Language and Dialects

Camilla Crane, Dan Hockstein, Nate Nihart, Elizabeth Prieto

Documentation of your CSV file(s).

This should include explanations of any new columns you added, and how to understand the values in these columns if it is not obvious.

Primary Spreadsheet

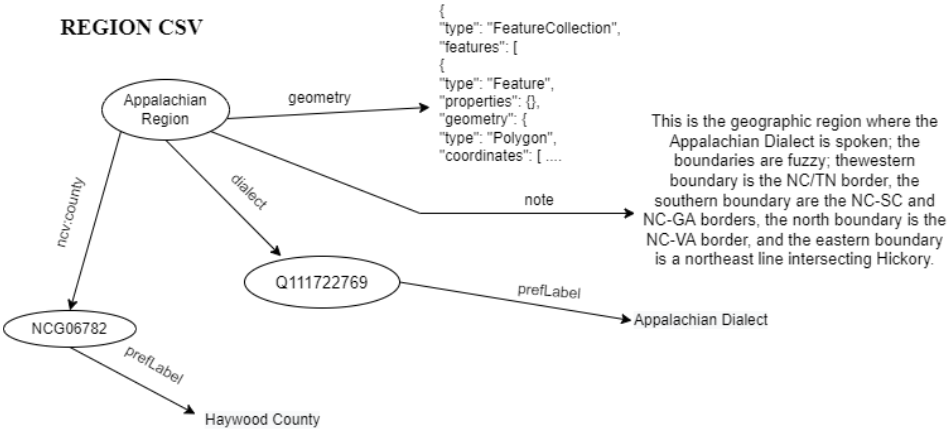
Column	Description	Notes
ncg_id		These will be newly minted as part of the added dialect regions.
skos:prefLabel	The names of the dialect regions.	
skos:altLabel	Alternative names for regions.	
dialect_wikidata_id	These are the wikidata identifiers for the dialects spoken in each dialect region entry.	
skos:note	This is equivalent to the skos:Note "Description" entity in the gazetteer.	
skos:closeMatch	Similarly named unique identifiers within the gazetteer.	This could be the output of a library or function that determines closeMatch data.
Geojson:geometry	These are the [potentially disputed] bounds for the dialect region, as noted in Talkin' Tarheel.	GeoJSON coordinates – may be less cumbersome to host them as URIs in subsequent versions of the data.
ncv:county	Counties included in the described dialect regions based on the “geometry” field. These may have overlap.	This is the domain in which we tie these entries to particular entries in the gazetter; rather than choosing a more precise but potentially more flawed method of distinction within the terms of the gazetter, it made sense to over-describe bounds based on included counties. Other methodologies to define bounds of these regions could be major cities, bounds based on features, etc. - but this seemed to be the best compromise between convenience and description.

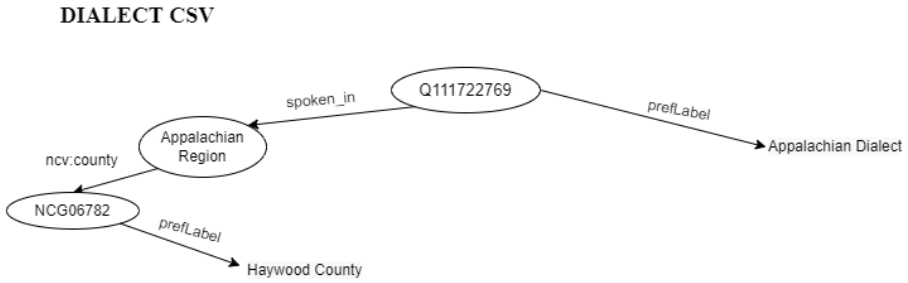
Dialect Spreadsheet

Column	Description	Notes
dialect_wikidata_id	These are the wikidata identifiers for the dialects spoken in each dialect region entry.	These can be expanded upon, and it would be ideal to have an administrative Wikidata linkage to the NCG dataset.
spoken_by	The group of people who speaks this dialect (only relevant for Lumbee Dialect)	This can be defined in a multitude of ways, but since the scope of our dataset focused almost entirely on English-speakers, this column felt as though it was rife with problems. Might there be overlap? Focusing on a population and associating it with a region could be useful, but is ultimately a level of granularity that would involve decisions surrounding the dataset that we didn't quite feel comfortable making.
spoken_in	The region in which this dialect is spoken in	[wikidata: indigenous to: https://www.wikidata.org/wiki/Property:P2341]
example_words	Words and phrases associated with the described dialect region.	A single example word was added to this field, along with the corresponding broadly used English language term; the source of this information is also included. More words can be added based on additional research and citations. A more elegant representation of this data would likely involve additional csv files to separate the example dialect word and the corresponding word.
talkin_tarheel_reference	Chapters in Talkin' Tarheel that discuss these dialects	
RepresentativeRecording	Recording of the dialect in the SLAAP database	
RecordedAt	Where the representative recording was made	

Documentation of how you think the resulting RDF should look.

This can take the form of examples in Turtle format of RDF describing 2 or 3 of your places.





Ncg:ncg_id [to be minted] skos:prefLabel “Appalachian Dialect Region”;

Ncg:dialect_wikidata_id wd: Q111722769;

Skos:note “This is the geographic region where...”;

Skos:closematch ncg:NCG00275;

Geojson:geometry “{ "type": "FeatureCollection", "features": […”;

Ncv:county “ncg:NCG397548|ncg:NCG8453948|ncg:NCG5479348”;

.

Documentation of your research process.

This should include details about the sources you consulted and tools you used to produce your CSV file. It should also include a description of sources you consulted that turned out to not be useful, or other things you tried that did not pan out, or that you did not have time to finish.

In our preliminary search of dialects in North Carolina, we found many popular sources based on the work of Walt Wolfram, a sociolinguistic researcher at NC State. We decided to use the book *Talkin' Tarheel,* authored by Wolfram and Jeffrey Reaser (another linguist at NCSU), as our main source of information. It was the most comprehensive source on NC dialects that we found.

We also reached out to Wolfram for additional sources. He pointed us to the SLAAP database, a database of sociolinguistic recordings run by researchers at NCSU and the University of Oregon. There were a multitude of recordings from North Carolina, but they unfortunately did not contain any metadata identifying the dialect being recorded. Because of this, we decided to manually listen to the recordings and create a column in our spreadsheet with one representative recording per dialect.

Talkin' Tarheel contained a lot of information outside of the scope of our project, including information on AAVE, Spanish speakers in NC/"Spanglish", and several now-extinct Indigenous languages. We decided to only focus on dialects of English that could be tied to a geographic area. However, one interesting additional source that we found for this was a 2019 report by the North Carolina Council for Women and Youth Involvement, "Limited English Proficiency and Less Commonly Spoken Languages in North Carolina," that provided county-level details on English proficiency as well as the population speaking other languages (for some counties). Another direction that didn't pan out due to a lack of time and the complexity of the resources included more focus on social dialects within NC (AAVE and Hispanic dialects); we located a dissertation (Taylor Jones, 2020, The Great Migration and Regional Differentiation in African American English) that looked at AAVE dialects in the US, but it was too dense to pick apart for use in such a short time frame). As part of this, we also wished to explore how we might label the data that we were proposing for addition to the gazetteer in a way that addressed potential harm that it might cause (i.e. if geographic bounds of dialects were based on research that didn't equally account for marginalized groups, then naming this).

Once we identified our scope, we manually created our spreadsheet headings as there was no already-organized dataset we could download and manipulate.

Ideas for future work.

This can include things you ran out of time to do, sources that you found but did not have time to investigate, or anything else that you would suggest a future group might work on.

As noted above, we communicated with professor Walt Wolfram, who hinted at but did not ultimately share with us a corpus of organized data that would be particularly useful to directly link or organize. It would be interesting to have access to the source material for works like *Talkin' Tarheel,* as there was clearly quantitative research behind the various charts and maps present in the book.

English speakers became the scope for this project, but additional Dialect Regions exist other than those documented here even within the realm of English. One project from Taylor W Jones at the University of Pennsylvania suggested regional distinctions of vernacular African American English. There were also extinct indigenous languages, which we opted to exclude from this dataset, that would be useful to examine within the context of existing Gazetteer data.

Particular words that were defined within a dialect felt like a more complex dataset that was separate from defining dialects, but would be a subsequent project associated with each dialect region.

Additionally, there is likely other authoritative data aside from *Talkin' Tarheel* even within the limited scope of our project that could be incorporated and subsequently analyzed/organized to a similar degree.

Finally, future work would benefit from contextualizing the regional data (how white is the data - which populations were a part of the research, do the dialects hold true across social groups; how do AAVE dialect regions match up)

Credit and acknowledgements.

This should list each member of your group and state what their contribution was. You can also acknowledge the assistance of others not in the group here.

Camilla Crane

- Retrieved *Talkin' Tarheel* digital copy and created an initial spreadsheet with all dialects mentioned in the book and some initial column headings (group would later narrow the scope and add headings collectively)
- Created GeoJSON geometries based on maps in *Talkin' Tarheel*
- Contributed to example Turtle and research documentation in report
- Emailed web editor of SLAAP database to gain access

Dan Hockstein

- Contacted Dr. Wolfram, sought additional references
- Added SLAAP recordings to dataset
- Delineated counties within the dialect csv
- Documented fields as they were idealized
- Assisted with in-class questions/walkthrough
- Contributed to additional documentation above

Nate Nihart

- Created documents for initial proposal and scoping, and final report
- Looked for AAVE dialect sources, reviewed dissertation mentioned above
- Contributed to documentation and ideas sections in final report
- Added wikidata entry for "Lumbee English" language dialect
- Retrieved non-electronically-available-book on Lumbee English from NC collection for review
- In dialect region csv, added skos:note description of regional bounds
- In dialect csv, added example words for dialects

Elizabeth Prieto

- Created WikiData entries for dialects
- RDF example