How America Talks: Quantifying and Predicting Dialect Variation Across the Continental United States

Christine Lee, Chris Yang, Nathanael Choi, and Julie Hohenberg

Abstract

American English exhibits regional variation, but the extent to which current dialect patterns align with geography remains an open question. This project analyzes decades of lexical survey collected by Dr. Bert Vaux, comprising nearly five million responses to 108 questions from over 47,000 participants across the continental United States. We combine statistical analysis, supervised and unsupervised machine learning, and interactive data visualization on a dashboard[1] to quantify dialect variation and evaluate the predictability of regional identity. Using chi-squared tests and Cramér's V, we find that age shows a consistently stronger association with lexical choice than geographic region, suggesting that generational change plays a major role in contemporary dialect usage. Supervised classification models trained on one-hot encoded lexical features outperform random and majority baselines, with XGBoost achieving the strongest performance at 59% accuracy across seven dialect regions. In contrast, K-Means clustering exhibits weak alignment with established dialect boundaries, indicating that dialects form gradual and overlapping regions rather than discrete clusters. We further apply SHAP-based feature attribution to interpret the XGBoost model, and use interpretable logistic regression feature rankings to construct a region prediction quiz hosted on our dashboard. Our results show that American dialects are measurable yet increasingly blended due to broader patterns of linguistic convergence. Our quantitative analysis and public interactive dashboard provide an accessible framework for exploring linguistic diversity in the continental United States.

1 Introduction

Our project explores how English dialects vary across the United States and how people communicate the same ideas using different regional vocabularies and grammatical patterns. We used a large dialect dataset collected by Professor Bert Vaux, consisting of thousands of survey participants, over 160 questions, numerous answer choices across these questions, and more than 11 million responses.

Clustering and predictive modeling were both conducted during this project. K-Means clustering showed very low alignment with true geographic regions, indicating that dialect-based patterns do not naturally recreate formal regional boundaries. Logistic regression predictions performed moderately well, with precision, recall, and F1-scores varying by region, with an overall accuracy of around 0.57. Despite this, the highest performance came from XGBoost with an accuracy of 0.59. This was what we used for the regional dialect prediction quiz. We also built an interactive dashboard that summarizes the dataset, provides visualizations, produces graphs, and includes the regional dialect prediction quiz, allowing users to see which region their own dialect most resembles.

Overall, our work highlights how language reflects culture, history, and identity, helping people appreciate everyday linguistic diversity. These findings can support educators, researchers, and organizations by revealing how dialects shift across regions and by offering insights for communication, cultural analysis, and future linguistic studies.

[1] Dashboard Link: https://dialectsofenglishcapstone2025.streamlit.app/

2 Prior Work

Prior work on American English dialects has taken several different approaches. Some studies focus on simply documenting how people across the United States actually speak, recording differences in sentence structure, word choice, and pronunciation (Axelrod and Scheibman, 2013). These studies show that features like saying *y'all*, using phrases such as *fixin' to*, combining multiple helping verbs (e.g., *might could*), or pronouncing vowels differently are often linked to region, even though their usage can vary depending on context.

Other work uses large surveys that ask people what words they use for everyday objects or ideas (Boberg, 2005). These studies focus on identifying which words are most useful for differentiating regions by measuring how unevenly a term is used across the country and how clearly it separates one area from another. Together, this research shows that while many dialect features are widespread, some specific vocabulary choices are especially informative for finding regional patterns.

Research on dialect perception suggests that people are not very good at identifying a speaker's region based on language alone. Early studies found that listeners often disagreed with one another when making regional or social judgments from speech samples (Lee, 1971). Later experiments showed only moderate success even when listeners were forced to choose from a fixed set of regions (Clopper et al., 2006). These results suggest that people tend to rely on general impressions, such as whether a voice sounds familiar or broadly regional, rather than on specific linguistic features.

More recent work has used larger datasets and computational methods to study regional English in a more systematic way. Projects such as the Yale Grammatical Diversity Project (Zanuttini et al., 2018) show that differences in sentence structure can be mapped geographically using large-scale surveys, revealing overlapping and non-discrete regional patterns. Other studies use statistical and machine learning techniques to identify dialect patterns across many linguistic variables (Grieve et al., 2011; Evanini, 2010; Hedges, 2017). These approaches demonstrate that regional distinctions can be detected in data, even when they are subtle and highly variable.

Taken together, this research shows that regional features of American English do exist, but they do not consistently align with clear geographic boundaries and are not always easy to understand or model. Our work builds on this foundation by applying clustering, geographic analysis, and supervised learning to a large lexical survey dataset. In doing so, we evaluate how reliably regional identity can be deduced from vocabulary choices and use these results to create an interactive region-prediction quiz that makes these patterns accessible at scale.

3 Data

The dataset originates from Dr. Bert Vaux, Professor of Phonology and Morphology at the University of Cambridge, who has been conducting dialect surveys since the mid 1990s. He began collecting them on paper before later migrating them to an online format with assistance from computer science students. Users in the survey were asked a question and had to choose one choice for their response (Figure 1). These data were stored in a regional database, and later the data were extracted into the following four flat CSV files shown in Table 1.

CSV File	Samples (rows)	Features (cols)
questions	181	7
choices	1484	6
users	385780	20
responses	11853462	7

Table 1: Dataset overview: number of samples and features in each CSV file prior to preprocessing.

Each CSV file contains identification fields stored as integers to link samples between files, text-based fields for objects like name and open-ended responses, boolean indicators encoded as integers, and float fields for geographic coordinates.

Prior to analysis, we performed several preprocessing steps. First, we deduplicated survey questions that appeared in multiple survey versions by mapping shared question IDs across *questions, choices,*

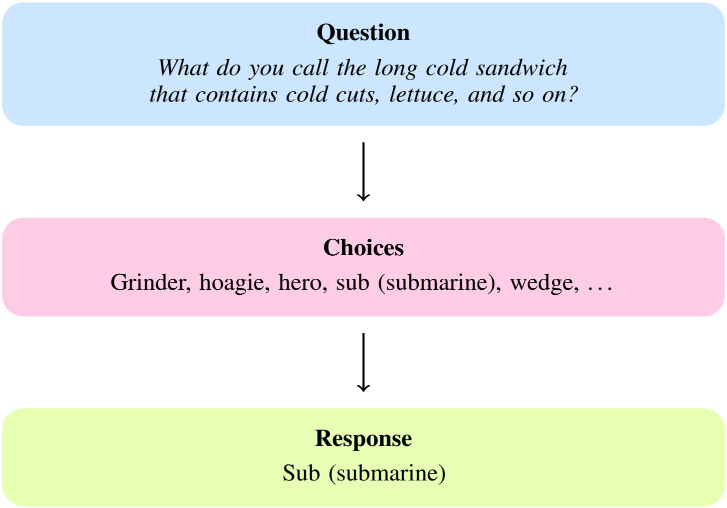

Question
What do you call the long cold sandwich that contains cold cuts, lettuce, and so on?
Choices
Grinder, hoagie, hero, sub (submarine), wedge, ...
Response
Sub (submarine)

Figure 1: Example of a dialect-survey question, available choices, and an individual user's response.

and *responses*. Additionally, we restricted the dataset to users whose reported latitude and longitude fell within the continental United States.

We restricted our analysis to questions that received between 40,000 and 60,000 responses. A histogram of response counts (Figure 2) showed that this range represented the majority of questions in the dataset, allowing us to retain the most consistently answered items while removing outliers with unusually low or high response rates. The full list of the 108 retained survey questions can be found in **Appendix A**.

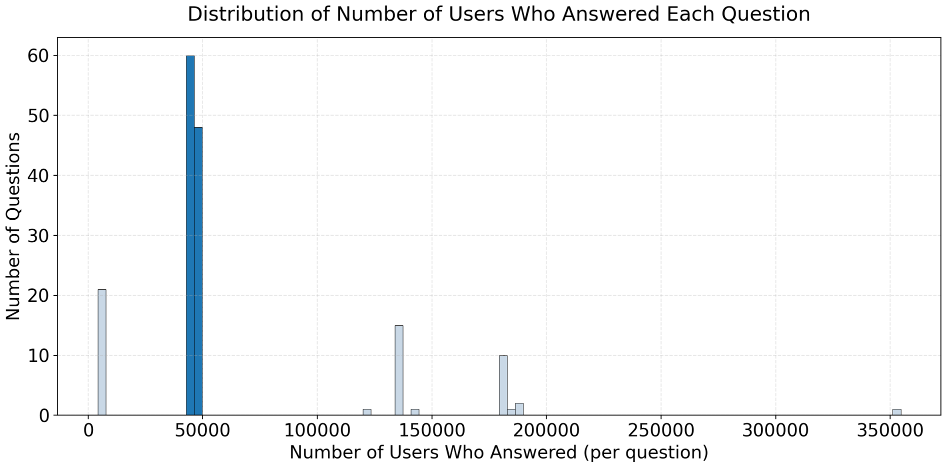


Figure 2: Histogram of user response counts per question. Only the questions in the 40-60k response interval(highlighted) were retained for modeling and analysis.

Ground-truth regions were assigned to each user based on dialect region boundaries defined by Aschmann (2014), which draw directly from Labov et al. (2006). Distribution of users across dialect regions is shown in Figure 3.

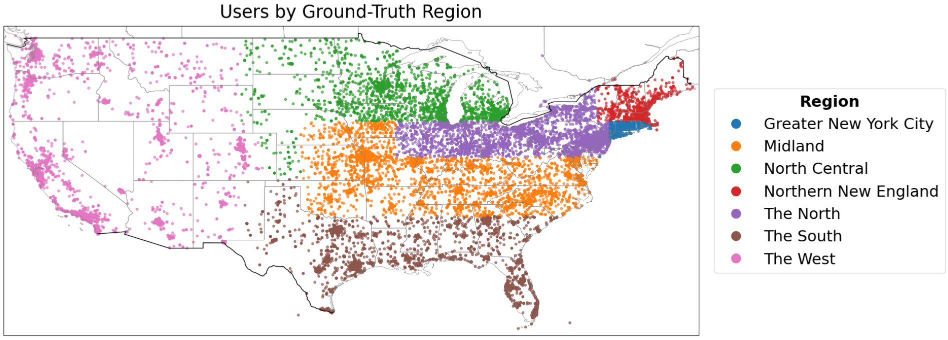


Figure 3: Post-Processing geographic distribution of users by ground-truth dialect region; Greater New York City, Northern New England, The North, Midland, and North Central are most strongly represented. The West is mainly represented by the West Coast.

After merging the four CSVs and applying all preprocessing steps, the final dataset contains nearly five million responses to more than 100 questions from over 47,000 unique users (Table 2).

Statistic	Value
Unique Questions	108
Choices per Question (min–max)	2 – 21
Choices per Question (mean)	5.16
Total Responses	4,973,285
Responses per Question (min–max)	43,387 – 47,034
Responses per Question (mean)	46,048.9
Unique Users in Final Dataset	47,218

Table 2: Dataset statistics after preprocessing and filtering.

4 Statistical Analysis

4.1 Techniques

For our statistical analysis we employed the chi-square test of independence to compare age groups and regions against dialect term usage and determine if they are statistically associated. This test evaluates whether the observed distribution of categorical variables differs from what can be expected if the two were completely independent.

The chi-square test works by first creating a contingency table showing the combinations, for example, how many people born within a certain age range respond with "roly poly" vs "pillbug." Then, it calculates the expected frequencies for each example to determine if the example was independent, which assumes that age and word choice are unrelated. The test computes the chi-square statistic using the formula

$$\chi^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{1}$$

summed across all the examples in the contingency table. The chi-square value is then compared against a theoretical chi-square distribution to determine the p-value, which represents the probability of getting the observed data in relation to the assumption that there is no relationship. So, a lower p-value score ($p < 0.05$) indicates statistical significant association, meaning we can conclude that age group and dialect choice are not independent.

Following the chi-square, we calculated Cramér's V as a standardized effect size measure. Cramér's V provides a measure of association strength ranging from 0 (no association) to 1 (perfect association).

This normalization allows meaningful comparison of effect sizes across different questions with different numbers of response options and age groups. We interpreted Cramér's V using these thresholds: negligible ($V < 0.10$), small effect ($0.10 \leq V < 0.30$), medium effect ($0.30 \leq V < 0.50$), and large effect ($V \geq 0.50$). Doing this approach with chi-square and Cramér's V identifies differences that are both statistically reliable and meaningful for understanding generational dialect usage.

4.2 Data Manipulation

The changes we made to our data primarily involved text normalization, as most answers had a variety of spellings, which were standardized to the most common form. For example, "roly poly", "rolly polly", and "rollie pollie" were changed to just "roly poly". Furthermore, some answers were just very long and were shortened to have the same meaning. Lastly, some questions had answers in different variations but shared the same meaning, such as "sub" and "sub (submarine)", which were merged into just "sub (submarine)". Then for our geographic analysis, the dataset was filtered to include only United States respondents and were classified into one of the nine US dialect regions based on their latitude and longitude coordinates. This regional dataset was merged back into the main dataset, and responses with missing geographic information were excluded. For our temporal analysis, the dataset was just split into two birth groups: users born before 1975 and those born after 1975.

4.3 Results

We evaluated our results using the techniques mentioned, such as the chi-square test, p-value, and Cramér's V score, on the 5 questions that had the most responses to ensure that our results had enough data. These questions were:

- What do you call gooey or dry matter that collects in the corners of your eyes?
- What do you call a small creature that rolls into a ball when you touch it?
- How do you pronounce the vowel sound in the word aunt?
- What is along cold sandwich with cold cuts and so on?
- What do you call the king of rain when the sun is shining?

4.3.1 Age groups

For all five top questions examining the relationship between age and dialect usage, we observed very large chi-square values, the lowest being 76,092. This indicates a very strong statistical association between age group and their word choice. The p-values for all questions were extremely small ($p < 0.0001$) showing that these patterns are also statistically significant and highly unlikely to occur by chance. All of the Cramér's V values fell within the medium effect size range $(0.3 \leq V < 0.5)$. This moderate effect size does confirm that age is a meaningful predictor of dialect variation, but it also suggests that there can be other factors that may explain why dialect usage can change beyond just generational differences.

Question	**Chi-square**	**p-value**	**Cramér's V**
Eye matter	134,222.75	<0.0001	0.369 (Medium)
Rolled up bug	79,231.26	<0.0001	0.420 (Medium)
Aunt pronunciation	119,202.50	<0.0001	0.418 (Medium)
Cold cut sandwich	126,545.27	<0.0001	0.403 (Medium)
Rain in sun	76,092.30	<0.0001	0.383 (Medium)

Table 3: Chi-square analysis results for age and dialect usage across five questions. All results show medium effect sizes, indicating age is a meaningful predictor of dialect variation.

4.3.2 Regions

Then, we deployed the same techniques when comparing regions to dialect term usage. For the same 5 questions, we observed chi-square values ranging from 3,031 to 7,041 with very low p-values ($p < 0.0001$), indicating that regional differences in word choice are statistically significant and not due to just chance. However, the Cramér's V value has a different story compared to the age analysis. The questions about the sandwich term, eye matter, and aunt pronunciation showed negligible effect size, while the other two questions showed small effect sizes. These extremely low effect sizes compared to the age analysis suggest that geographic region is a weaker predictor of dialect variation than age.

Question	Chi-square	p-value	Cramér's V
Eye matter	3,080.24	<0.0001	0.070 (Negligible)
Rolled up bug	5,113.95	<0.0001	0.146 (Small)
Aunt pronunciation	3,031.41	<0.0001	0.092 (Negligible)
Cold cut sandwich	3,559.63	<0.0001	0.092 (Negligible)
Rain in sun	7,040.70	<0.0001	0.174 (Small)

Table 4: Chi-square analysis results for region and dialect usage across five questions. Effect sizes are substantially smaller than age analysis, suggesting region is a weaker predictor of dialect variation.

4.3.3 Conclusion

While both factors show statistically significant associations with dialect usage ($p < 0.001$), the effect sizes differ dramatically. Age consistently displays a medium effect size, while region shows a negligible to small effect size. This gap indicates that generational shift could be the primary mechanism driving dialect variation in America. The findings suggest that dialect boundaries are becoming less geographically bound, reflecting broader patterns in linguistic change due to mobility, media, and wide ranges of communication. This comparison points to the idea that the type of English someone speaks depends on "when you were born" rather than "where you're from".

5 Machine Learning

To prepare the survey data for modeling, we transformed all user responses into a high-dimensional one-hot encoded matrix aligned across all users and all possible answer choices. Each feature corresponds to a unique (question, answer) pair drawn from the preprocessed dataset. If a user did not answer a particular question, all features corresponding to that question's answer choices were set to 0. This creates a "missing" category and keeps the dimensional structure fixed, which is necessary for linear models. The dataset contains 47,128 users, each represented by 108 questions which were encoded into 665 question–choice features for each user. An example of this is shown in Table 5. Both supervised and unsupervised machine learning techniques were conducted using this dataset with different task-specific preprocessing steps applied.

User ID	243_MISSING	243_[o:]	243_[u:]	Region
1126637	1	0	0	West
1126639	0	1	0	Northern New England

Table 5: Example of one-hot encoded inputs for Question 243 *(the first vowel in "Bowie knife")* showing a subset of response categories. Ground-truth region labels are excluded from training and used only for evaluation; user IDs are likewise omitted from training and retained only for record-keeping.

5.1 Region Prediction Quiz

Inspired by the *New York Times* region-prediction quiz (Katz and Andrews, 2024), we aimed to create a shorter and more effective quiz using ML techniques. Rather than relying on the full set of 108 questions, our goal was to determine whether a small group of especially diagnostic lexical variables could approximate a user's dialect region with reasonable accuracy.

We trained and tested five classification models on the question-answer choice features, sampled in Table 5. The data was split into 80% for training and 20% for testing. Classifier models and their hyperparameters are shown in Table 6.

Model	**Hyperparameters**
Logistic Regression	multinomial; `lbfgs`; `max_iter`=1000
Random Forest	300 trees; unlimited depth; `n_jobs`=-1; `random_state`=42
Gradient Boosting	300 estimators; learning rate = 0.05; max depth = 3; subsample = 0.8
Linear SVM	`C`=1.0; class weight = balanced; `max_iter`=5000

XGBoost	300 trees; learning rate = 0.05; max depth = 6; subsample = 0.8;
	colsample_bytree = 0.8; objective = `multi:softprob`

Table 6: Supervised learning models and their hyperparameters used for dialect region classification.

Majority and random baselines were also used for comparison. The majority baseline predicts the most common region for all users (The North), which provides a benchmark for class imbalance. The random baseline assigns regions by sampling from the training-set class distribution, simulating a random guess. Results are shown in Table 7.

Model	**Precision**	**Recall**	**F1-Score**	**Accuracy**
Random Baseline	–	–	–	0.17
Majority Baseline	–	–	–	0.24
Random Forest	0.54	0.49	0.51	0.53
Logistic Regression	0.57	0.54	0.55	0.57
Gradient Boosting	0.59	0.55	0.57	0.58
Linear SVM	0.53	**0.56**	0.54	0.55
XGBoost	**0.61**	**0.56**	**0.58**	**0.59**

Table 7: Macro-averaged supervised classification performance, with random and majority baselines for reference. Strongest number for each metric in bold. All models outperformed baselines. XGBoost achieves the highest precision, recall, F1, and accuracy.

All models achieved higher accuracy than both baselines, but XGBoost performed the strongest, with an accuracy of 0.59. It also maintained the strongest precision (0.61) and recall (0.56) across regions relative to the other models. Table 8 shows detailed performance metrics for XGBoost across regions.

Region	**Precision**	**Recall**	**F1-Score**
Greater New York City	0.58	0.45	0.51
Midland	0.57	0.59	0.58
North Central	0.66	0.59	0.63
Northern New England	**0.70**	0.51	0.59
The North	0.53	0.65	0.58
The South	0.54	0.46	0.50
The West	0.65	**0.66**	**0.66**
Overall Accuracy		0.59	

Table 8: Per-region performance of the XGBoost classifier. Strongest number in each metric in bold. XGBoost performs best in the West (highest recall and F1), while Northern New England shows the highest precision, but lower recall.

Once XGBoost was determined to be the strongest model, SHAP (Shapley Additive Explanations) was used to find prominent lexical variables for each region, as determined by XGBoost. Results are shown in Table 9. SHAP assigns each feature a value indicating how much it contributes to a model's prediction. For a model prediction $f(x)$ and feature i, the SHAP value is shown below:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|! \, (|F| - |S| - 1)!}{|F|!} \left[f(x_{S \cup \{i\}}) - f(x_S) \right],$$

F is the full set of features, S is any subset of features not including i, $f(x_S)$ is the model prediction using only the features in S, and ϕ_i is the SHAP value quantifying feature i's contribution.

Region	**Most Predictive Feature**	**Mean** $|$SHAP$|$
Greater New York City	*Q255:* pronouncing Mary, merry, and marry all differently (Mary/merry/marry)	0.2370
Midland	*Q305:* calling the glowing summer insect a "firefly"	0.2076
North Central	*Q316:* calling something diagonally across the street "kitty-corner"	0.2828
Northern New England	*Q266:* pronouncing "route" to rhyme with "hoot"	0.1256
The North	*Q305:* calling the glowing summer insect a "firefly"	0.1669
The South	*Q343:* calling the school drinking fixture a "water fountain"	0.1448
The West	*Q319:* using "highway" as the general term for a big, fast road	0.2938

Table 9: Top region-distinguishing lexical variables according to mean absolute SHAP value.

Ten quiz questions were chosen and are shown in **Appendix B**. These were selected from the largest logistic regression coefficients, which offer a clearer, more interpretable ranking of discriminative features. Although the quiz itself uses the XGBoost model as its backend, SHAP values diffuse importance across many interactions, making them less suitable for isolating a small, human-readable set of questions.

5.2 Clustering

In addition to supervised learning, we also carried out unsupervised learning to determine if the features would naturally cluster by region. K-Means with Euclidean distance was selected as the clustering algorithm because it is well-suited for high-dimensional and sparse data like our one-hot encoded survey matrix. The one-hot encoded features (Table 5) were scaled using Z-score scaling for clustering so that each answer category contributed equally to Euclidean-distance calculations.

The K-Means algorithm partitions the dataset into k clusters by minimizing the total within-cluster variance. Its objective function is shown below, where μ_j is the centroid of cluster C_j:

$$\min_{\{C_1,\ldots,C_k\}} \sum_{j=1}^{k} \sum_{x_i \in C_j} \|x_i - \mu_j\|^2,$$

K-Means clustering was evaluated for $k \in \{3, 4, 5, 6, 7\}$. Two different feature representations were used: (1) the full scaled one-hot encoded response matrix and (2) a reduced set of the top 50 most region-predictive features from Logistic Regression in section 5.1 (see **Appendix C** for top-50 question list). Performance was assessed using the silhouette score, Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI).

The silhouette score measures within-cluster cohesion relative to between-cluster separation. Values close to 1 indicate well-separated clusters, while values near 0 indicate overlapping clusters. Negative values imply that points may be assigned to the wrong cluster. It is defined below, where $a(i)$ is the average intra-cluster distance and $b(i)$ is the nearest-cluster distance:

$$\text{Silhouette}(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}},$$

The ARI evaluates agreement between cluster assignments and ground-truth dialect regions, correcting for chance. ARI ranges from 0 (random labeling) to 1 (perfect recovery), with negative values indicating worse-than-chance alignment. It is defined below, where RI is Rand Index (unadjusted):

$$\text{ARI} = \frac{\text{RI} - \mathbb{E}[\text{RI}]}{\max(\text{RI}) - \mathbb{E}[\text{RI}]},$$

The NMI measures shared information between the clusters and true labels. NMI ranges from 0 (no shared information) to 1 (perfect correspondence). It is defined below, where C is the predicted cluster assignment, R is the ground-truth region label, $I(C; R)$ is the mutual information between I and R, and $H(\cdot)$ denotes entropy:

$$\text{NMI}(C, R) = \frac{2\,I(C; R)}{H(C) + H(R)}.$$

k	Silhouette (Full)	ARI (Full)	NMI (Full)	Silhouette (Top)	ARI (Top)	NMI (Top)
3	0.0235	0.0228	0.0377	0.0504	0.0373	0.0574
4	0.0233	0.0190	0.0347	0.0185	0.0931	0.1251
5	0.0117	0.0736	0.1009	0.0274	0.0907	0.1207
6	0.0130	0.0543	0.0823	−0.0399	0.0841	0.1159
7	0.0162	0.0711	0.1133	0.0140	0.0821	0.1155

Table 10: Comparison of K-Means clustering performance on the full feature set versus the top region-predictive features, evaluated using Silhouette, ARI, and NMI.

Across all values of k, both the full feature set and top-feature subset produce very weak clustering structure. The Silhouette score ranges from approximately -0.04 to 0.05, which indicates that there is no meaningful geometric cluster structure among the encoded questions and answers. Similarly, the ARI ranges from approximately 0.02 to 0.09, indicating that the performance was equivalent to random assignment. Lastly, NMI ranged from 0.03 to 0.11 for the full feature clustering and ranged from 0.05 to 0.13 for the top-feature clustering. Although these values indicate that clusters share minimal information with the true dialect regions, the top-feature subset shows slightly higher NMI.

The strongest clustering solution was $k = 5$ using the top-question feature subset. It included one of the strongest ARI (0.09) and NMI (0.12) values, and maintained a competitive silhouette score (0.027) relative to other k-values. The number of users in each region in these clusters is shown in Table 10, and a visual distribution of cluster users is shown in Figure 4. Clusters 1 and 3 are the largest and dominate the map, with Cluster 1 dominating the North Central and West regions and Cluster 3 dominating the Midland, South, North, Greater New York City, and Northern New England regions.

Region	C0	C1	C2	C3	C4	Total Users
Greater New York City	53	442	125	384	2,547	3,551
Midland	118	2,530	176	5,328	845	8,997
North Central	90	4,674	198	1,100	526	6,588
Northern New England	43	458	103	378	1528	2,510
The North	173	3,913	309	3,963	3,136	11,494
The South	458	1,138	154	3,186	637	5,573
The West	120	6,037	253	1,405	690	8,505
Cluster Totals	**1,055**	**19,192**	**1,318**	**15,744**	**9,909**	**47,218**

Table 11: Region-by-cluster distribution for K-Means with $k = 5$ using the top region-predictive features. The final column reports total users per region (not used in clustering); the bottom row shows total users per cluster.

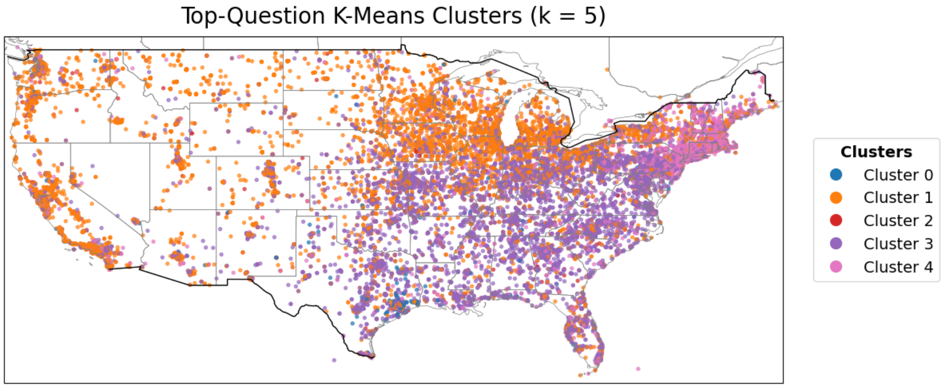


Figure 4: Geographic distribution of K-Means clusters for $k = 5$ using the top region-predictive features. Colors indicate cluster users within each dialect region.

The overall weak clustering alignment with established dialect boundaries indicates that dialects form gradual and overlapping regions rather than discrete clusters. This reflects the limitations of one-hot lexical representations and suggests that alternative feature representations may better capture these patterns.

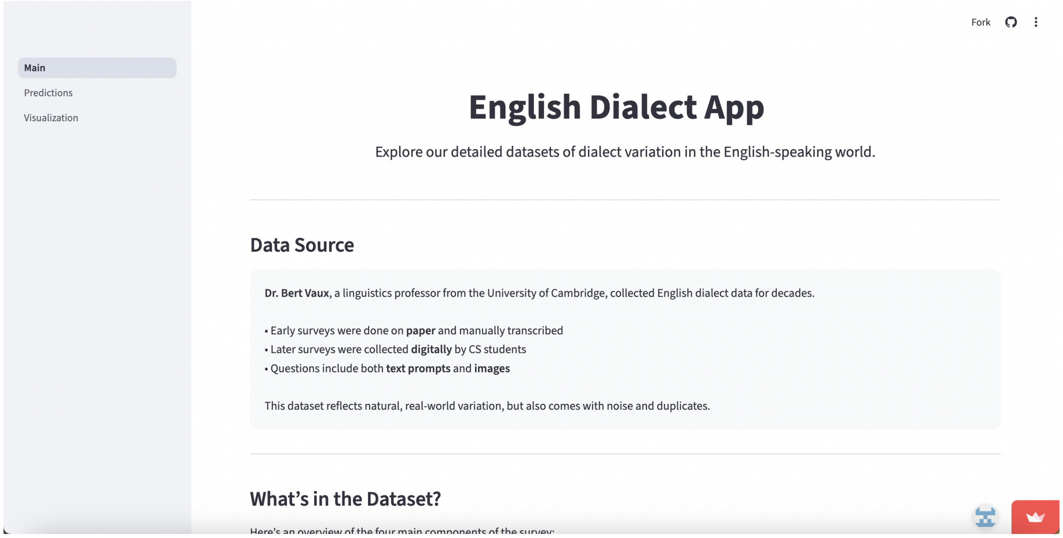
6 Data Communication

To communicate our findings clearly and give users an interactive way to explore American dialect patterns, we created a Streamlit dashboard that summarizes our dataset, visualizes key trends, and includes a prediction quiz, allowing users to see which region their speech most closely aligns with.

The dashboard helps users living in the United States understand how language reflects culture, history, and identity while highlighting familiar variations in everyday speech. They will be able to discover how language can change throughout time between different age groups and regions of the United States. In addition, users will be able to take our prediction quiz to determine if their choice of speech can choose which region of America they are from. Using GitHub for version control, we also used Google Drive for hosting data, Streamlit Cloud for deployment, and the following Python packages: streamlit, pandas, numpy, plotly, gdown, re, and pathlib. Using our clean datasets, users will be able to interact with our dashboard's visualizations and prediction tools, which can be accessed through a simple user interface. See Figure 5, 6, 7, and 8 for pictures from the dashboard.

Our dashboard provides advantages for students and researchers by increasing linguistic awareness, providing the ability to examine how dialects change through time, and allowing groups to adjust how they communicate within a regionally based audience. The dashboard will be demonstrated during our poster presentation, and it is publicly available here (must be in light mode, can be changed in settings), with the full codebase on GitHub[2].

[2] https://github.com/christin3l33/dialects-of-english-capstone-2025



English Dialect App

Explore our detailed datasets of dialect variation in the English-speaking world.

Data Source

Dr. Bert Vaux, a linguistics professor from the University of Cambridge, collected English dialect data for decades.

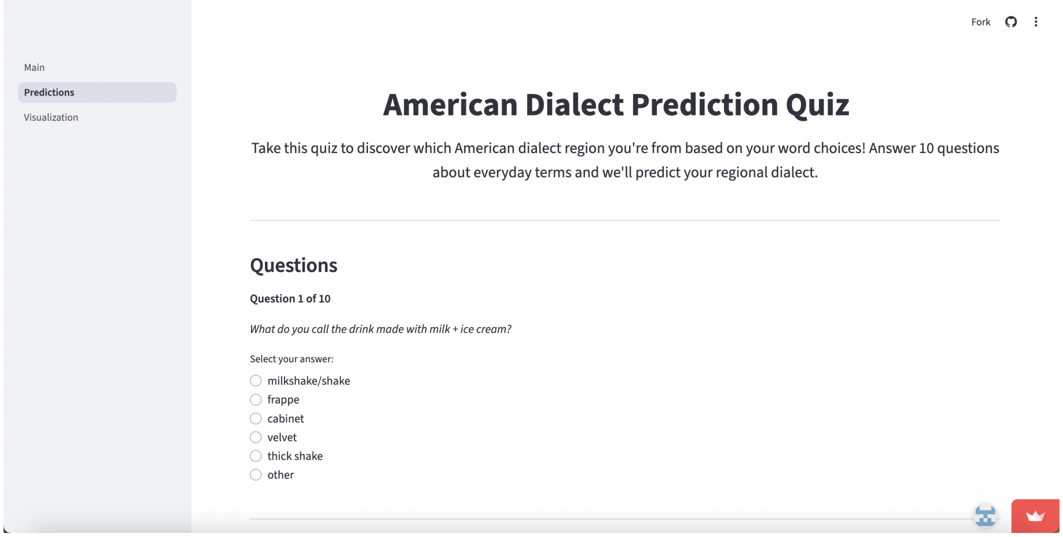
- Early surveys were done on **paper** and manually transcribed
- Later surveys were collected **digitally** by CS students
- Questions include both **text prompts** and **images**

This dataset reflects natural, real-world variation, but also comes with noise and duplicates.

What's in the Dataset?

Here's an overview of the four main components of the survey:

Figure 5: Main page of the dashboard that provides basic information about our data.



American Dialect Prediction Quiz

Take this quiz to discover which American dialect region you're from based on your word choices! Answer 10 questions about everyday terms and we'll predict your regional dialect.

Questions

Question 1 of 10

What do you call the drink made with milk + ice cream?

Select your answer:

○ milkshake/shake
○ frappe
○ cabinet
○ velvet
○ thick shake
○ other

Figure 6: Prediction page of the dashboard that allows users to take and determine their regional dialect.

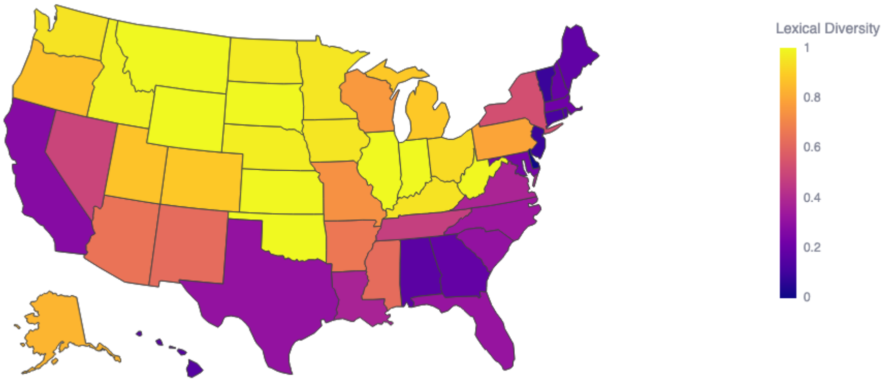


Figure 7: Lexical Diversity (Shannon Entropy) by U.S. State — Soda vs. Pop on the visualization page of the dashboard. Colors indicate lexical diversity for each state.

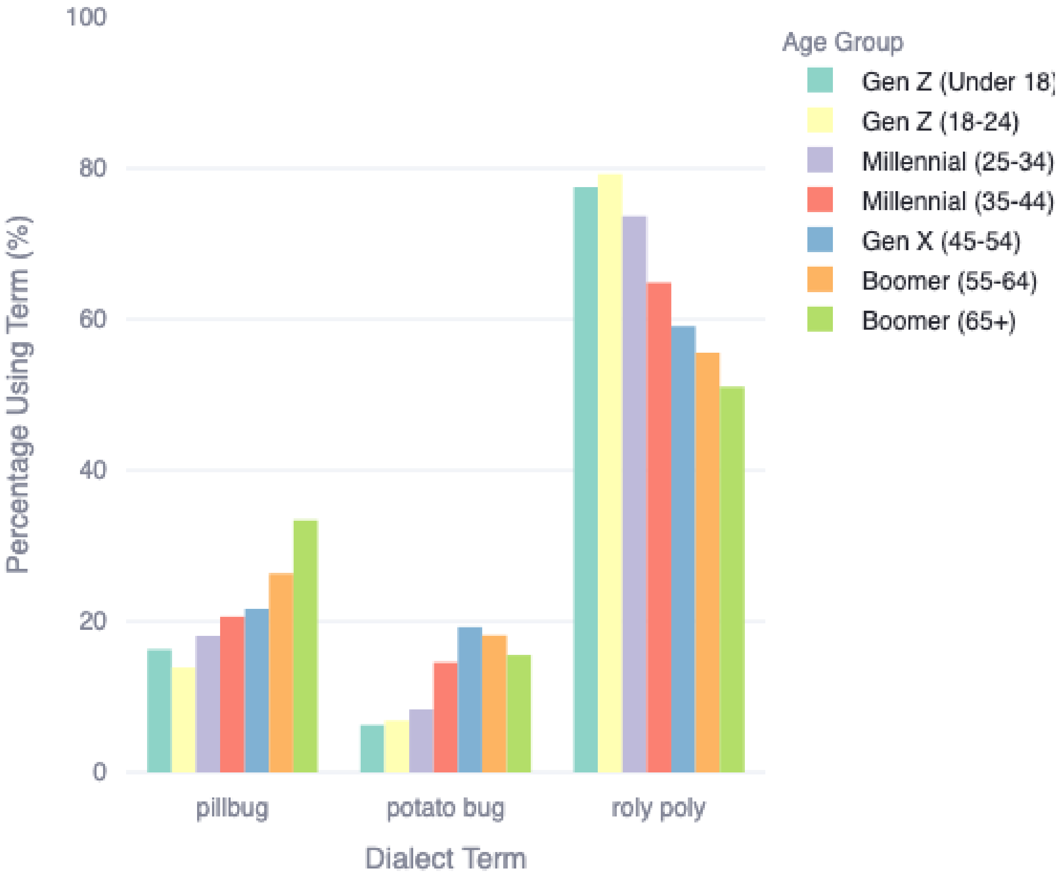


Figure 8: Roly Poly Question bar graph of usage by age group on the visualization page of the dashboard. Colors indicate different age groups.

7 Conclusions and Future Work

Our project set out to explore how English dialects vary across the United States and to examine whether regional identity can be inferred from patterns in everyday vocabulary. Using a large survey dataset with nearly five million responses, we cleaned, merged, and standardized data from multiple sources, and then applied statistical analysis, machine learning, and interactive visualization through a Streamlit dashboard. Together, these all paint a picture of how language reflects regional identity.

From a statistical perspective, our chi-square tests and Cramér's V analyses revealed clear generational differences in dialect usage, showing that age plays a meaningful role in vocabulary choice. In contrast, our unsupervised clustering results showed very weak structure, as k-means clusters did not align well with established dialect regions. This reinforced the idea that dialect boundaries are gradual and overlapping rather than sharply defined. Our supervised learning models performed better, and XGBoost achieved the strongest results at 59% accuracy and relatively balanced precision across regions. Though these results are far from perfect, they demonstrate that having a small set of lexical features can still reveal broad regional patterns. Our SHAP analysis also helped clarify which specific linguistic variables contributed most to these predictions.

These findings are important because they confirm that American dialects are both real and measurable, yet far more blended and murky than traditional maps often suggest. Our project also provides a practical and accessible way for researchers, educators, and the general public to engage with linguistic diversity. The interactive dashboard allows users to explore regional language trends and see where their own speech comes from, encouraging reflection on language as a marker of culture, migration, and identity.

The project also came with various challenges. Our initial dataset was extremely large and required extensive preprocessing to address duplicated questions, inconsistent spellings, missing responses, and multiple versions of the survey. It was also difficult to manage more than eleven million raw responses across four interconnected CSV files and often called for many iterations. We also ran into some difficulty designing machine-learning features for a sparse, high-dimensional one-hot encoded dataset. We addressed these issues by standardizing response formats, keeping only consistently answered questions, and retaining users who were within the continental United States.

If we were to continue this project, there are several adjustments that could be made. First, we would aim to predict location at a finer geographic accuracy, such as the state or city level. While this would require additional features or more complex models, it would produce more personalized and intuitive results. Second, instead of presenting users with regional probability percentages alone, we could choose the top few cities or states that most closely match their responses, allowing users to reflect on whether these results align with their background or influences. Finally, we would redesign the quiz to move beyond a fixed set of questions. A larger pool of items could allow the quiz to dynamically select questions based on previous responses or user characteristics, making the experience more flexible. Additionally, allowing users to provide their actual home region could help the system collect new labeled data and improve its predictions over time.

Overall, this project highlights both the promise and the complexity of modeling American English dialects. By combining statistical analysis, machine learning, and a publicly accessible interactive platform, we offer insight into how Americans speak and how those linguistic patterns reflect broader cultural and geographic influences.

8 Appendix

Appendix A: Survey Questions

Number of questions: 108

1. Question 242: *been*
   (a) as in "sit"
   (b) i as in "see"
   (c) as in "set"
   (d) other

2. Question 243: *the first vowel in "Bowie knife"*
   - (a) o as in "Bo"
   - (b) u as in "boo"
   - (c) I have seen this word in print, but have no idea how to pronounce it
   - (d) I have never seen or heard this word
   - (e) other
3. Question 244: *caramel*
   - (a) with 2 syllables ("car-ml")
   - (b) with 3 syllables ("carra-mel")
   - (c) I use both interchangeably
   - (d) I have both forms, but the two have different meanings (please state how in the comments box)
   - (e) other
4. Question 245: *the vowel in the second syllable of "cauliflower"*
   - (a) i as in "see"
   - (b) as in "sit"
   - (c) other
5. Question 246: *the last vowel in "centaur"*
   - (a) as in "car" ("sen-tar")
   - (b) as in "caught"
   - (c) I use the same vowel in "car", "caught", and "centaur"
   - (d) rhymes with "sore" and "more" ("sen-tore")
   - (e) other
6. Question 247: *coupon*
   - (a) with u as in "coop" ("coopon")
   - (b) with ju as in "cute" ("cyoopon")
   - (c) other
7. Question 248: *Craig (the name)*
   - (a) as in "set"
   - (b) e as in "say"
   - (c) I say something in between the vowels in "set" and "say", but closer to the one in "say"
   - (d) I say something in between the vowels in "set" and "say", but closer to the one in "set"
   - (e) other
8. Question 249: *crayon*
   - (a) as in "man" (1 syllable, "cran")
   - (b) ej (2 syllables, "cray-ahn")
   - (c) ej (2 syllables, "cray-awn", where the second syllable rhymes with "dawn")
   - (d) aw (I pronounce this the same as "crown")
   - (e) other
9. Question 250: *creek (a small body of running water)*
   - (a) i as in "see"
   - (b) as in "sit"
   - (c) I use both interchangeably
   - (d) I don't know how to pronounce this word
   - (e) I use both, but they mean two different things (please state how they differ in the comments box)
   - (f) other
10. Question 251: *the first vowel in "Florida"*

(a) o as in "flow" ("flow-ri-da")

(b) as in "ah" ("flah-ri-da")

(c) as in "saw" ("flaw-ri-da")

(d) as in "sore" ("flore-i-da")

(e) other

11. Question 252: *flourish*

(a) as in "bird" ("flurr-ish")

(b) as in "sore" ("flore-ish")

(c) as in "sun" ("fluh-rish")

(d) other (including if you use one pronunciation for the verb and a different pronunciation for the noun)

12. Question 253: *the last vowel in "handkerchief"*

(a) i as in "see"

(b) as in "sit"

(c) other

13. Question 254: *lawyer*

(a) with j as in "boy" ("loyer")

(b) with as in "saw" ("law-yer")

(c) I use both interchangeably

(d) other

14. Question 255: *How do you pronounce Mary/merry/marry?*

(a) all 3 are the same

(b) all 3 are different

(c) Mary and merry are the same marry is different

(d) merry and marry are the same Mary is different

(e) Mary and marry are the same merry is different

15. Question 256: *mayonnaise*

(a) with as in "man" (2 syllables–"man-aze")

(b) with ej (3 syllables–"may-uh-naze")

(c) I use both interchangeably

(d) other

16. Question 257: *the first vowel in "miracle"*

(a) i as in "near"

(b) as in "knit"

(c) as in "net"

(d) I say something in between and

(e) other

17. Question 258: *mischievous vs. mischievious*

(a) mischievous (3 syllables)

(b) mischievious (4 syllables)

(c) I write "mischievous" but say "mischievious"

(d) I use both

(e) other

18. Question 259: *the final vowel in "Monday," "Friday," etc.*

(a) e as in "say"

(b) i as in "see"

(c) I use e with the words in isolation, but i in compounds (such as "Sunday school")

(d) other (e.g. do you use one vowel in some day names, and another in the other names?)

19. Question 260: *the second vowel in "pajamas"*
    - (a) as in "jam"
    - (b) as in "father"
    - (c) other
20. Question 261: *pecan*
    - (a) pikn with stress on the first syllable ("PEE-can")
    - (b) pikn with stress on the second syllable ("pee-CAN")
    - (c) pikn with stress on the first syllable ("PEE-Kahn")
    - (d) pikn with stress on the second syllable ("pee-KAHN")
    - (e) pkn ("pick Ann")
    - (f) pkn ("pick Ahn")
    - (g) I pronounce it differently when it's alone than when it's in a compound like "pecan pie" (please state how you pronounce the two variants in the comments box)
    - (h) other
21. Question 262: *poem*
    - (a) one syllable
    - (b) two syllables
22. Question 263: *really*
    - (a) i as in "see" ("reely")
    - (b) as in "sit" ("rilly")
    - (c) i ("ree-l-y")
    - (d) other (including if you use two or more of these interchangeably)
23. Question 264: *realtor (a real estate agent)*
    - (a) syllables ("reel-ter")
    - (b) syllables (realtor, in other words "reel-uh-ter")
    - (c) syllables (ree-l-ter)
    - (d) I don't use this word I use "estate agent"
    - (e) other
24. Question 265: *roof, room, broom, root*
    - (a) u as in "food"
    - (b) as in "foot"
    - (c) these four words do not all have the same vowel (please use the comments box to let us know which is which)
25. Question 266: *route (as in, "the route from one place to another")*
    - (a) rhymes with "hoot"
    - (b) rhymes with "out"
    - (c) I can pronounce it either way interchangeably
    - (d) I say it like "hoot" for the noun and like "out" for the verb.
    - (e) I say it like "out" for the noun and like "hoot" for the verb.
    - (f) other
26. Question 267: *the first vowel in "syrup"*
    - (a) i "sear-up"
    - (b) "sih-rup"
    - (c) as in "sir"
    - (d) other
27. Question 268: *Do you pronounce "cot" and "caught" the same?*
    - (a) different
    - (b) same

28. Question 269: *almond*
    (a) all-mond (first syllable sounds like "all")
    (b) ah-mond (no l)
    (c) aw-mond (if different from "ah-mond")
    (d) I say something in between l and nothing
    (e) other

29. Question 270: *the "s" in "anniversary"*
    (a) s as in "sock"
    (b) as in "shock"

30. Question 271: *asterisk*
    (a) asteriks
    (b) asterisk
    (c) asterik (with no s in the final cluster)
    (d) other

31. Question 272: *candidate*
    (a) I pronounce the first d
    (b) I don't pronounce the first d
    (c) I vary freely between pronouncing the first d and not doing so
    (d) I only pronounce the first d when I'm speaking slowly/carefully
    (e) Depends whether it refers to a political or generic candidate, as in "that assignment looks like a good candidate for elimination" (please state how the two pronunciations differ)
    (f) other

32. Question 273: *the "s" in "chromosome"*
    (a) s
    (b) z
    (c) both are acceptable to me
    (d) other

33. Question 274: *et cetera*
    (a) pronounced etsetera (4 syllables)
    (b) pronounced etsetra (3 syllables)
    (c) pronounced eksetera (4 syllables)
    (d) pronounced eksetra (3 syllables)
    (e) other

34. Question 275: *the final consonant in "garage"*
    (a) as in the middle consonant of "measure"
    (b) as in "edge"
    (c) I use both interchangeably
    (d) other

35. Question 276: *the "c" in "grocery"*
    (a) s as in "sock"
    (b) as in "shock"
    (c) other

36. Question 277: *huge, humor, humongous, human...*
    (a) I pronounce the h
    (b) I don't pronounce the h
    (c) I can pronounce the h or not
    (d) other

37. Question 278: *the "s" in "nursery"*

    (a) s as in "sock"

    (b) as in "shock"

    (c) other

38. Question 279: *the "s" in the last name of Elvis Presley*

    (a) s

    (b) z

39. Question 280: *quarter*

    (a) with kw

    (b) with k ("cor-ter")

    (c) I use both interchangeably

    (d) other

40. Question 281: *Do you use "spigot" or "spicket" to refer to a faucet or tap that water comes out of?*

    (a) spicket

    (b) spigot

    (c) I use both interchangeably

    (d) I say "spicket" but spell it "spigot"

    (e) I use both with different meanings (please explain how in the comments box)

    (f) I don't use either version of this word

    (g) other

41. Question 282: *strength*

    (a) the "g" is pronounced as

    (b) the "g" is pronounced as k

    (c) the "g" is silent

42. Question 283: *the final consonant in "Texas"*

    (a) s

    (b) z

    (c) either one

    (d) other

43. Question 284: *cream cheese*

    (a) CREAM cheese (stress on the first syllable)

    (b) cream CHEESE (stress on the second syllable)

    (c) it sounds right either way

    (d) other

44. Question 285: *insurance*

    (a) INsurance (stress on the first syllable)

    (b) inSURance (stress on the second syllable)

    (c) I can stress either the first or the second syllable

    (d) other

45. Question 286: *New Haven (the city in Connecticut where Yale University is located)*

    (a) NEW Haven

    (b) New HAVEN

    (c) I use both interchangeably

    (d) other

46. Question 287: *Thanksgiving*

    (a) THANKSgiving

    (b) ThanksGIVing

   (c) I use both interchangeably
   (d) other

47. Question 288: *umbrella*
    (a) UMbrella
    (b) umBRELLa

48. Question 289: *I her lifeless body from the pool*
    (a) dragged
    (b) drug
    (c) I use both interchangeably
    (d) other

49. Question 291: *Would you say "Are you coming with?" as a full sentence, to mean "Are you coming with us?"*
    (a) yes
    (b) no
    (c) other

50. Question 292: *Would you say "where are you at?" to mean "where are you?"*
    (a) yes
    (b) no
    (c) I can use "where are you at" in contexts such as asking someone how s/he is coming along on a project, but not in the general sense of "where are you physically located in the world at this moment".

51. Question 293: *Modals are words like "can," "could," "might," "ought to," and so on. Can you use more than one modal at a time? (e.g., "I might could do that" to mean "I might be able to do that" or "I used to could do that" to mean "I used to be able to do that")*
    (a) yes (please consider adding which combinations of modals you use in the comments box)
    (b) no
    (c) other

52. Question 294: *He used to nap on the couch, but he sprawls out in that new lounge chair anymore*
    (a) this use of "anymore" is acceptable
    (b) this use of "anymore" is unacceptable
    (c) not sure

53. Question 295: *I do exclusively figurative paintings anymore*
    (a) acceptable
    (b) unacceptable
    (c) not sure

54. Question 296: *Pantyhose are so expensive anymore that I just try to get a good suntan and forget about it.*
    (a) acceptable
    (b) unacceptable
    (c) not sure

55. Question 297: *Forget the nice clothes anymore (referring to babies eating messily after a certain age)*
    (a) acceptable
    (b) unacceptable
    (c) not sure

56. Question 299: *What do you call the game wherein the participants see who can throw a knife closest to the other person (or alternately, get a jackknife to stick into the ground or a piece of wood)?*

(a) mumblety-peg
(b) mumbledy-peg
(c) mumbly peg
(d) mumbly pegs
(e) mumblely peg (with 2 l's)
(f) mumble peg
(g) mummety-peg
(h) mumble-the-peg
(i) fumbledy peg
(j) numblety peg
(k) peggy
(l) baseball jackknife
(m) stick-knife
(n) stick-frog
(o) stretch
(p) chicken
(q) knifey
(r) splits
(s) Russian roulette
(t) I have never heard of this "game" and have no idea what it's called
(u) other (state here if you have heard one or more of these terms but never knew what they meant)

57. Question 300: *What do you call the area of grass between the sidewalk and the road?*

(a) berm
(b) parking
(c) tree lawn
(d) terrace
(e) curb strip
(f) beltway
(g) verge
(h) I have no word for this
(i) other

58. Question 301: *What do you call the area of grass that occurs in the middle of some streets?*

(a) boulevard
(b) midway
(c) traffic island
(d) island
(e) neutral ground
(f) I have no word for this
(g) other

59. Question 302: *What do you call the long narrow place in the middle of a divided highway?*

(a) median strip
(b) median
(c) boulevard
(d) mall
(e) traffic island
(f) neutral ground
(g) island
(h) park strip
(i) I have no word for this

(j) other

60. Question 303: *What do you call the drink made with milk and ice cream?*

    (a) milkshake/shake
    (b) frappe
    (c) cabinet
    (d) velvet
    (e) thick shake
    (f) other

61. Question 305: *What do you call the insect that flies around in the summer and has a rear section that glows in the dark?*

    (a) lightning bug
    (b) firefly
    (c) I use lightning bug and firefly interchangeably
    (d) peenie wallie
    (e) I have no word for this
    (f) other

62. Question 307: *What do you call the kind of spider (or spider-like creature) that has an oval-shaped body and extremely long legs?*

    (a) daddy long leg(s)
    (b) daddy big legs
    (c) daddy (bug)
    (d) father longlegs
    (e) granddaddy
    (f) daddy graybeard
    (g) daddy spider
    (h) harvestman
    (i) moskeet spider
    (j) pointer
    (k) shepherd spider
    (l) other

63. Question 308: *What nicknames do/did you use for your maternal grandmother?*

    (a) grandmother
    (b) granny
    (c) grandma
    (d) nana
    (e) mimi
    (f) grammy/grammie/grammi
    (g) other

64. Question 309: *What about your paternal grandmother (is there a distinction?)*

    (a) grandmother
    (b) granny
    (c) grandma
    (d) gramma
    (e) nana
    (f) other

65. Question 310: *What do/did you call your maternal grandfather?*

    (a) gramps
    (b) grandpa
    (c) grampa

(d) grandad, granddad
(e) pap
(f) I spell it "grandpa" but pronounce it as "grampa"
(g) other (including if you use a different term to address him directly than you do when speaking about him to a third party)

66. Question 311: *paternal grandfather?*

    (a) gramps
    (b) grandpa
    (c) grampa
    (d) pap
    (e) other

67. Question 312: *What do you call the big clumps of dust that gather under furniture and in corners?*

    (a) dust bunnies
    (b) dust kittens
    (c) dust mice
    (d) kitties
    (e) dust balls
    (f) other

68. Question 316: *What term do you use to refer to something that is across both streets from you at an intersection (or diagonally across from you in general)?*

    (a) kitty-corner
    (b) kitacorner
    (c) catercorner
    (d) catty-corner
    (e) kitty cross
    (f) kitty wampus
    (g) I can only use "diagonal" for this
    (h) I have no term for this
    (i) other

69. Question 317: *What do you call the activity of driving around in circles in a car?*

    (a) doing donuts
    (b) doing cookies
    (c) whipping shitties
    (d) other

70. Question 318: *What do you call paper that has already been used for something or is otherwise imperfect?*

    (a) scratch paper
    (b) scrap paper
    (c) scratch paper is still usable (for example, the paper you bring to do extra work on a test) scrap paper is paper that isn't needed anymore and can be thrown away.
    (d) other

71. Question 319: *What is your general term for a big road that you drive relatively fast on?*

    (a) highway
    (b) freeway
    (c) parkway
    (d) turnpike
    (e) expressway
    (f) throughway/thru-way

(g) a freeway is bigger than a highway
(h) a freeway is free (i.e., doesn't charge tolls) a highway isn't
(i) a freeway has limited access (no stop lights, no intersections), whereas a highway can have stop lights and intersections
(j) other

72. Question 321: *When you are cold, and little points of skin begin to come on your arms and legs, you have-*

    (a) goose bumps
    (b) goose flesh
    (c) goose pimples
    (d) chill bumps
    (e) chill bugs
    (f) chilly bumps
    (g) cold-chill bumps
    (h) other

73. Question 323: *What do you call an easy course?*

    (a) gut
    (b) crypt course
    (c) crip course
    (d) bird
    (e) blow-off
    (f) meat
    (g) other

74. Question 325: *What is the thing that women use to tie their hair?*

    (a) (hair) elastic
    (b) rubber band
    (c) horsetail
    (d) hair thing
    (e) hair tie
    (f) other

75. Question 326: *Do you use the word cruller?*

    (a) yes
    (b) no, but I know what it means
    (c) I have no idea what this means

76. Question 327: *Do you use the term "bear claw" for a kind of pastry?*

    (a) yes
    (b) no, but I know what it means
    (c) I have no idea what this means

77. Question 328: *What do you call someone who is the opposite of pigeon-toed (i.e. when they walk their feet point outwards)?*

    (a) duck-footed
    (b) slue-footed
    (c) splay-footed
    (d) bow-legged
    (e) toed out
    (f) other
    (g) I have no word for this

78. Question 329: *Can you call coleslaw "slaw"?*

    (a) yes

(b) yes, but I can also use it in other forms such as apple slaw or broccoli slaw
(c) no
(d) I have never heard that usage before
(e) other

79. Question 330: *What do you call the box you bury a dead person in?*
    (a) coffin
    (b) casket
    (c) a coffin and a casket are not the same, and I know the difference
    (d) other

80. Question 331: *Do you say "vinegar and oil" or "oil and vinegar" for the type of salad dressing?*
    (a) vinegar and oil
    (b) oil and vinegar
    (c) both sound equally good to me
    (d) neither
    (e) other

81. Question 332: *What do you call it when a driver changes over one or more lanes way too quickly?*
    (a) Chinese lane change
    (b) Chinese fire drill
    (c) other

82. Question 333: *When you stand outside with a long line of people waiting to get in somewhere, are you standing "in line" or "on line" (as in, "I stood in the cold for two hours before they opened the doors")?*
    (a) on line
    (b) in line
    (c) both sound equally good
    (d) neither
    (e) other

83. Question 334: *Do you say "frosting" or "icing" for the sweet spread one puts on a cake?*
    (a) frosting
    (b) icing
    (c) icing is thinner than frosting, white, and/or made of powdered sugar and milk or lemon juice
    (d) both
    (e) neither
    (f) other

84. Question 335: *What is "the City"?*
    (a) New York City
    (b) Boston
    (c) DC
    (d) LA
    (e) Chicago
    (f) other

85. Question 336: *What is the distinction between dinner and supper?*
    (a) supper is an evening meal while dinner is eaten earlier (lunch, for example)
    (b) supper is an evening meal, dinner is the main meal
    (c) dinner takes place in a more formal setting than supper
    (d) there is no distinction they both have the same meaning

(e) I do not use the term supper
(f) I don't use the term dinner
(g) other

86. Question 337: *Which of these terms do you prefer?*
    (a) trash can
    (b) garbage can
    (c) rubbish bin
    (d) waste(paper) basket
    (e) These words refer to different things
    (f) other
87. Question 338: *Which of these terms do you prefer?*
    (a) By accident
    (b) On accident
    (c) both
    (d) neither
    (e) other
88. Question 339: *Which of these terms do you prefer for the small road parallel to the highway?*
    (a) frontage road
    (b) service road
    (c) access road
    (d) feeder road
    (e) gateway
    (f) we have them but I have no word for them
    (g) I've never heard of this concept
    (h) other
89. Question 340: *Do you cut or mow the lawn or grass?*
    (a) cut the grass
    (b) cut the lawn
    (c) mow the grass
    (d) mow the lawn
    (e) other
90. Question 341: *Do you pass in homework or hand in homework?*
    (a) pass in
    (b) hand in
    (c) both
    (d) neither
    (e) other
91. Question 343: *What do you call the thing from which you might drink water in a school?*
    (a) bubbler
    (b) water bubbler
    (c) drinking fountain
    (d) water fountain
    (e) other
92. Question 344: *What do you call a public railway system (normally underground)?*
    (a) the subway
    (b) the L, or the El
    (c) the T
    (d) the metro

    (e) BART
    (f) other

93. Question 346: *What do you call the act of covering a house or area in front of a house with toilet paper?*

    (a) tp'ing
    (b) rolling
    (c) toilet papering
    (d) wrapping
    (e) papering
    (f) bog rolling
    (g) I have no word for this
    (h) other

94. Question 347: *What do you call a traffic jam caused by drivers slowing down to look at an accident or other diversion on the side of the road?*

    (a) rubberneck
    (b) rubbernecking
    (c) rubbernecking is the activity (slowing down and gawking) that causes the traffic jam, but I have no word for the traffic jam itself
    (d) gapers' block
    (e) gapers' delay
    (f) Lookie Lou
    (g) curiosity delay
    (h) gawk block
    (i) I have no word for this
    (j) other

95. Question 348: *What vowel do you use in bag?*

    (a) as in "sat"
    (b) as in "set"
    (c) e as in "say"
    (d) other

96. Question 349: *What do you call the paper container in which you might bring home items you bought at the store?*

    (a) bag
    (b) sack
    (c) poke
    (d) other

97. Question 350: *What do you call the night before Halloween?*

    (a) gate night
    (b) trick night
    (c) mischief night
    (d) cabbage night
    (e) goosy night
    (f) devil's night
    (g) devil's eve
    (h) I have no word for this
    (i) other

98. Question 351: *What do you call the end of a loaf of bread?*

    (a) end
    (b) heel

(c) crust
(d) nose
(e) butt
(f) shpitzel
(g) I have no word for this
(h) other

99. Question 352: *How do you pronounce the word for the type of drug that acts as central nervous system depressant and is used as a sedative or hypnotic? (Please do not look up the word in a dictionary before answering this question.)*

    (a) barbituate
    (b) barbiturate
    (c) I don't use either of these
    (d) other

100. Question 353: *amphitheater*

    (a) f
    (b) p
    (c) other

101. Question 354: *citizen*

    (a) s
    (b) z
    (c) other

102. Question 355: *What do you call a point that is purely academic, or that cannot be settled and isn't worth discussing further?*

    (a) a moot point
    (b) a mute point
    (c) either one of the above
    (d) I have no idea
    (e) other

103. Question 356: *How do you pronounce the -sp- sequence in "thespian" (the word meaning "actor")?*

    (a) sp (as in "desperate")
    (b) zb (rhymes with "lesbian")
    (c) other

104. Question 357: *What do you call the level of a building that is partly or entirely underground?*

    (a) basement
    (b) cellar
    (c) I use both, and they mean the same thing
    (d) A basement is finished (for example with plastered or painted walls, carpets, etc.), whereas a cellar is unfinished (made up of bare stone or cement, used only for storage).
    (e) A cellar has an outside entrance (some call this a "bulkhead"), whereas a basement does not
    (f) other

105. Question 358: *What do you call a drive-through liquor store?*

    (a) brew thru
    (b) party barn
    (c) bootlegger
    (d) beer barn
    (e) beverage barn
    (f) we have these in my area, but we have no special term for them

   (g) I have never heard of such a thing
   (h) other

106. Question 360: *What do you say when you want to lay claim to the front seat of a car?*
   (a) dibs
   (b) shotgun
   (c) hosey
   (d) high hosey
   (e) I have no term for this
   (f) other

107. Question 361: *What word do you use for gawking at someone in a lustful way?*
   (a) ogle
   (b) oogle
   (c) oggle (pronounced to rhyme with "boggle", but may still be spelled "ogle")
   (d) I use both oogle and ogle interchangeably
   (e) I use both ogle and "oggle"
   (f) I have no word for this activity
   (g) other

108. Question 362: *Do you say "expecially", or "especially"?*
   (a) expecially (or "ecspecially" or "ekspecially")
   (b) especially
   (c) I use both interchangeably
   (d) neither
   (e) other

Appendix B: Quiz Questions

1. Question 303: *What do you call the drink made with milk + ice cream?*
   (a) milkshake/shake
   (b) frappe
   (c) cabinet
   (d) velvet
   (e) thick shake
   (f) other

2. Question 300: *Grass between sidewalk + road?*
   (a) berm
   (b) parking
   (c) tree lawn
   (d) terrace
   (e) curb strip
   (f) beltway
   (g) verge
   (h) other

3. Question 335: *What is “the City”?*
   (a) New York City
   (b) Boston
   (c) DC
   (d) LA
   (e) Chicago
   (f) other

4. Question 358: *Drive-through liquor store?*
    (a) party barn
    (b) brew thru
    (c) bootlegger
    (d) beer barn
    (e) beverage barn
    (f) no special term
    (g) never heard
    (h) other
5. Question 316: *Diagonal across the street?*
    (a) kitty-corner
    (b) kitacorner
    (c) catercorner
    (d) catty-corner
    (e) kitty cross
    (f) kitty wampus
    (g) diagonal
    (h) other
6. Question 350: *Night before Halloween?*
    (a) mischief night
    (b) devil's night
    (c) cabbage night
    (d) goosy night
    (e) gate night
    (f) trick night
    (g) I have no word
    (h) other
7. Question 343: *Thing you drink water from in school?*
    (a) bubbler
    (b) drinking fountain
    (c) water fountain
    (d) water bubbler
    (e) other
8. Question 319: *General term for a big road you drive fast on?*
    (a) highway
    (b) freeway
    (c) parkway
    (d) turnpike
    (e) expressway
    (f) throughway/thru-way
    (g) other
9. Question 302: *Median of a divided highway?*
    (a) median
    (b) median strip
    (c) neutral ground
    (d) mall
    (e) traffic island
    (f) island
    (g) park strip

    (h) other

10. Question 305: *Glow-in-the-dark bug?*
    (a) lightning bug
    (b) firefly
    (c) both
    (d) peenie wallie
    (e) I have no word
    (f) other

Appendix C: Top 50 Survey Questions Identified by Logistic Regression Feature Coefficients

1. Question 245: *the vowel in the second syllable of "cauliflower"*
    (a) i as in "see"
    (b) as in "sit"
    (c) other
2. Question 249: *crayon*
    (a) as in "man" (1 syllable, "cran")
    (b) ej (2 syllables, "cray-ahn")
    (c) ej (2 syllables, "cray-awn", where the second syllable rhymes with "dawn")
    (d) aw (I pronounce this the same as "crown")
    (e) other
3. Question 250: *creek (a small body of running water)*
    (a) i as in "see"
    (b) as in "sit"
    (c) I use both interchangeably
    (d) I don't know how to pronounce this word
    (e) I use both, but they mean two different things (please state how they differ in the comments box)
    (f) other
4. Question 253: *the last vowel in "handkerchief"*
    (a) i as in "see"
    (b) as in "sit"
    (c) other
5. Question 255: *How do you pronounce Mary/merry/marry?*
    (a) all 3 are the same
    (b) all 3 are different
    (c) Mary and merry are the same marry is different
    (d) merry and marry are the same Mary is different
    (e) Mary and marry are the same merry is different
6. Question 267: *the first vowel in "syrup"*
    (a) i "sear-up"
    (b) "sih-rup"
    (c) as in "sir"
    (d) other
7. Question 268: *Do you pronounce "cot" and "caught" the same?*
    (a) different
    (b) same

8. Question 272: *candidate*
    (a) I pronounce the first d
    (b) I don't pronounce the first d
    (c) I vary freely between pronouncing the first d and not doing so
    (d) I only pronounce the first d when I'm speaking slowly/carefully
    (e) Depends whether it refers to a political or generic candidate, as in "that assignment looks like a good candidate for elimination" (please state how the two pronunciations differ)
    (f) other
9. Question 277: *huge, humor, humongous, human...*
    (a) I pronounce the h
    (b) I don't pronounce the h
    (c) I can pronounce the h or not
    (d) other
10. Question 278: *the "s" in "nursery"*
    (a) s as in "sock"
    (b) as in "shock"
    (c) other
11. Question 281: *Do you use "spigot" or "spicket" to refer to a faucet or tap that water comes out of?*
    (a) spicket
    (b) spigot
    (c) I use both interchangeably
    (d) I say "spicket" but spell it "spigot"
    (e) I use both with different meanings (please explain how in the comments box)
    (f) I don't use either version of this word
    (g) other
12. Question 283: *the final consonant in "Texas"*
    (a) s
    (b) z
    (c) either one
    (d) other
13. Question 285: *insurance*
    (a) INsurance (stress on the first syllable)
    (b) inSURance (stress on the second syllable)
    (c) I can stress either the first or the second syllable
    (d) other
14. Question 286: *New Haven (the city in Connecticut where Yale University is located)*
    (a) NEW Haven
    (b) New HAVEN
    (c) I use both interchangeably
    (d) other
15. Question 299: *What do you call the game wherein the participants see who can throw a knife closest to the other person (or alternately, get a jackknife to stick into the ground or a piece of wood)?*
    (a) mumblety-peg
    (b) mumbledy-peg
    (c) mumbly peg
    (d) mumbly pegs

(e) mumblely peg (with 2 l's)
(f) mumble peg
(g) mummety-peg
(h) mumble-the-peg
(i) fumbledy peg
(j) numblety peg
(k) peggy
(l) baseball jackknife
(m) stick-knife
(n) stick-frog
(o) stretch
(p) chicken
(q) knifey
(r) splits
(s) Russian roulette
(t) I have never heard of this "game" and have no idea what it's called
(u) other (state here if you have heard one or more of these terms but never knew what they meant)

16. Question 300: *What do you call the area of grass between the sidewalk and the road?*

    (a) berm
    (b) parking
    (c) tree lawn
    (d) terrace
    (e) curb strip
    (f) beltway
    (g) verge
    (h) I have no word for this
    (i) other

17. Question 301: *What do you call the area of grass that occurs in the middle of some streets?*

    (a) boulevard
    (b) midway
    (c) traffic island
    (d) island
    (e) neutral ground
    (f) I have no word for this
    (g) other

18. Question 302: *What do you call the long narrow place in the middle of a divided highway?*

    (a) median strip
    (b) median
    (c) boulevard
    (d) mall
    (e) traffic island
    (f) neutral ground
    (g) island
    (h) park strip
    (i) I have no word for this
    (j) other

19. Question 303: *What do you call the drink made with milk and ice cream?*

    (a) milkshake/shake
    (b) frappe

(c) cabinet
(d) velvet
(e) thick shake
(f) other

20. Question 305: *What do you call the insect that flies around in the summer and has a rear section that glows in the dark?*

(a) lightning bug
(b) firefly
(c) I use lightning bug and firefly interchangeably
(d) peenie wallie
(e) I have no word for this
(f) other

21. Question 307: *What do you call the kind of spider (or spider-like creature) that has an oval-shaped body and extremely long legs?*

(a) daddy long leg(s)
(b) daddy big legs
(c) daddy (bug)
(d) father longlegs
(e) granddaddy
(f) daddy graybeard
(g) daddy spider
(h) harvestman
(i) moskeet spider
(j) pointer
(k) shepherd spider
(l) other

22. Question 310: *What do/did you call your maternal grandfather?*

(a) gramps
(b) grandpa
(c) grampa
(d) grandad, granddad
(e) pap
(f) I spell it "grandpa" but pronounce it as "grampa"
(g) other (including if you use a different term to address him directly than you do when speaking about him to a third party)

23. Question 312: *What do you call the big clumps of dust that gather under furniture and in corners?*

(a) dust bunnies
(b) dust kittens
(c) dust mice
(d) kitties
(e) dust balls
(f) other

24. Question 316: *What term do you use to refer to something that is across both streets from you at an intersection (or diagonally across from you in general)?*

(a) kitty-corner
(b) kitacorner
(c) catercorner
(d) catty-corner
(e) kitty cross

(f) kitty wampus
(g) I can only use "diagonal" for this
(h) I have no term for this
(i) other

25. Question 317: *What do you call the activity of driving around in circles in a car?*
    (a) doing donuts
    (b) doing cookies
    (c) whipping shitties
    (d) other

26. Question 319: *What is your general term for a big road that you drive relatively fast on?*
    (a) highway
    (b) freeway
    (c) parkway
    (d) turnpike
    (e) expressway
    (f) throughway/thru-way
    (g) a freeway is bigger than a highway
    (h) a freeway is free (i.e., doesn't charge tolls) a highway isn't
    (i) a freeway has limited access (no stop lights, no intersections), whereas a highway can have stop lights and intersections
    (j) other

27. Question 321: *When you are cold, and little points of skin begin to come on your arms and legs, you have-*
    (a) goose bumps
    (b) goose flesh
    (c) goose pimples
    (d) chill bumps
    (e) chill bugs
    (f) chilly bumps
    (g) cold-chill bumps
    (h) other

28. Question 323: *What do you call an easy course?*
    (a) gut
    (b) crypt course
    (c) crip course
    (d) bird
    (e) blow-off
    (f) meat
    (g) other

29. Question 325: *What is the thing that women use to tie their hair?*
    (a) (hair) elastic
    (b) rubber band
    (c) horsetail
    (d) hair thing
    (e) hair tie
    (f) other

30. Question 333: *When you stand outside with a long line of people waiting to get in somewhere, are you standing "in line" or "on line" (as in, "I stood in the cold for two hours before they opened the doors")?*

    (a) on line
    (b) in line
    (c) both sound equally good
    (d) neither
    (e) other

31. Question 334: *Do you say "frosting" or "icing" for the sweet spread one puts on a cake?*
    (a) frosting
    (b) icing
    (c) icing is thinner than frosting, white, and/or made of powdered sugar and milk or lemon juice
    (d) both
    (e) neither
    (f) other

32. Question 335: *What is "the City"?*
    (a) New York City
    (b) Boston
    (c) DC
    (d) LA
    (e) Chicago
    (f) other

33. Question 336: *What is the distinction between dinner and supper?*
    (a) supper is an evening meal while dinner is eaten earlier (lunch, for example)
    (b) supper is an evening meal, dinner is the main meal
    (c) dinner takes place in a more formal setting than supper
    (d) there is no distinction they both have the same meaning
    (e) I do not use the term supper
    (f) I don't use the term dinner
    (g) other

34. Question 337: *Which of these terms do you prefer?*
    (a) trash can
    (b) garbage can
    (c) rubbish bin
    (d) waste(paper) basket
    (e) These words refer to different things
    (f) other

35. Question 338: *Which of these terms do you prefer?*
    (a) By accident
    (b) On accident
    (c) both
    (d) neither
    (e) other

36. Question 339: *Which of these terms do you prefer for the small road parallel to the highway?*
    (a) frontage road
    (b) service road
    (c) access road
    (d) feeder road
    (e) gateway
    (f) we have them but I have no word for them
    (g) I've never heard of this concept

(h) other

37. Question 341: *Do you pass in homework or hand in homework?*
    (a) pass in
    (b) hand in
    (c) both
    (d) neither
    (e) other

38. Question 343: *What do you call the thing from which you might drink water in a school?*
    (a) bubbler
    (b) water bubbler
    (c) drinking fountain
    (d) water fountain
    (e) other

39. Question 344: *What do you call a public railway system (normally underground)?*
    (a) the subway
    (b) the L, or the El
    (c) the T
    (d) the metro
    (e) BART
    (f) other

40. Question 346: *What do you call the act of covering a house or area in front of a house with toilet paper?*
    (a) tp'ing
    (b) rolling
    (c) toilet papering
    (d) wrapping
    (e) papering
    (f) bog rolling
    (g) I have no word for this
    (h) other

41. Question 347: *What do you call a traffic jam caused by drivers slowing down to look at an accident or other diversion on the side of the road?*
    (a) rubberneck
    (b) rubbernecking
    (c) rubbernecking is the activity (slowing down and gawking) that causes the traffic jam, but I have no word for the traffic jam itself
    (d) gapers' block
    (e) gapers' delay
    (f) Lookie Lou
    (g) curiosity delay
    (h) gawk block
    (i) I have no word for this
    (j) other

42. Question 349: *What do you call the paper container in which you might bring home items you bought at the store?*
    (a) bag
    (b) sack
    (c) poke
    (d) other

43. Question 350: *What do you call the night before Halloween?*
    (a) gate night
    (b) trick night
    (c) mischief night
    (d) cabbage night
    (e) goosy night
    (f) devil's night
    (g) devil's eve
    (h) I have no word for this
    (i) other
44. Question 351: *What do you call the end of a loaf of bread?*
    (a) end
    (b) heel
    (c) crust
    (d) nose
    (e) butt
    (f) shpitzel
    (g) I have no word for this
    (h) other
45. Question 352: *How do you pronounce the word for the type of drug that acts as central nervous system depressant and is used as a sedative or hypnotic? (Please do not look up the word in a dictionary before answering this question.)*
    (a) barbituate
    (b) barbiturate
    (c) I don't use either of these
    (d) other
46. Question 355: *What do you call a point that is purely academic, or that cannot be settled and isn't worth discussing further?*
    (a) a moot point
    (b) a mute point
    (c) either one of the above
    (d) I have no idea
    (e) other
47. Question 356: *How do you pronounce the -sp- sequence in "thespian" (the word meaning "actor")?*
    (a) sp (as in "desperate")
    (b) zb (rhymes with "lesbian")
    (c) other
48. Question 357: *What do you call the level of a building that is partly or entirely underground?*
    (a) basement
    (b) cellar
    (c) I use both, and they mean the same thing
    (d) A basement is finished (for example with plastered or painted walls, carpets, etc.), whereas a cellar is unfinished (made up of bare stone or cement, used only for storage).
    (e) A cellar has an outside entrance (some call this a "bulkhead"), whereas a basement does not
    (f) other
49. Question 358: *What do you call a drive-through liquor store?*
    (a) brew thru

    (b) party barn
    (c) bootlegger
    (d) beer barn
    (e) beverage barn
    (f) we have these in my area, but we have no special term for them
    (g) I have never heard of such a thing
    (h) other

50. Question 360: *What do you say when you want to lay claim to the front seat of a car?*
    (a) dibs
    (b) shotgun
    (c) hosey
    (d) high hosey
    (e) I have no term for this
    (f) other

Acknowledgments

We thank our mentor, Dr. Emily Prud'hommeaux, whose guidance was essential to shaping this project. We are also grateful to Dr. Bert Vaux for providing access to his high-quality dataset.

References

Rick Aschmann. North american english dialects, 2014. URL `https://aschmann.net/AmEng/`. Accessed: 2025-010-31.

Melissa Axelrod and Joanne Scheibman. Contemporary english in the usa. *English Faculty Publications*, (31), 2013. URL `https://digitalcommons.odu.edu/english_fac_pubs/31`.

Charles Boberg. The north american regional vocabulary survey: New variables and methods in the study of north american english. *American Speech*, 80(1):22–60, 2005.

Cynthia G. Clopper, Susannah V. Levi, and David B. Pisoni. Perceptual similarity of regional dialects of american english. *The Journal of the Acoustical Society of America*, 119(1):566–574, 2006. doi: 10.1121/1.2141171.

Keelan Evanini. *The Role of Phonological Variation in Predicting Linguistic Variables*. PhD thesis, University of Pennsylvania, 2010.

Jack Grieve, Dirk Speelman, and Dirk Geeraerts. A statistical methodology for the identification and characterization of regional dialect variation. *Dialectologia*, Special Issue II:1–28, 2011.

Stephanie Nicole Hedges. Dialect regions in north america identified from a survey of lexical and phonological variables. Master's thesis, Brigham Young University, 2017.

Josh Katz and Wilson Andrews. The U.S. dialect quiz: How y'all, youse and you guys talk. *The New York Times*, April 11 2024. URL `https://www.nytimes.com/interactive/2014/upshot/dialect-quiz-map.html`.

William Labov, Sharon Ash, and Charles Boberg. *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Mouton de Gruyter, Berlin, 2006.

Richard R. Lee. Dialect perception: A critical review and re-evaluation. *Quarterly Journal of Speech*, 57(4):410–417, 1971. doi: 10.1080/00335637109383086.

Raffaella Zanuttini, Laurence Horn, Emily M. Bender, Bill Haddican, Rebecca Nye, Jenny Liu, and Jim Wood. The yale grammatical diversity project: Morphosyntactic variation in north american english. *Language and Linguistics Compass*, 12(3), 2018.