Today’s talk

• What do corpus studies stand to contribute to heritage language studies?
• Why are spoken corpora important?
• How does one create a spoken heritage language corpus?
• What considerations should one make when doing so?
• How can spoken corpora be used in research?

Corpus linguistics

• Corpus linguistics is the study of linguistic phenomena through the analysis of corpora
  • Corpus (“body”): a collection of
    • Texts (most common)
    • Transcriptions of speech (e.g., CHILDES, Switchboard)
    • Speech acts (multi-modal corpora)
  • Analysis: typically computer-based
  • Corpus linguistics is a method of studying language (Gries 2009)

Why might someone say this?

• Corpora consist of a speaker’s performance of language rather than a pure measure of her competence
• If we’re trying to get at the underlying grammar of the speaker(s), acceptability judgments would be a more direct approach
• Much work in corpus linguistics is descriptive and atheoretical
• (theoretically agnostic)

Why should even Chomsky like corpus studies for HLs?

• Corpus linguistic methods are increasingly used by linguists of all types, just like GJT’s and experimental data
• Further, due to heritage speakers’ lack of metalinguistic knowledge about their HL, it may be difficult to establish competence through other means
• see, e.g., Orfitelli & Polinsky 2012 on issues with use of GJT’s with heritage speakers
The methods themselves may be atheoretical, but they are principled, and the results become meaningful when applied to theory.
- Corpus data is still data!
- Many of the results provide important descriptive information about the language not obtainable otherwise (esp. for spoken language)
- Corpora can be shared, enabling more people access to relevant data

Why spoken corpora?
- Until recently, there has been a “written bias in linguistics” (Linell 2005)
- Speech has been seen as an inferior form of language
- However, it has become increasingly clear that speech and writing are different modes of language, with different characteristics and some separate rules, and merit study independently (Gilquin & de Cock 2011)

Speech corpora may provide insight into features especially important for heritage language studies, including:
- On-line sentence planning strategies
- Variation/uncertainty
- Evidence of extent of transfer from dominant language/“naturalness”
- Speaker strategies for coping with grammatical and lexical uncertainties and gaps

Why speech corpora in addition to texts?
- HSs may not be able to write the language well (or at all)
- Even spontaneous texts may more closely represent off-line production
- Spoken language is different!
- E.g., “disfluencies” may have communicative or speech management purpose
- Speech corpora of family interactions can also shed light on the baseline input for heritage speakers

Sample narrative - 1
Heritage English in France (age 10)
So there’s a sort of wolf who’s working in the street and he takes out a cigarette and he smokes, but a drop of water goes onto the cigarette, off down his nose. He looks upwards, and he sees a rabbit who’s putting water on his flowers, and he imagines the rabbit in the plate. So he sees clothes who are attached at a rope. He takes the clothes off, takes a rope, attach it to a flat, and starts to climb. But the rabbit who was getting his flowers he sees the rope and cuts the rope, so the wolf falls and rolls on the policeman.

Sample narrative - 2
Heritage English in Israel (age 8)
S: em there’s a dog, whatever, and he stepped on things and he found a cigarette so he got it, and then he saw that water’s coming and he wasxxx
I: and he what?
S: he saw water fell on him, and then he thought that who that is planting the food, and he took the string, and he took the where that they put the clothes to dry, and he run away, and then he climb then then the rabbit who’s planting the plants, so he cut the string and he fell down
PolLab Spoken HL Corpora

- For past 2 years, we have been developing corpora of several spoken heritage languages
- Chinese, English, Japanese, Korean, Russian, Spanish
- Generally, 30-45 mins per speaker; a brief interview and video narrations
- Videos selected to be culturally appropriate

Sample HL spoken corpora

- Polinsky Language Sciences Lab
  - Chinese, English, Korean, Japanese, Russian, Spanish
  - Heritage, L2 interviews, narratives
  - http://dvn.iq.harvard.edu/dvn/dv/polinsky

- New England Corpus of Heritage and Second Language Speakers (NECHSLS)
  - Spanish & Portuguese
  - Immigrant, heritage, L2 interviews
  - http://digitalhumanities.umass.edu/nechsls

The Heritage Language Variation and Change Project (UToronto)

- Cantonese, Faetar, Hungarian, Italian, Korean, Polish, Russian, Ukrainian
- Interviews, naming task, picture-elicited narratives
- http://projects.chass.utoronto.ca/ngn/HLVC/1_4_corpus.php

Three primary stages in corpus development

- Planning
- Developing
- Producing

Planning

- What is your goal in creating the corpus?
- What features are you interested in?
- What would you like to do with the data?

Developing

- Collecting data
- Transcribing data
- Categorizing/annotating data
Planning

What is the purpose of the study?
What features are you interested in?
What would you like to do with the data?
The answers will determine:
  - How much data you should collect
  - How that data should be elicited
  - How that data should be collected
  - How that data should be transcribed
  - Which consents/approvals must be obtained

Purpose of the study

Document the speech of a group/community?
Investigate the use of particular morphological/syntactic constructions?
Analyze the phonetic realization of segments?
Develop a private/restricted/publicly accessible corpus?

How much data?

Aren’t corpora huge?
Especially with lesser documented languages or for a specialized purpose, it’s not always possible to have a large corpus
In general, larger = better, but smaller corpora may nonetheless be adequate if carefully put together
Quality of data is at least as important as size (Kennedy 1998)

How should spoken data be elicited?

Three common approaches:
  - Interviews
  - Elicited narratives
  - Naturalistic recordings

Interviews:

Interaction between interviewer and subject in the language
Content can be guided by the interviewer
Conversational speech, but style/register may be influenced by speech of interviewer or the setting
Often useful for obtaining linguistic background info simultaneously
Elicited narratives

- E.g. picture-based narrative tasks (“frog stories”, Berman & Slobin 1994)
- Data isn’t fully natural, but enables researchers to reduce the effect of content as a variable, improving fluency of speech

PolLab and others: elicited narratives based on cartoons/silent movie clips
- May select culturally appropriate videos

May result in even more “natural” narratives than pictures
- Content completely removed as variable
- Speakers are recounting rather than inventing
- Recall of content does not appear to hinder performance

Naturalistic recordings

- Recording of conversation without presence/participation of interviewer
- May avoid “interviewer” effect on language so may obtain most natural speech samples
- But, little control over speech participants and content of recording

How should the data be collected?

- May be useful for examining interactions between speakers
  - Input from parents/family members to children
  - Interaction between heritage speakers
  - Codeswitching in natural discourse
  - May be more difficult for making cross-speaker comparisons

- If interested in phonetics, recording quality is of prime importance to obtain accurate measurements
- Otherwise, quality less critical so long as speech can be perceived clearly for transcription
- Consider effect of recording setting on naturalness of speech and potential for disruptions, background noise
Ethical considerations

- IRB approval required
  - IRB approval required for text-based and spoken corpus collection activities
  - May place restrictions on length of archiving audio/video
  - Will require confidentiality and anonymity for subjects
  - Stricter scrutiny if wish to make recordings available
    - E.g. make all participants (subjects, investigators, transcribers, researchers) sign agreements relating to the intended and permitted use
- Written consents required from subjects
  - Consent to record
  - Consent to distribute

Anonymity

- Measures must be taken to ensure anonymity of the data to prevent subject from being identified
  - Use subject codes or pseudonyms rather than actual names
    - HR1, HR2, HR3…HK1, HK2, HK3…
    - H1, H2, H3…N1, N2, N3…
    - AB, DI, DL, PR (but do not use subjects’ initials)
- Anonymity more difficult when want to share audio or video
  - Voices and images may very easily enable the identification of the subject
  - Modification of the data may be made (voice alteration, pixelation of video), but may limit the utility of the data
  - Measures taken may depend on intended distribution of materials

Ensuring anonymity within corpus

- Review transcribed material to ensure personally identifying material removed
  - Personal names (self, siblings, teachers, etc.)
  - Street addresses
  - School names
  - If recordings will be made available, remove material from recordings as well

Transcribing/Annotating

- Corpus studies depend on the existence of a textual corpus that can be analyzed
  - But, the preparation of this corpus is undoubtedly the most resource-consuming parts of the process
    - Transcribing the recorded data
    - Annotating the recorded data
    - Verifying the accuracy of transcriptions and notations
  - Various approaches can be taken to each
Transcribing data

- Manual transcription
  - Done by native speakers? L2/heritage?
  - Linguistically trained?
- Automated transcription (speech recognition)
  - Several projects in the works that may enable this in the future

Accuracy of transcription is key

- Various approaches to ensuring accuracy, from statistical sampling to multiple checks
- Approach taken should depend partly on resources and amount of data

Orthography

- Standard orthography for language vs. phonetic orthography?
  - Who is intended audience?
  - What is the purpose of the study?
  - Does the standard orthography allow you to convey information relevant to your purposes?

Whether/how to indicate deviations/errors in pronunciation?

- three [fɹi] vs. free
- says [sɛz] vs. says [sej]
- Depends on the purpose of the study
- Can give negative impression of speaker if not related to the questions being examined

Annotations

- "Annotations": markings to add extra-linguistic information to the corpus
  - error and disfluency markings
  - POS tagging
  - Comments (setting, "stage notes")
- Enable searches for and analysis of these items, including errors and disfluencies, in context
- Which annotations are made will depend on the purpose of the project

Various systems exist – each corpus contains a key to its annotations

- Which annotations you use may depend on your plans for the data
- PolLab annotations based on CHAT conventions (CHILDES corpus)
- Olesya and I are currently working on a uniform system for use in heritage corpora
NECHLS (speaker 17 – heritage)

PolLab (subject AR)

S: In his cart, and he started walking and then he took all kinds of um@fp food that you dip "like" meat, to dip "like" xxx, and then he took the pan and "like" mushed everything so everything fell, and then he started walking "like" was slipping on everything, and then the don't know how it's called the food that sold everything "like" the owner saw him so he got angry, and then he stood up and he got [//] he bumped in him and then they started to dance like balerínas and then they did a circle around the fountain [//] water fountain and then the rabbit took um@fp

Sample PolLab annotations

- @fp filled pause
- #, ##, ### unfilled pause
- & false start
- [/] retracing without correction
- [//] retracing with correction
- <> filler/placeholder words
- xxx unintelligible in recording

Resource requirements

- For text-based corpora, resource requirements may be minimal
- Processing spoken data requires more resources than processing textual data
- But it contains information not obtainable from texts!

PolLab approach: 2½ steps

- Step 1: plain transcription by a native or high-level heritage or L2 speaker, with minimal linguistic training required but basic training in transcription (4-12 hours per hour of recording, for Heritage English)
- Step 2: verification of transcription and addition of annotation by linguistically trained person (2-3 hours per hour of recording, for HE)
- Step 2½: verification of annotation by second linguistically trained person (5-1.5 hours per hour of recording, for HE)

Use of spoken corpora in research

- As mentioned earlier, corpora are increasingly used for data supporting linguistic analyses
- Much remains to be discovered about the features of heritage grammars, and spoken corpora may provide new, useful data
- Data mining spoken corpora for new research questions
• Enable study of features of spoken language of heritage speakers, insight into on-line processing/production
• Findings from corpus studies are often followed up with other methodologies (e.g. experimental work)

Developing cross-linguistic correspondences
• Speech corpora of the sort described can be easily duplicated with other heritage languages/other dominant languages
• Enable various comparisons
  • Same HL, different dominant languages
  • Same dominant language, different HLs

• Most existing work is on HLs with English as the dominant language
• Are some shared features (e.g. simplification/loss of morphology) due to effect of the dominant language?

PolLab current projects:
• Heritage English in France and Israel
• Heritage Russian in the U.S. and Germany
• Hope to add more soon
• These types of comparisons may help reveal “heritage” features and add to our understanding of heritage grammars

PolLab spoken heritage corpus
• As we said, in progress over the past two years
  • Chinese, English, Japanese, Korean, Russian, Spanish
  • Files are constantly being processed; a new batch should be uploaded within the next couple of weeks
  • The project is the result of a collaboration with a number of people at PolLab and outside
  • If anyone would be interested in participating, please let us know!

Summary
• Corpus studies may provide many useful benefits to heritage studies
  • More, shareable data on HLs
  • Spoken studies may provide us with data on features not accessible from text-based corpora
  • Additional means of attempting to identify features of “heritage” grammar
• Developing a corpus need not be a daunting task!
• Olesya walked you through the beginning steps of setting up a corpus in GOLD
• The resource requirements to develop a spoken corpus may be more significant, but spoken language contains a rich amount of information not otherwise available

Thank you!