Preaching to the Choir: What all linguists need to know about defining and operationalizing ethnicity and race in research

Alicia Beckford Wassink, Robert Squizzero, Martin Horst, Alex Panicacci, Emily M. Bender
University of Washington
1. Equipping linguists in all fields with a critical lens for understanding (and using) race and ethnicity
   - Use of terms
   - Recognizing racializing practices
   - Collecting demographic data
2. Who are we?
3. Why this paper?
4. The study of textbook guidance
5. “Defining” race and ethnicity
6. “How much demographic data do I need?”
Who we are:

Dialectology
Variationist Sociolinguistics
Sociophonetics
Syntax
Interactional Sociolinguistics/French and Italian studies
Computational linguistics
Social Psychology
Phonetics

Language submission title: Commentary “Collecting and using race and ethnicity data in linguistic studies” (Squizzero, Horst, Wassink, Panicacci, Jensen, Moroz, Conrod, and Bender, in prep.)
Why this study?

1. Linguistics has been criticized for undertheorized application of the notions of race and ethnicity (from outside and inside)
   - 2019 LSA Statement on Race

2. Improved research ethics & respect for communities we represent

3. Colleagues’ requests for recommendations and templates

4. Better alignment with our sister fields (anthropology, archaeology, psychology, sociology)
“All linguistic research has the potential to reproduce or challenge racial notions.”

- Linguistic Society of America Statement on Race (2019)
New researchers are most likely to find guidance about conceptualizing (speech/language) community in critical sociolinguistics, applied sociolinguistics, language documentation, language variation and change.

New researchers are most likely to find general guidance about designing demographic prompts in methods texts about studying language variation and change.

All subfields queried (including core subfields, applied linguistics, corpus linguistics, language documentation, anthropological linguistics)

62 textbooks found and reviewed
Publication years: 1951 and 2020

Actually “went there”
(Milroy & Gordon 2003; Heller et al. 2018)
My study is “just about language.” Sociolinguistic issues are not at the heart of the linguistic enterprise, so it’s not my concern. (Kiesling 2011, Milroy 1987, Kibrik 1977)

My research is socially “neutral”/uncontaminated by investigation.

I can judge who my subjects should be/how to describe the community.

Race = Ethnicity

We act in the social world and [must] reflect upon ourselves and our actions as objects in that world.” (Hammersley & Atkinson 2006)
Race & ethnicity defined

- Race and ethnicity are viewed as overlapping and used interchangeably.
- In linguistic research, it is important to distinguish between:
  1) Approaches that are static and essentializing (usually race-based)
  2) Approaches that are practice-based (usually ethnicity-based)
Race

- *Race* refers to a group sharing physical features, especially skin color, facial features, eye shape, and hair texture (Bobo 2001; Spears 2020)
- Definitions are pointless outside of an acknowledgement of *racism* (Lanehart, 2023)
- *Race essentialism* is the tendency to view race as biologically based, immutable, and informative (Haslam, Rothschild & Ernst 2000; Prentice & Miller 2007)
  - Race essentialism has been linked to racial stereotyping and prejudice (Levy & Dweck 2003; Williams & Eberhardt 2008)
- Essentialist racial classification schemes based in biology or genetics are:
  - unreliable (Garcia 2020; Relethford 2009)
  - severely flawed (Keita et al. 2004)
  - completely arbitrary (Omi & Winant 2014)
Ethnicity

- *Ethnicity* refers to a grouping based upon:
  - Shared signs (in the semiotic sense),
  - Shared aspects of a common culture, or
  - Shared practice

- Material manifestations of shared aspects of culture may include:
  - Following patterns of dress
  - Adhering to diets or eating particular foods
  - Observing holidays
  - Practicing religions, and crucially
  - Speaking *languages and language varieties*

  (Garcia 2020, writing on behalf of the American Anthropological Association and the Society for Anthropology in Community Colleges)
So should I ask about race or ethnicity?

- *Probably ethnicity.* When describing language, useful to adopt framing, practice-based approaches
- less likely to be essentializing
- less exclusionary
  - Sometimes, language users do not possess the phenotypic characteristics stereotypically associated the speech community to which they belong
  - Practice-based approaches describe what people **do**, not what people look like
  - Main exception: if you are investigating racism, you might ask about race
How did we come up with our recommendations?

1. Reviewer comments we’ve received on journal manuscript submissions (including when we’ve been criticized)
2. Conversations about research ethics (with our community partners, on Twitter, in the Media, online)
3. Two years of laboratory group discussions
4. Reviewer reactions to two earlier versions of this paper
5. Texts mentioned earlier that provided guidance (or examples where these would have been helpful)
6. Wisdom shared by some of you...
How much demographic information might I need?

**Large Corpus-Driven:**
- Computational Linguistics
- Historical Linguistics
- Corpus Linguistics

**Formal:**
- Syntax
- Semantics
- Morphology
- Phonology
- Typology
- Pragmatics

**Experimental:**
- Phonetics
- Phonology
- L1 & L2
- Acquisition
- Bilingualism
- Psycholinguistics
- Neurolinguistics
- Sociolinguistics

**Qualitative:**
- Discourse Analysis
- Sociolinguistics
- Sociocultural Linguistics
- Language Documentation
- Raciolinguistics

Probably less

Probably more
Typical types and amounts of data:
- Digitized
- Internet-sourced
- Vast corpora (millions of observations), largely anonymous

- Typical style of research question involves **hypothesis testing**
- The need for race and ethnicity data can arise for specific tasks which implicate social identities
  - When working with “unlabeled” social media data (Twitter or Reddit)
  - Linking language use and social identity (Hate speech detection)
  - NLP: building broadly useful tech (e.g. speech to text)

(Abreu 2015)
Bender and Friedman (2018)
Gonen and Goldberg 2019
Recommendations

1. Closed dataset? **Be transparent** when self-identity data are not known
2. Your dataset? let participants opt-in, give informed consent, **self-report demographics**
3. **Avoiding essentializing** linguistic features as THE markers of racialized language varieties
4. **Beware of linguistic appropriation** in datasets
5. Compare large datasets to relevant studies drawn from the same user population

Abreu (2015)
Charity Hudley (2017)
Typical types and amounts of data:
  - Grammatical intuitions (unnamed consultants)
  - Small number of consultants
  - Data from preexisting studies

- Typical style of research question involves **deductive reasoning and explanation**
- The need for reporting ethnicity data arises because such methods may:
  - Misrepresent the provenance of the phenomenon
  - Misreporting the state of the grammar of interest
Our Recommendations:
1. Collect minimal demographic information to establish regional and social location (region, ethnicity, gender identification and language background)
2. Identify your sources (judgements and examples)
   How many speakers? From what regions? Of what ethnicity?
3. If your source is yourself, write a brief positionality statement
4. Ask your sources what relevant social categories would apply if they were to write a brief positionality statement or bio for themselves. Include this information in a footnote or an appendix

Legate et al. (2020)
Experimental:

Typical types and amounts of data:
- Large judgement or random sample (primary data)
- Inferential, time series and descriptive analysis
- Coding for age, gender, language exposure, interlocutor type, treatment, group, etc.

Typical style of research question is **descriptive** or **correlational**:
- Sociolinguistics: Distribution of some linguistic feature, e.g., “Is use of Avertive *liketa* disfavored in constructions displaying multiple-negation in Mississippi AAE?”
- Acquisition: “Does X feature (FL learning) occur differently in group Y (intervention children) than in group Z (CPC)?” (Ferjan-Ramirez & Kuhl, 2020)
Experimental:

Recommendations:
1. understand which ethnic labels might be relevant in/to the community of interest. Immediate social network useful.
2. allow participants/caregivers to not answer questions.
3. allow participants/caregivers to choose multiple options.
4. multiple choice with well-justified categories, free-response or interview-style (consider including an option where participants/caregivers can name one or more labels not already included which are relevant to them)
5. Report analyst’s positionality.
Qualitative Research Questions vary. Characterization of some linguistic phenomenon within an individual language.

- Community participation
- Lapierre (p.c.): Age, gender, clan, familial and social roles (marriage), village of origin
- Analyst records key facts

**Typical types and amounts of data:**
- Qualitative - Ethnographic observation of “lived routines of daily living”
- Period of observation spans years
- Large amounts of data recorded by the analyst

**Research Questions vary.** Characterization of some linguistic phenomenon within an individual language.

- Community participation
- Lapierre (p.c.): Age, gender, clan, familial and social roles (marriage), village of origin
- Analyst records key facts

**Key:** Ethnographic approach

Ethnos > Gk. “belonging”

**Recommendations:** See previous slide; include practices
Conclusion: What linguistics has to gain by critical treatment of race & ethnicity data

- Better design considerations
- Better transparency (e.g., [under]representation of speaker types in NLP modeling)
- Better representation of under-sampled groups
- Avoiding social harm to participants
- Appropriate level of generalization for theory building
- ... and more!
Acknowledgements

Anne Charity Hudley,
Arthur Spears,
Sonja Lanehart,
Laada Bilaniuk,
Maya Angela Smith,
Sara B. Ng,
Laura Munger
References (and reading suggestions)


Fasold, R. 2019. Comment on: LSA Statement on Race


