian	57	3/15/2018	30	27	No
TO-in-NYC_F_62_20_VL	TO-in-NYC	F	Jewish	62	10/9/2018	20	42	No
TO-in-NYC_F_62_23_FA	TO-in-NYC	F	Canadian	62	3/6/2018	23	39	No
TO-in-NYC_F_69_30_RE	TO-in-NYC	F	Caucasian	69	10/17/2018	30	39	No
TO-in-NYC_M_26_18_DA	TO-in-NYC	M	Caucasian	26	6/13/2018	18	8	No
TO-in-NYC_M_28_18_IK	TO-in-NYC	M	Asian	28	3/11/2018	18	10	No
TO-in-NYC_M_29_25_AM	TO-in-NYC	M	Indian	29	12/7/2018	25	4	No
TO-in-NYC_M_30_23_FR	TO-in-NYC	M	white	30	5/12/2018	23	7	No
TO-in-NYC_M_33_20_JO	TO-in-NYC	M	(none entered)	33	2/22/2018	20	13	Yes, friend data redacted
TO-in-NYC_M_34_24_MW	TO-in-NYC	M	white	34	8/16/2018	24	10	No
TO-in-NYC_M_35_20_TR	TO-in-NYC	M	white	35	3/5/2018	20	15	No
TO-in-NYC_M_40_32_EL	TO-in-NYC	M	white	40	8/23/2018	32	8	No
TO-in-NYC_M_41_31_CO	TO-in-NYC	M	eastern european/middle eastern	41	3/28/2018	31	10	No

TO-in-NYC_M_49_30_GR	TO-in-NYC	M	white/caucasian	49	4/19/2018	30	19	No
TO-in-NYC_M_62_22_NO	TO-in-NYC	M	white	62	4/4/2018	22	40	No
TO-in-NYC_M_62_27_AB	TO-in-NYC	M	Caucasian/Jewish	62	7/18/2018	27	35	No
TO-in-NYC_M_66_31_ED	TO-in-NYC	M	Caucasian	66	8/15/2018	31	35	Yes

Data collection and recording

Participants were recruited through online fora such as expat community boards and social media (e.g., Facebook, Twitter). Participants took part in an approximately 60-minute long conversational interview with a local consultant which focused on growing up in Toronto, moving to their new country and city, and impressions of the people and culture there. In six cases, a friend or family member of the participant who is native to the new city (i.e. was born in and lived there until at least age 18, and has spent most or all of their subsequent adult life in that region) also participated in the conversation.

After the conversational portion of the interview, participants completed word list readings and minimal pair tasks. In the minimal pair task, word pairs distinguished by low back vowels (e.g. *cot* and *caught*; *dawn* and *don*) were presented to participants via an iPad, with other potential pairs (e.g. *Mary/merry, pin/pe*n) interspersed to pull focus away from the contrasts of low back vowels. Interviewers instructed participants to read each pair out loud and then say whether the two words sounded the same or different to them.

All interviews were recorded to 44.1kHz/16-bit wav files using a Zoom H4N solid state recorder and an Audio-Technica AT831b lavalier microphone. Interviews took place in the participant's home or a quiet public space (such as in a meeting room in the New York Public Library). Consent was obtained and participants received an Amazon gift card worth $20 in US dollars upon completion of the interview.

Files

Each audio file is stored in wav format, and is accompanied by an orthographic transcription in both .eaf (ELAN) and .TextGrid (Praat) format with the same filename stem. Transcripts are divided into breath units (roughly, units of speech occurring between pauses or breaths), with the transcription of each breath unit time-aligned to the audio file. Files are stored individually in a Georgetown Box folder, and can be downloaded individually or in batches/all at once. A full list of files for this corpus can be found in the TO-in-NYC Files tab of this spreadsheet.

How to cite this corpus

Nycz, Jennifer. 2024. *The Torontonians-in-NYC Corpus.* Version 2024.1. Georgetown University. Accessible online at https://sites.georgetown.edu/corms/

The New Yorkers-in-Toronto Corpus (NYC-in-TO)

Author: Jennifer Nycz
Release date: June 2024 (v.2024.1)
Interviewers: Nicole Hildebrand-Edgar, Katharina Pabst, Emilie LeBlanc
Transcribers: Natalie Bazata, Amelia Becker, Nicole Rybak
Redaction assistance: Matthew Dearstyne, Ping Hei Yeung

Participants

NYC-in-TO consists of 28[2] primary speakers (19 women, 9 men) across 56 audio files, collected as part of the *Second Dialect Acquisition and Stylistic Variation in Mobile Speakers* project (NSF-BCS-1651108). The speakers were recorded between February 2018 and June 2020. All participants had been living in Toronto for at least 4 years at time of interview, though time in that city as well as the age at which speakers moved there varied across the sample; the mean age at time of interview was 52.6 years (sd=17.3), the mean age at which speakers moved to Toronto is 28.6 (sd=10.5), and the mean years in Toronto is 24 (sd=16.8). Metadata for NYC-in-TO includes broad information about speakers' demographic backgrounds (Table 2). No restrictions on social class or other characteristics were placed, other than those implied by the inclusion of participants who are able to move internationally. The speakers reported no speech or hearing disorders.

Metadata details

Speaker Code: each participant is assigned a unique code which denotes their corpus, gender, age at interview, and age moved to Toronto, and ends with an two-letter code assigned by the fieldworker (typically the first two letters of the participant's chosen pseudonym).
Corpus: indicates that the speaker's data is part of the NYC-in-TO corpus
Gender: each participant was asked to report their gender early in the interview while filling out a short demographic questionnaire; all participants identified as either male/man (M) or female/woman (F).
Ethnicity: participants were also asked to describe their ethnicity on the questionnaire; this column reflects the wording that each person used

[2] Data was collected from 30 speakers total; only 28 primary participants gave unambiguous consent for their data to be shared with other researchers.

Age at Interview: the age of the participant when they took part in the interview
Date of Interview: when data collection took place
Age Moved to TO: the age of the participant when they moved to Toronto
Years in TO: the number of years the participant has been living in Toronto
Friend present: "Yes" if a local friend of the participant was present for the interview, "No" if the interview was one-on-one. If the participant's friend did not also give clear consent to have their data shared, their contributions were redacted from the audio and transcript.

Table 2. Speakers in the NYC-in-TO corpus.

Speaker Code	Corpus	Gender	Ethnicity	Age at Interview	Date of Interview	Age Moved to TO	Years in TO	Friend present
NYC-in-TO_F_27_18_DO	NYC-in-TO	F	white	27	3/15/2018	18	9	No
NYC-in-TO_F_34_30_LY	NYC-in-TO	F	Lebanese Jewish-NYer	34	9/25/2018	30	4	Yes
NYC-in-TO_F_35_29_GM	NYC-in-TO	F	white	35	9/30/2018	29	6	Yes, friend data redacted
NYC-in-TO_F_36_25_MA	NYC-in-TO	F	white	36	10/24/2018	25	11	Yes
NYC-in-TO_F_40_36_MA	NYC-in-TO	F	caucasian (American)	40	2/20/2018	36	4	Yes
NYC-in-TO_F_43_26_LD	NYC-in-TO	F	caucasian	43	2/20/2020	26	17	No
NYC-in-TO_F_43_37_GR	NYC-in-TO	F	Korean	43	8/30/2019	37	6	Yes
NYC-in-TO_F_48_15_NI	NYC-in-TO	F	Polish, Irish, Croatian	48	7/19/2018	15	33	No
NYC-in-TO_F_49_42_MV	NYC-in-TO	F	Irish/Jewish	49	11/23/2018	42	7	No
NYC-in-TO_F_51_34_EL	NYC-in-TO	F	White Ashkenazi Jewish	51	3/28/2019	34	17	Yes
NYC-in-TO_F_60_26_IN	NYC-in-TO	F	Norwegian American Canadian	60	9/20/2018	26	34	Yes
NYC-in-TO_F_61_26_DS	NYC-in-TO	F	Eastern European heritage, Jewish	61	6/1/2018	26	35	Yes
NYC-in-TO_F_64_28_DI	NYC-in-TO	F	Caucasian Greek-Polish-English-Irish-American-Canadian	64	9/30/2019	28	36	Yes
NYC-in-TO_F_64_32_SA	NYC-in-TO	F	white	64	2/9/2019	32	32	No

NYC-in-TO_F_69_52_RH	NYC-in-TO	F	Jewish/Caucasian	69	4/16/2018	52	17	Yes
NYC-in-TO_F_75_45_JA	NYC-in-TO	F	Jewish	75	1/17/2019	45	30	Yes
NYC-in-TO_F_76_53_ME	NYC-in-TO	F	Caucasian	76	6/27/2020	53	23	No
NYC-in-TO_F_80_22_MY	NYC-in-TO	F	Jewish	80	1/29/2019	22	58	Yes
NYC-in-TO_F_80_25_GI	NYC-in-TO	F	Jewish	80	1/29/2019	25	55	No
NYC-in-TO_M_29_22_JO	NYC-in-TO	M	White (N. European)	29	3/17/2018	22	7	Yes
NYC-in-TO_M_30_23_CH	NYC-in-TO	M	White	30	3/1/2018	23	7	Yes
NYC-in-TO_M_32_17_KY	NYC-in-TO	M	Caucasian	32	11/20/2018	17	15	Yes
NYC-in-TO_M_35_25_DR	NYC-in-TO	M	Chinese	35	9/30/2019	25	10	Yes, friend data redacted
NYC-in-TO_M_49_31_BO	NYC-in-TO	M	Half Irish Half Eastern European Jewish	49	10/3/2019	31	18	No
NYC-in-TO_M_51_8_MA	NYC-in-TO	M	Caucasian	51	1/10/2019	8	43	No
NYC-in-TO_M_62_15_TH	NYC-in-TO	M	Jewish	62	9/24/2019	15	47	No
NYC-in-TO_M_74_27_AJ	NYC-in-TO	M	Italian-American	74	1/22/2019	27	47	Yes
NYC-in-TO_M_75_31_WA	NYC-in-TO	M	White (Danish/Irish)	75	1/24/2019	31	44	Yes

Data collection and recording

Participants were recruited through online fora such as expat community boards and social media (e.g., Facebook, Twitter). Participants took part in an approximately 60-minute long conversational interview with a local graduate student consultant which focused on growing up in New York City, moving to their new country and city, and impressions of the people and culture there. In eighteen cases, a friend or family member of the participant who is native to the new city (i.e. was born in and lived there until at least age 18, and has spent most or all of their subsequent adult life in that region) also participated in the conversation.

After the conversational portion of the interview, participants completed word list readings and minimal pair tasks. In the minimal pair task, word pairs distinguished by low back vowels (e.g. *cot* and *caught*; *dawn* and *don*) were presented to participants on paper or via an iPad, with other potential pairs (e.g. *Mary/merry, pin/pe*n) interspersed to pull focus away from the contrasts of low back vowels. Interviewers instructed participants to read each pair out loud and then say whether the two words sounded the same or different to them.

All interviews were recorded to 44.1kHz/16-bit wav files using a Zoom H4N solid state recorder and an Audio-Technica AT831b lavalier microphone. Interviews took place in the participant's home or a quiet public space. Consent was obtained and participants received an Amazon gift card worth $20 in Canadian dollars upon completion of the interview.

Files

Each audio file is stored in wav format, and is accompanied by an orthographic transcription in both .eaf (ELAN) and .TextGrid (Praat) format with the same filename stem. Transcripts are divided into breath units (roughly, units of speech occurring between pauses or breaths), with the transcription of each breath unit time-aligned to the audio file. Files are stored individually in a Georgetown Box folder, and can be downloaded individually or in batches/all at once. A full list of files for this corpus can be found in the NYC-in-TO Files tab of this spreadsheet.

How to cite this corpus

Nycz, Jennifer. 2024. *The New-Yorkers-in-Toronto Corpus.* Version 2024.1. Georgetown University. Accessible online at https://sites.georgetown.edu/corms/

3. Transcription conventions

English transcriptions of sound files were produced manually by humans using the ELAN transcription software. In order to ensure consistency among transcribers and to facilitate later automatic alignment processes, the following transcription conventions were followed (largely drawn from the Penn/LDC *Automatic Alignment and Analysis of Linguistic Change - Transcription Guidelines*):

- Lexical and morphosyntactic variation is represented faithfully.. Omitted words, false starts, repetitions, slang, or disfluencies are recorded as the speaker said it; that is, the trascriptons do not "clean up" what was said.
- Standard orthography and standard punctuation conventions were used.
- If a speaker abruptly stops in the middle of a word or restarts a word, the word is transcribed to the point until the speaker breaks off. A dash is used to indicate a word that was not completed.
- Numbers were written out fully (e.g. one hundred twenty three, seventy five). Similarly, words were transcribed in full rather than abbreviated (e.g. "Saint Joseph" instead of "St. Joseph").
- If a speaker used a common lexical variant of common word collocations, special spellings that could be recognized by the aligner were implemented as opposed to the standard orthography: *gonna, wanna, yknow, sorta, kinda, gotta, tryna.*

- Acronyms that are normally written as a single word but are pronounced as a sequence of individual letters were written in all capital letters, with each individual letter surrounded by spaces (e.g. C D C, Y M C A, B C).
- Incomprehensible speech is indicated by double parentheses (()). If it is possible to reasonably guess the speaker's words, possible transcriptions were surrounded by double parentheses during these stretches of uncertain words (e.g. And ((I just don't know))).
- For laughter, coughs, lip smacks, or other relevant noises of this nature, special codes within brackets were implemented in the transcription. A complete table of these special codes can be found in Table 3.
- Filled pauses were transcribed using the following limited set of common filled pauses recognized by the aligner: *ah, eh, er, uh, um.*
- For various yes/no interjections pronounced by a speaker, the following common spellings were used with a space after the first syllable: *mm- hmm, uh- huh, nuh- uh.*
- The following standardized spellings were also used to transcribe the following interjections: *duh, eee, ew, ha, hee, huh, hm, mm, nah, eh, ooh, woah, whew, whoops, yay, yeah, yep, yup.*
- If a speaker used and gave meaning to a word that is not an actual word, the word was spelled out as it sounded.

Table 3. Non-speech labels

Label	**Description**
{LG}	The speaker laughs.
{BR}	The speaker takes an audible breath.
{LS}	The speaker smacks their lips.
{NS}	Loud background noise.
{CG}	The speaker coughs, or clears their throat.
(())	Incomprehensible speech.

4. Redaction notes

Interviews were anonymized using a modified version of the CORAAL procedure (Kendall, Tyler and Charlie Farrington, 2021), where slashes and a redaction code are used to obscure real names, addresses, places of work, and schools. Redaction codes are as follows, with the # being the number of syllables obscured:

RD-WORK, RD-SCHOOL, RD-PLACE, RD-NAME
(e.g., Richard = /RD-NAME-2/; Elm Street = /RD-PLACE-2/)

All identifying information including the names of schools, universities, work places, and street addresses were redacted. Any names of neighborhoods, cities, states, and countries were not redacted.

5. Accessing the data

To obtain access to the linguistic data, interested parties must register by providing their name, email address, occupation, and their intent to use the corpora here. Once your registration is approved, you will receive a follow up email that will confirm your access to the corpus or corpora you have requested, with instructions for how to download files.

Corpus files for TO-in-NYC and NYC-in-TO are currently housed in a secure Georgetown University Box account; to view and download these files you will need to create a free Box account using an email address. Note that other users who have been added to a given corpus folder will be able to see who else has been invited to that folder. If you would like to be removed from a specific Box folder, email jennifer.nycz@georgetown.edu.

File formats

A key goal of CorMS is to make data as accessible as possible across different platforms and users. Accordingly, all files associated with this site use open, non-proprietary formats. Audio files were recorded as and are presented here as .wav files; transcript files are given in ELAN and Praat formats, both of which are readable as plain text files. ReadMe and metadata files are also saved in formats that may be opened as text files (.txt or .csv).

Time-aligned, orthographic transcripts delimiting breath units are provided. No other annotations are provided here (e.g., word- or phone-level segmentations, phonetic transcriptions, morphosyntactic parses, etc.). Any additional segmenting or coding along these lines necessarily involves additional theoretical assumptions or methodological decisions which may or may not be shared by other researchers, and approaches to data processing are constantly evolving. We at CorMs suspect that many linguists will want to "roll their own" analyses using whatever methods of data processing and analysis are considered to be best practices at the time, or that allows them maximal comparability with other studies they may be carrying out.

ELAN

Our interview recordings were transcribed by humans using ELAN, free audio and video annotation software provided by **The Language Archive at the Max Planck Institute for Psycholinguistics**). Our corpora transcripts are stored as .eaf (ELAN annotation format) files containing annotations which are time-linked to accompanying .wav audio files. The .eaf files may be viewed and manipulated in various ways in ELAN itself, or exported to a number of different formats , but they are essentially (marked-up) text files, which means that they can be opened in any text editor and manipulated by any number of programs or scripts.

Praat

Praat is free software commonly used for speech analysis. Our ELAN transcripts have also been exported as Praat .TextGrid files, for easy opening and manipulation alongside the audio .wav files in Praat. Praat TextGrids are also just fancy text files which can be opened in and manipulated by any program that deals with text.

Terms of Use

We hope this data will be of interest to a wide range of students and researchers who are interested in language use and mobility. CorMS is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike (4.0) International license. This means you are free to use and reuse the corpus for non-commercial purposes, but that you must cite the original corpus and any derivative versions of CoRMS corpora you develop and wish to share with others must be distributed using the same license.

Appendix: Word list used in NYC-in-TO and TO-in-NYC

Instructions: Please read each of the following words in your normal speaking voice. When you have finished a word, swipe left to move on to the next one.

hire
cot
side
doll
couch
heed
bot
decaf
pour
house
marry
slavic
pass
beg
knotty
out
father
cam
caught
bode
plaza
rice
had
put
mouth
pawned
tight
copy
hand
llama
mary
water
eyes
auto
hide
poor
caller
avowed
talk
but
pond
pen
housed
dies
odd
very
cap
tide
doubt
mat
hood
dowd
awed
mad
pen
man
bid
vary
head
mouthed
pin
drama
cloth
bead
collar
ice
bother
coffee
who’d
avocado
bite
off
how’d
taught
math
on
about
bade
don
booed
picasso
hut
bag
height
hat
cab
bide
pasta
pan
naughty
pajamas
hid
lava
bad
hoed
gouged
rise
merry
half
otto
sight
bed
dog
bat
tot
dice
cog
colorado
lot

Appendix: Minimal pair list used in NYC-in-TO and TO-in-NYC

Instructions: A pair of words will appear on each of the following slides. Please read each pair out loud, then say whether they sound the same or different.

bet	bit
latter	ladder
odd	awed
putt	put
rider	writer
talk	tock
merry	marry
pond	pawned
bid	bed
higher	hire
naughty	knotty
merry	mary
don	dawn
pole	pull

beg	bag
bought	bot
poor	pore
otto	auto
vary	very
born	barn
caught	cot
pill	peel
collar	caller
pen	pan
taught	tot
pen	pin