Dialect mapping of grammatical features in North American English using a corpus of geotagged Youtube transcriptions

Brett Hashimoto, Brigham Young University
Joseph A. Stanley, Brigham Young University
Jack Grieve, University of Birmingham

BYU
BRIGHAM YOUNG
UNIVERSITY

UNIVERSITY OF
BIRMINGHAM

Introduction

- Dialect mapping primarily based on phonetic features nowadays (e.g. Labov, Ash & Boberg 2006)
- Some (but not much) corpus-based socio (Grieve 2016)
- Some grammatical/lexical research, but the minority and mostly elicited rather than naturalistic (e.g. Kurath 1939, Carver 1987, Leemann et al 2018, Leemann et al 2020)
- Little that maps large areas or focuses on multiple features simultaneously (though see Kim et al 2019 and Stanley 2022)

Introduction

- Grieve (2016)
  - Regional variation in written American English
  - 200,000 letters to the editor (36+ million words)
  - 240 cities across the US
  - 135 lexico-grammatical alternation variables
  - Mapped variation according to each of these variables
  - Uncovered five primary modern American dialect regions

Research Purpose

- Generate maps of the distributions of 100+ lexico-grammatical feature alternations in **spoken North American English**
- Bigger project:
    - Compare our work with previous dialect mapping
    - Multivariate analyses
        - Factor analysis
        - Cluster analysis

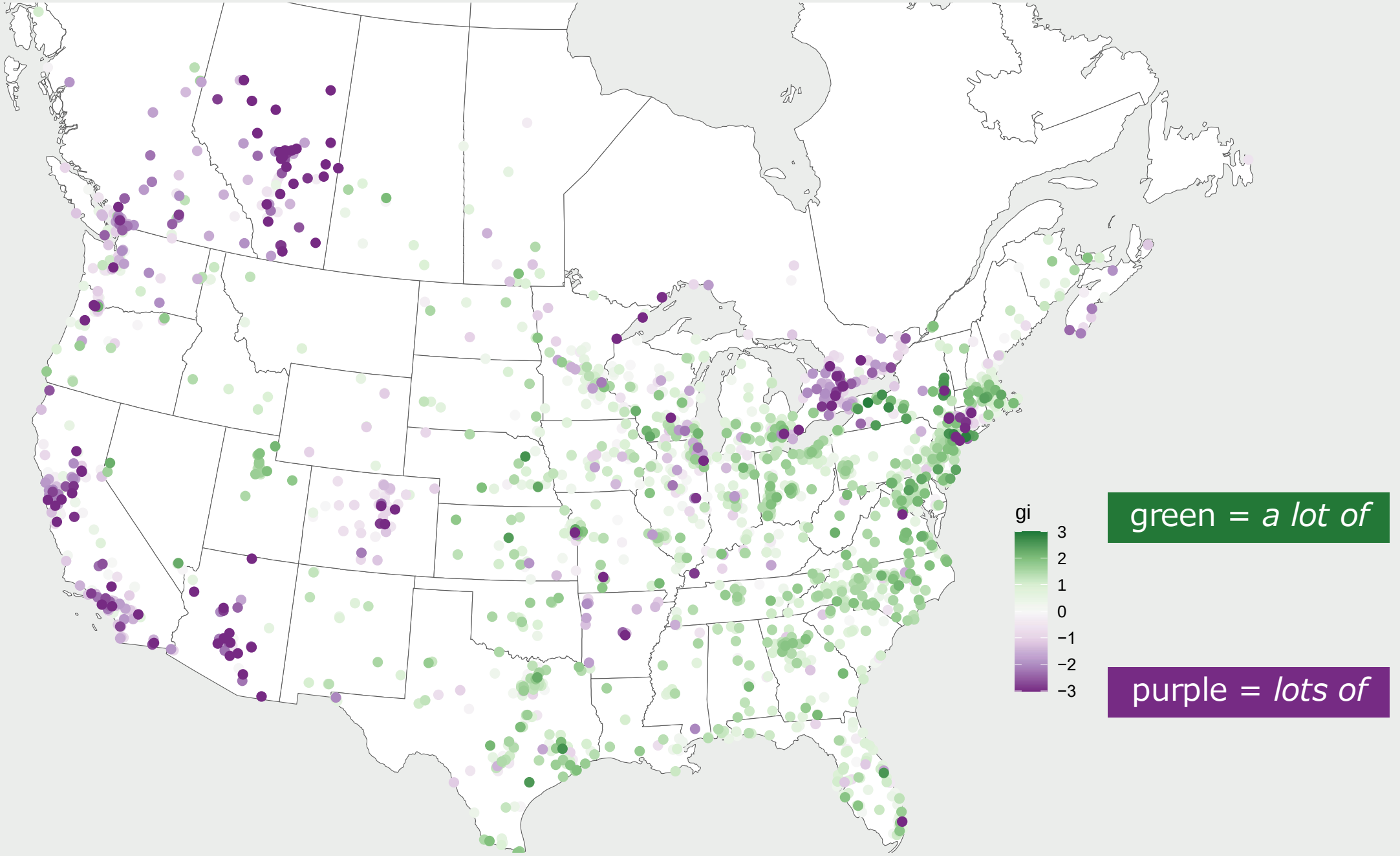
Corpus

- Corpus of North American Spoken English (CoNASE: Coats, 2019; 2023)
  - YouTube channels of mainly regional and local government entities or other governmental/civic organizations
  - Stratified sampling from counties across the US and Canada
  - 301,847 texts; 154,041 hours of spoken language; 1,252,066,371 words
  - Autotranscribed and geotagged
  - Stanza lemmatized; Part-of-speech tagged
- Same 135 grammatical alternation variables as Grieve (2016)
- Algorithms for feature identification were altered from Grieve (2016) to be more suitable
- Accuracy checking of features

Quantitative Analysis

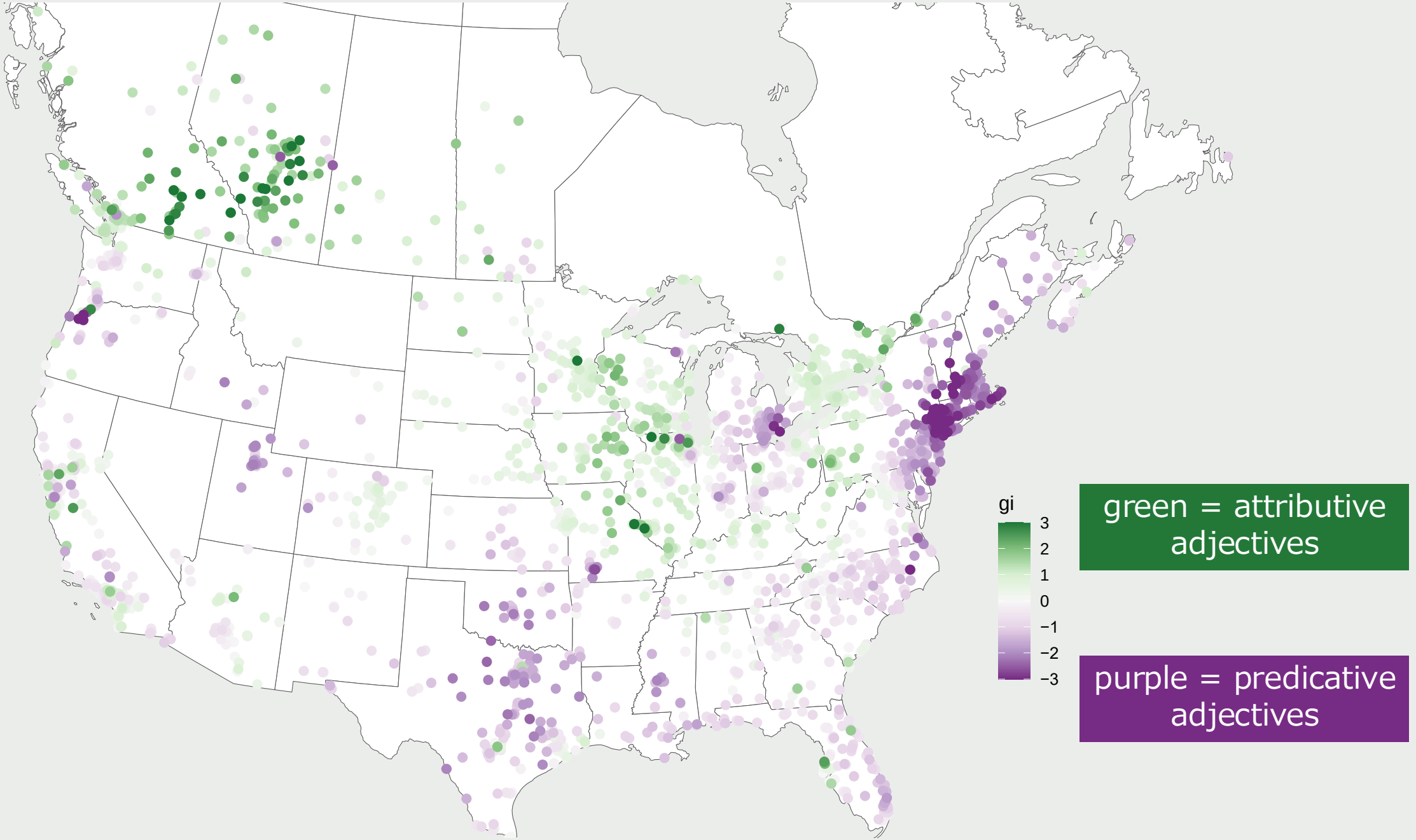
- Proportions by location
  - We calculated the proportion of each variant for each feature, i.e. A/(A+B).
  - Weighted average per location (301K texts → 2,537 locations)
- Spatial stats following Grieve (2016)
  - Getis Ord-G_i statistic: For each location, indicates whether there is high/low clustering at that location (without regard to political boundaries)
  - Interpret this like a *z*-score, so high absolute values = statistically significant.
- Maps
  - Plot points (if there's enough data).
  - One variant is green; the other is purple.

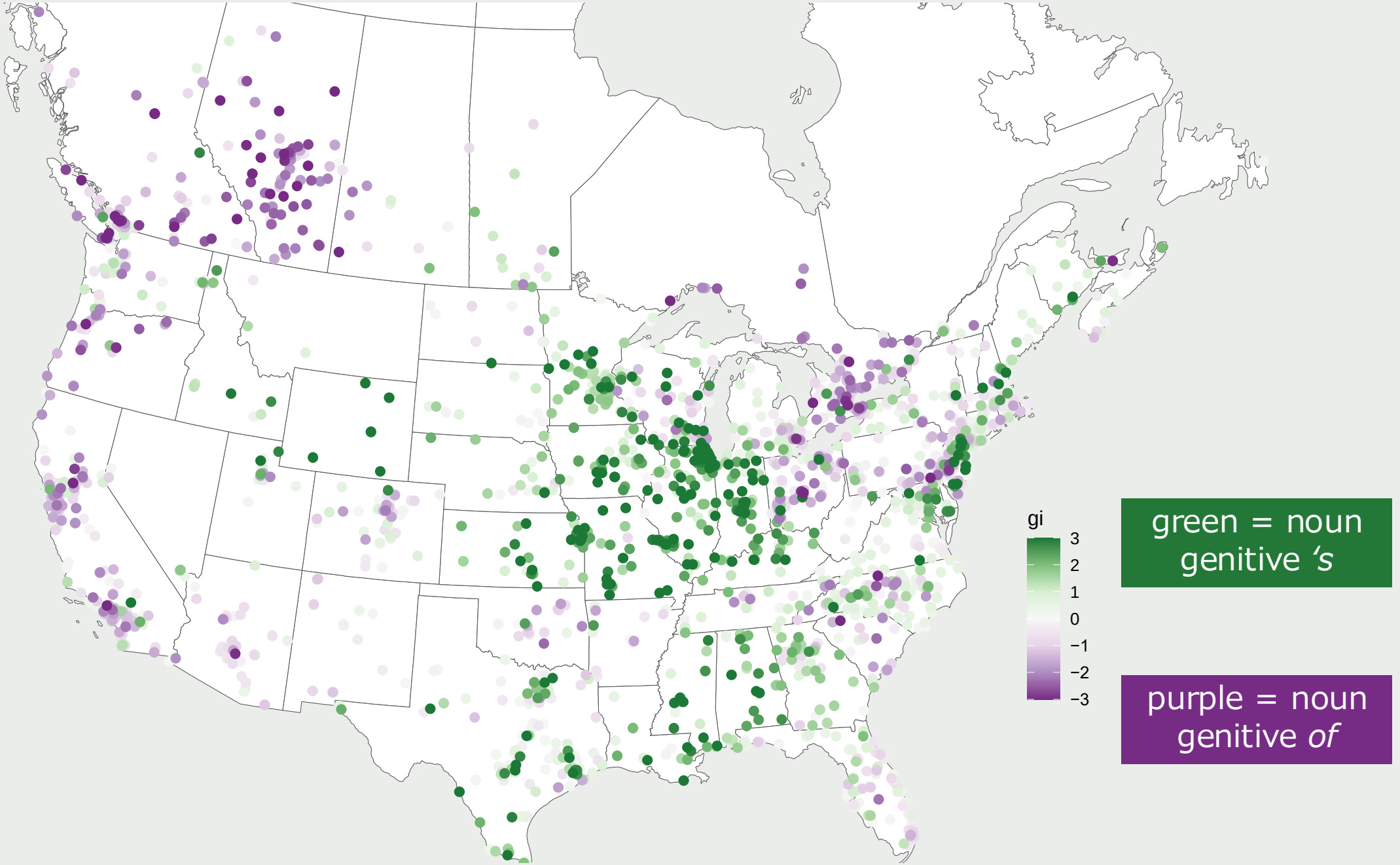
RESULTS



green = *a lot of*

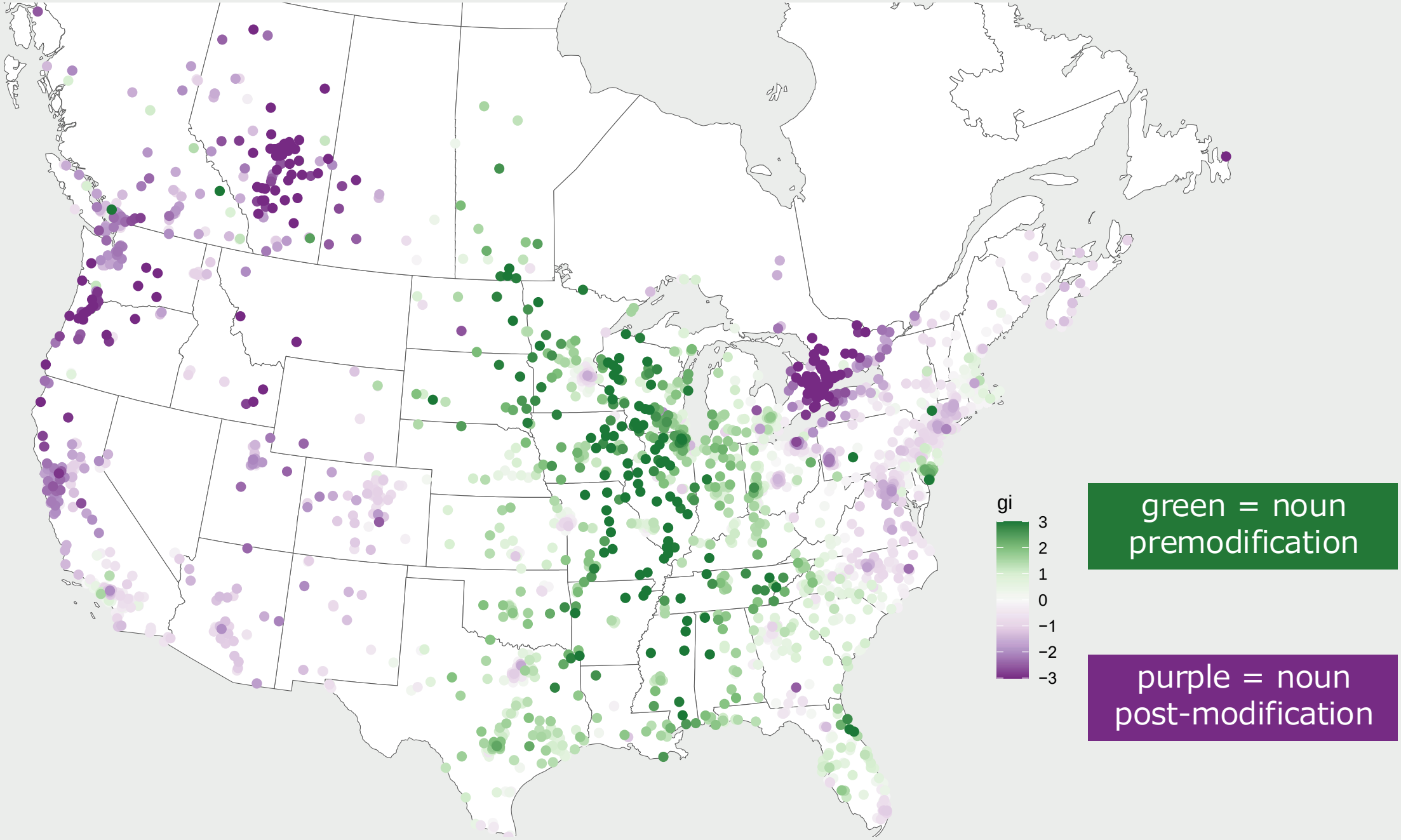
purple = *lots of*


gi
3
2
1
0
−1
−2
−3
green = attributive adjectives
purple = predicative adjectives



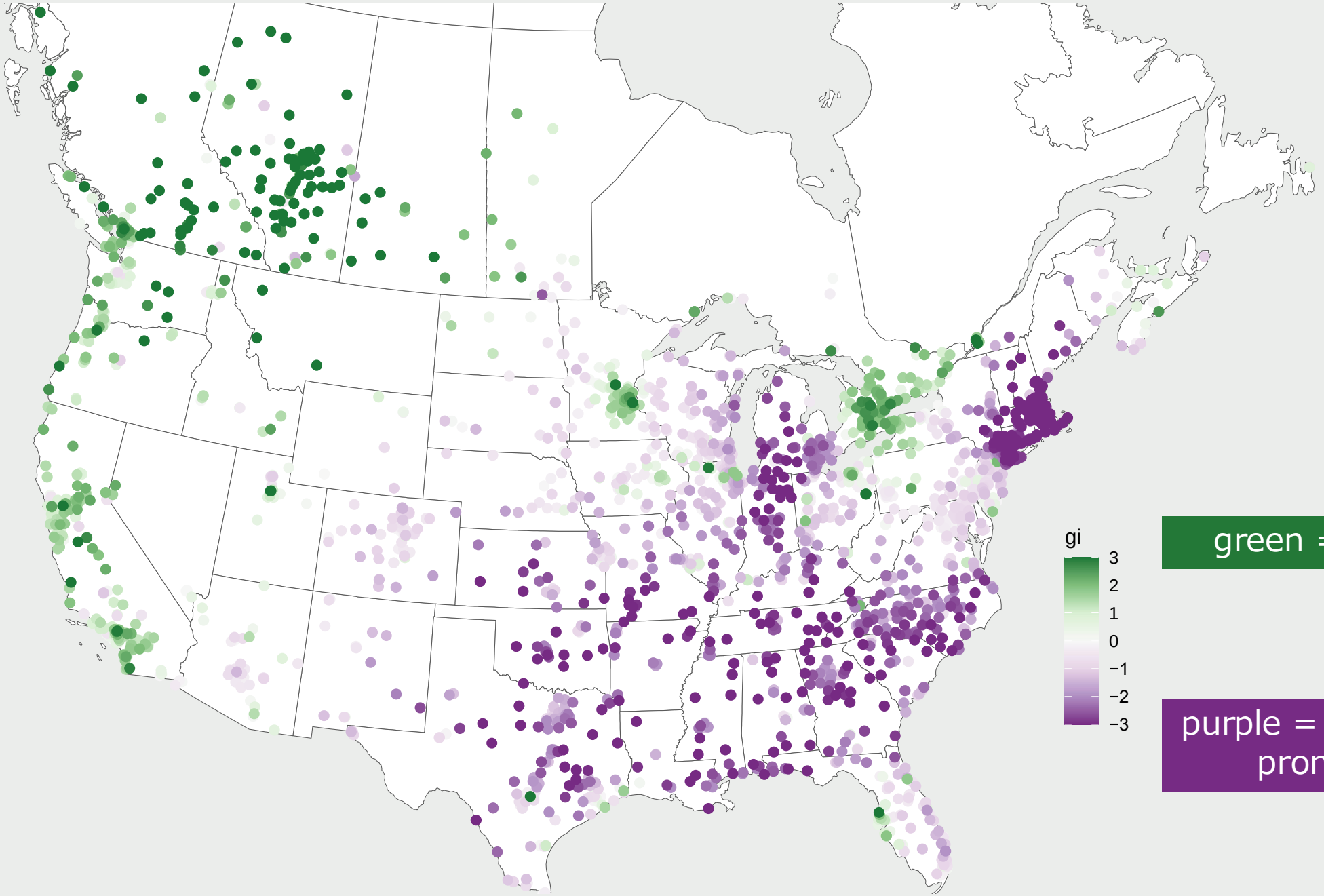
green = noun genitive *'s*

purple = noun genitive *of*



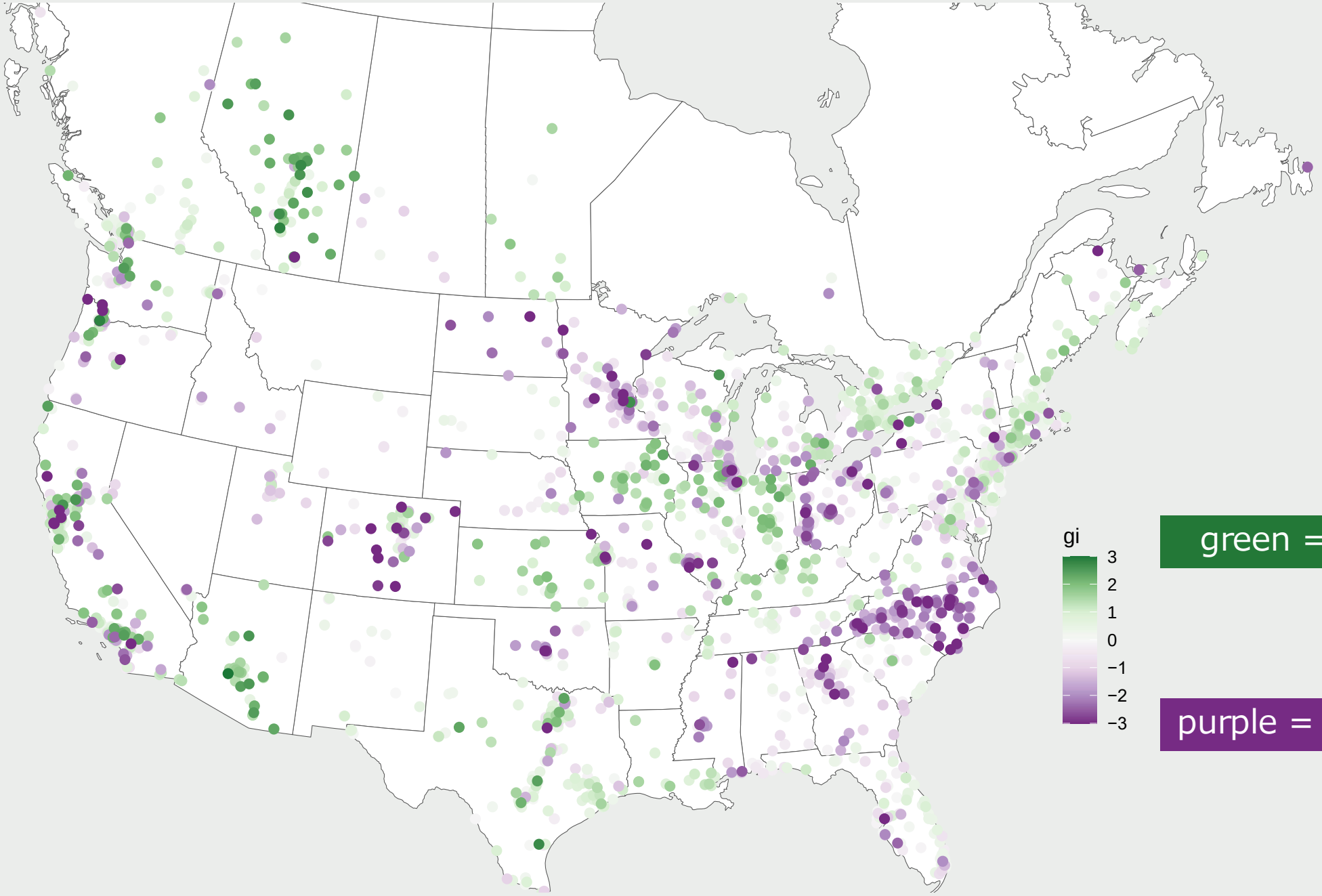
green = noun premodification

purple = noun post-modification



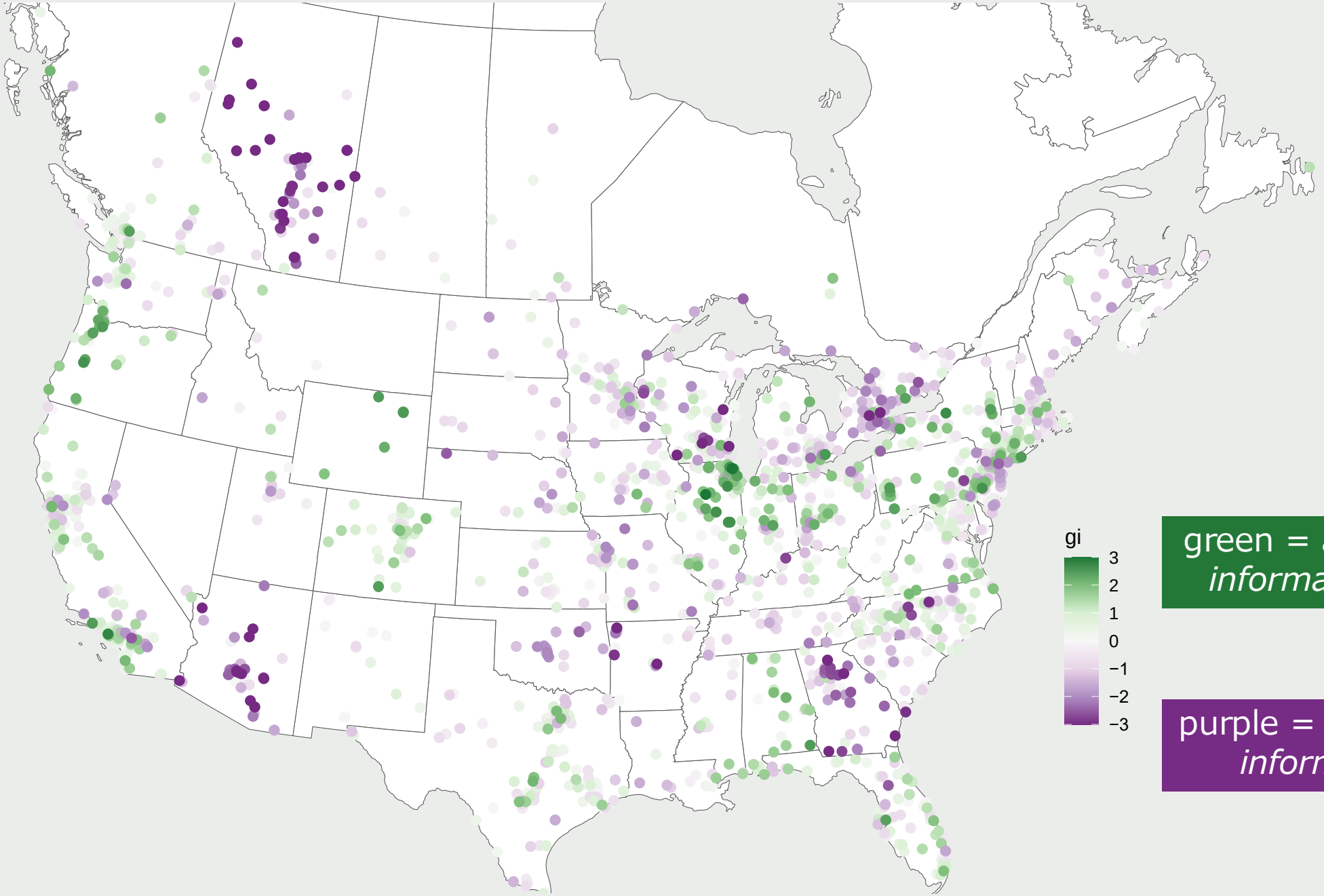
green = noun

purple = personal pronoun



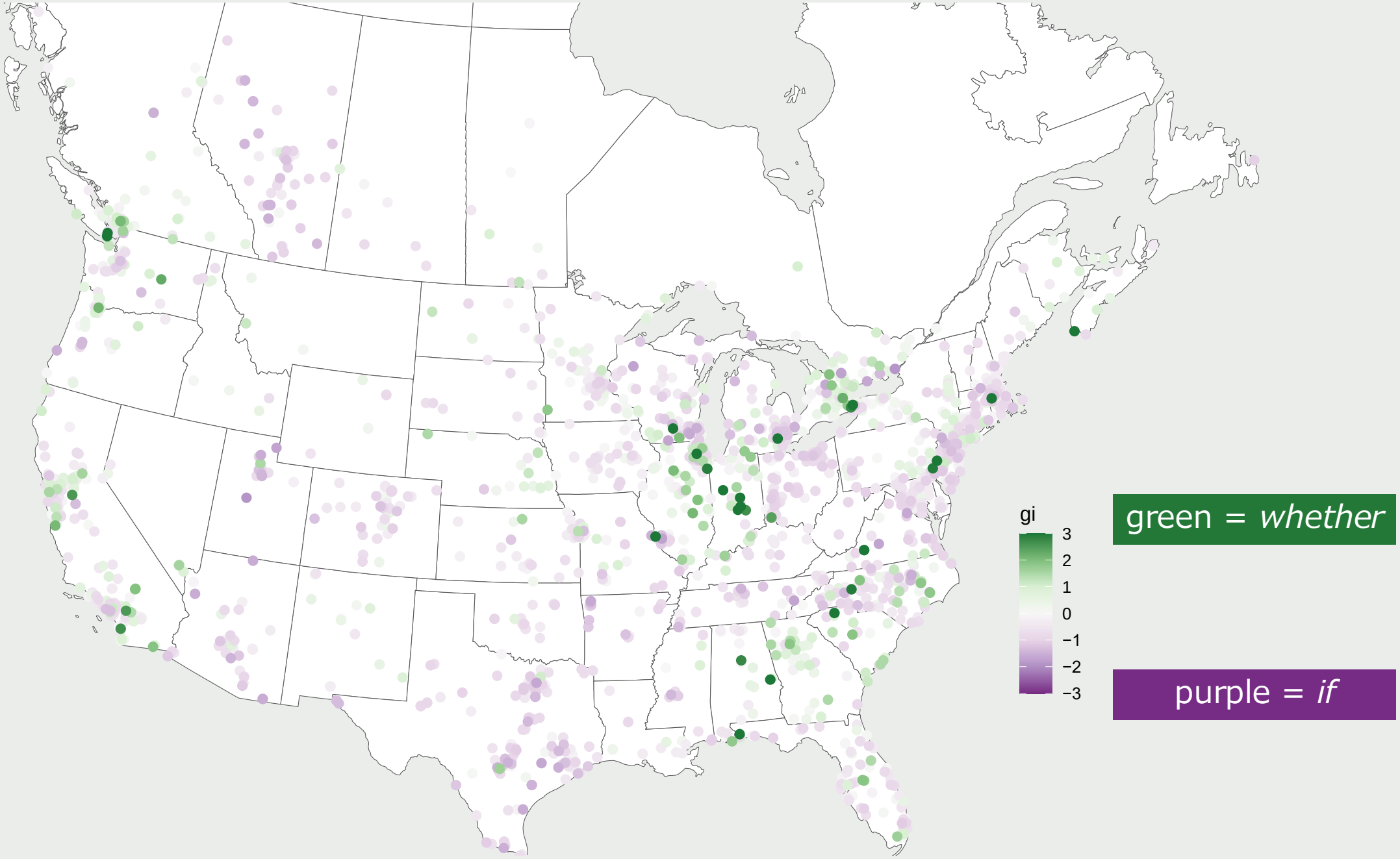
green = *compare to*

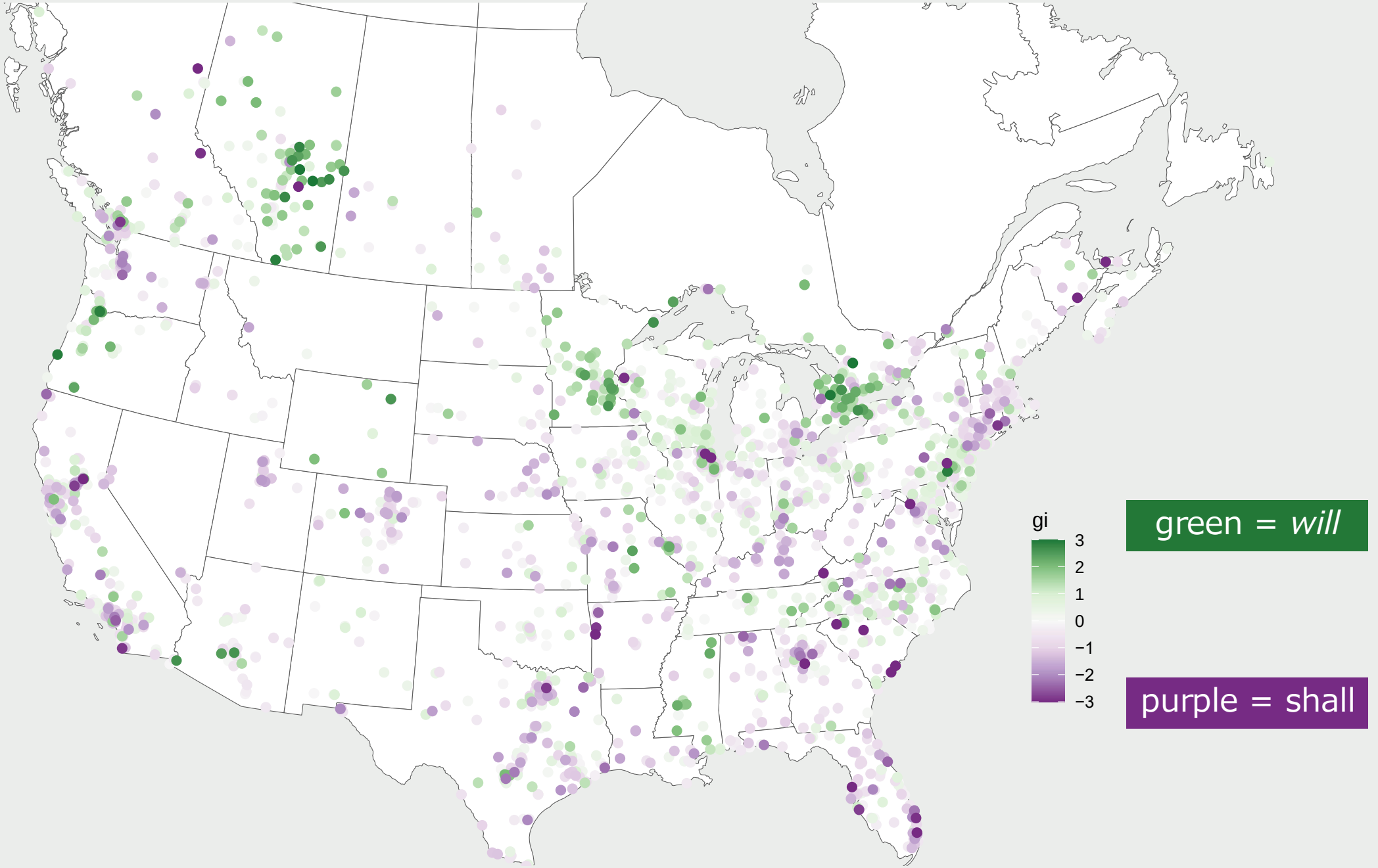
purple = *compare with*

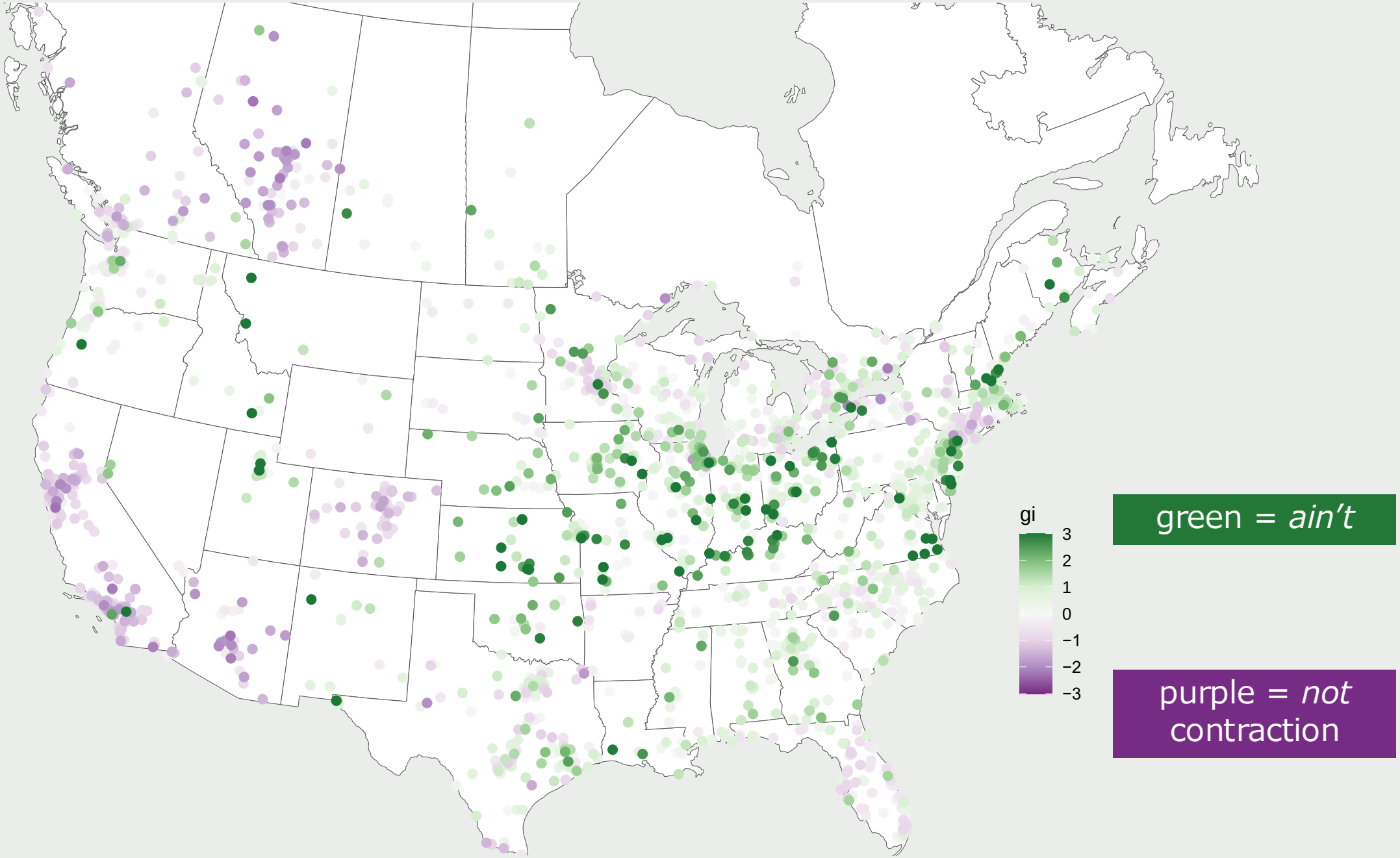


green = *article/book/ information about*

purple = *article/book/ information on*

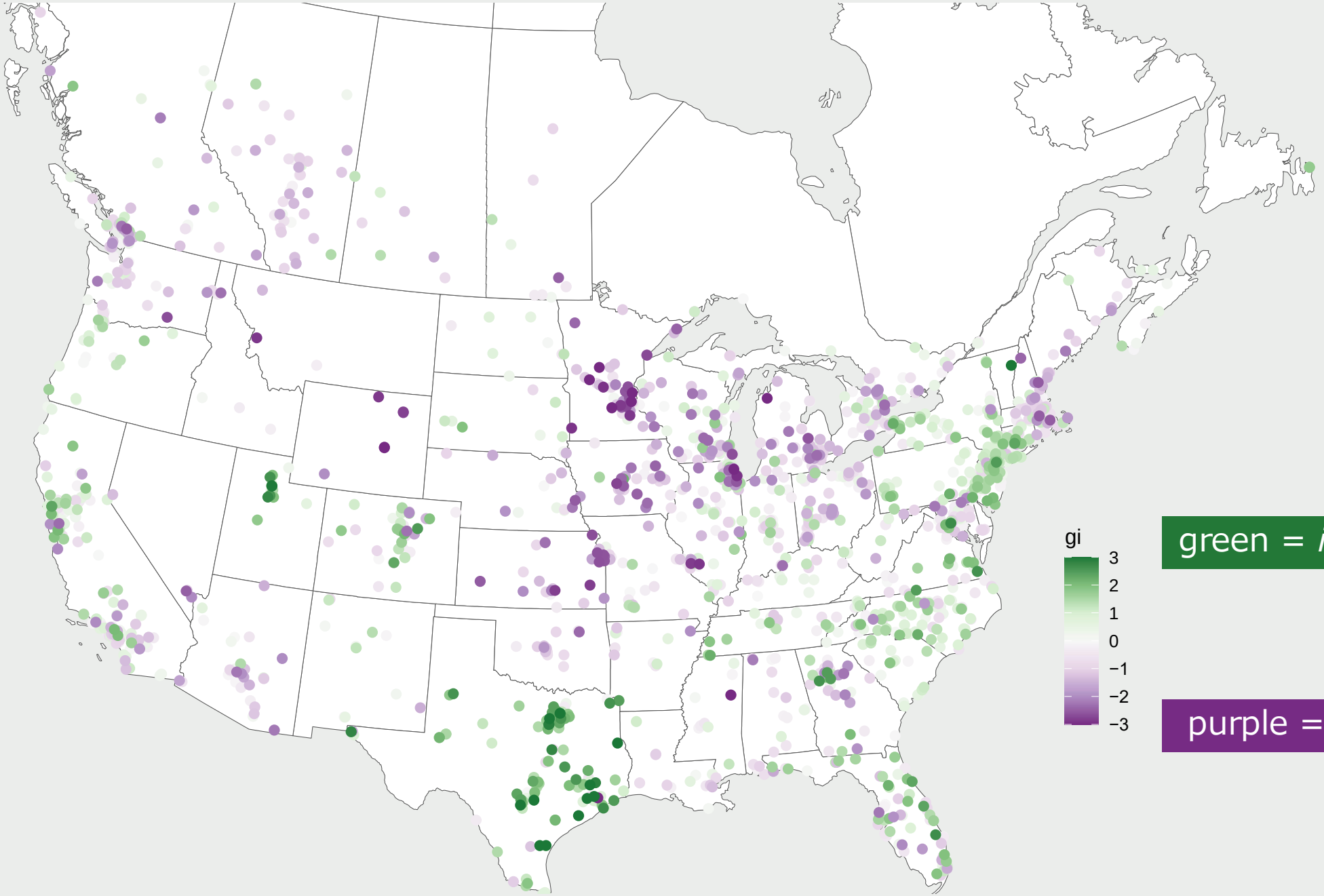

gi
3
2
1
0
−1
−2
−3
green = *whether*
purple = *if*


gi
3
2
1
0
−1
−2
−3
green = *will*
purple = shall



green = *ain't*

purple = *not* contraction



green = *immediately*

purple = *right away*

Conclusions

- Many features align nicely with known dialectal isoglosses
- Many features are highly interpretable
- Many reveal interesting new geographic patterns
- Future research
  - Improve accuracy of features
  - Multivariate analyses
  - Additional features
  - Sample underrepresented regions

THANK YOU

Download these slides at
joeystanley.com/aacl2024

Brett Hashimoto, BYU,
brett_hashimoto@byu.edu

Joseph A. Stanley, BYU,
joey_Stanley@byu.edu

Jack Grieve, University of Birmingham,
j.grieve@bham.ac.uk