A Sociolinguistic Analysis of Lexical Variation on Reddit

Harrison Finkelstein-Hynes (260969949)

Simon Risman (260973208)

Harrison Finkelstein-Hynes
Simon Risman

Background

The internet is one of the primary places we can find natural speech that does not require any elicitation. *Reddit* is an online forum with a seemingly limitless subdivision of communities catering to geographic regions, hobbies, educational institutions, and more. These community divisions can provide us with valuable insights into how language varies across them and how language spreads between them. The multitude of divisions also creates an opportunity to observe and test the homophily hypothesis from a linguistic perspective. We will test our prediction that the more similar the lexical choices within a Reddit community are, the more similar their user base; it follows that the lexical choices can be used to draw insights about the demographic of the subreddit.

The homophily hypothesis has two important components. First is the similarity attraction hypothesis (Byrne, 1971) claims that individuals who share similar traits are more likely to interact with one another. Second, the theory of self-categorization (Turner, 1987) proposes that an individual A's perception of their similarity to an individual B stems from their perceived proximity across a series of categories, including race, gender, age, level of education, and ethnic background, among others.

Reddit is an interesting way to study homophily, because users' identities are kept anonymous. A Reddit user's identity is only available if they explicitly choose to share it, whether that be in a post or their bio. Users are thus generally unaware of the explicit identity of others on the site, instead of finding groups based on identity, users differentiate themselves by topic of discussion in sub-communities called subreddits. Subreddits communities often center around a topic (r/nfl, r/nba), but sometimes that topic says much more about ideology or identity, than it does about interest (r/conservative, r/lgbt). Subreddits communities can then be seen as a collection of posts by people who at the very least have the same interest in the subreddits given topic

Methodology

Instead of using the Reddit API to gather most of our data, due to computational and time constrictions, we chose to use the Reddit (by subreddit) corpus compiled using the Convokit python package at the Stanford University Social Media. This corpus includes the complete record of 100 subreddits from 2005 to 2018. It provides a wealth of metadata on the speaker, conversation, utterance, and corpus levels.

Using data from (Kumar et al., 2018) it is possible to systematically compute the overlap of users between subreddits. This similarity score will be how we measure the similarity of the categories, in this case, the subreddits, that people have sorted themselves into.

In addition, the website https://subredditstats.com/ takes an input subreddit and returns a list of similarity scores that are computed based on how likely a user that posted in the input subreddit is to post in the other subreddits. It is clarified by the website as "a score of 2 means that users of [subreddit] are twice as likely to post and comment on that subreddit. A score of 1 means that users of [subreddit] are no more likely to frequent that subreddit than the average Reddit user. A score of 0 means that users of [subreddit] never post/comment on that subreddit." It uses the same general principle as the academic paper, just with an older data set. We found this tool adequate given our computational bottlenecks. .

The independent variable assessed in this investigation was only the subreddit we chose to analyze, as it is the primary attribute we are using to divide Reddit users.

The dependent variables we will observe are as follows: Incidence of Contractions, markers of regional dialects of North American English, British and American spellings, and Acronyms in each selected subreddit. The incidence of contractions is measured by the ratio of the selected contraction to its expanded form, for example, 'gonna' to 'going to.' We assigned each variable a predicted group or trait it is associated with.

Harrison Finkelstein-Hynes
Simon Risman

Contractions are likely to be a marker of formality, with a higher incidence of contractions in subreddits with less formal subject matter. Subreddits related to hobbies, like anime and hiking, are likely to include a less formal style of language than, for example, debate or philosophy.

Markers of regional variants of English will be reflective of the geographic background of a subreddit's users. This effect will likely be more pronounced in subreddits explicitly associated with a region, like a city or a sports team. Also, regional markers can help us identify if a subreddit values covert prestige. As found by (Trudgill, 1978), regional linguistic markers are tied to an increase in covert prestige, allowing us to comment on how a subreddit views the balance of overt and covert prestige.

Results

Contractions

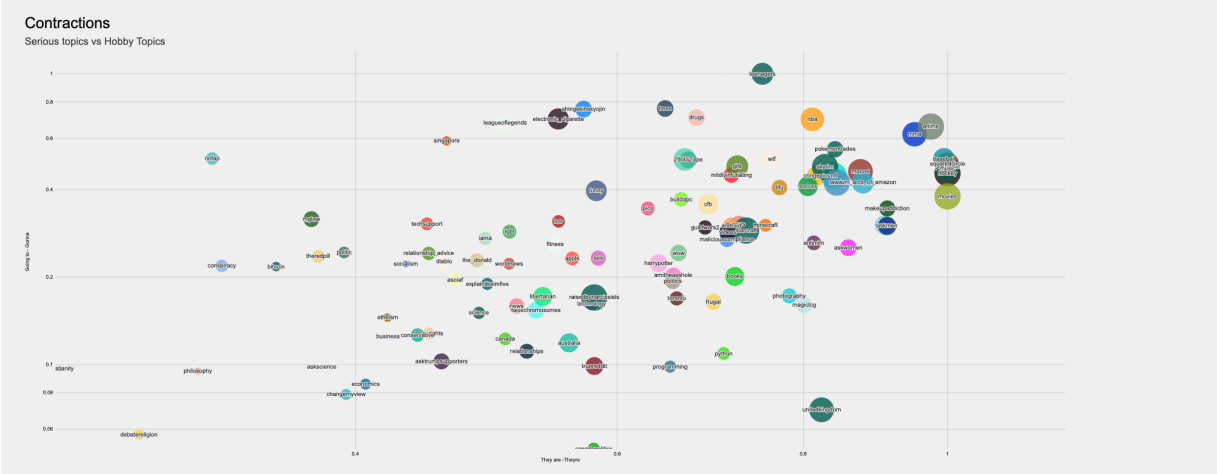

Fig 1: A bubble chart where the X axis is a scale of “They are” to “They’re” and the Y axis is a scale of “Going to” to “Gonna”. The size of the bubble is a scale of “Cannot” to “Can’t”.

This bubble chart helps us analyze the incidence of contractions. Among the contractions we searched for, the three represented in the bubble chart show interesting variation. As we predicted during a preliminary analysis of the data, subreddits that deal with more formal topics

exhibit a low incidence of contractions, and those related to hobbies, general interests, and other informal aspects of the users' lives. Moreover, these trends have uncovered information regarding the age demographic of these subreddits. Particularly extreme results occurred on the low end in r/christianity, r/debatereligion, and r/philosophy. Particularly extreme results occurred on the opposite end of the spectrum in r/anime, r/teenagers, and r/baseball. To see the differences more clearly, we can plot their individual contraction frequencies.

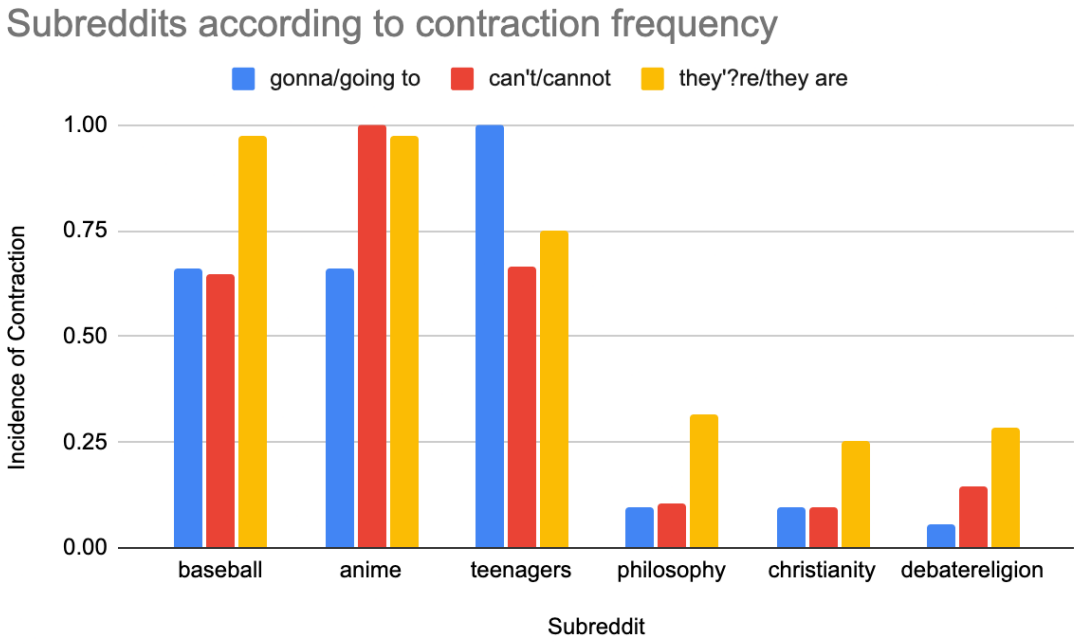

Fig 2: A Bar chart displaying the frequencies of contractions in selected subreddits.

We can clearly see that subreddits associated with more formal topics have a much lower occurrence of these contractions. This is an example of code switching, as when people switch to more formal topics they tend to exhibit more formal speech. Furthermore, these communities likely have very different age demographics. Teenagers explicitly targets a younger audience,

while it is also safe to assume that most of the activity in communities discussing religion and debate can be attributed to an older audience. Now we can observe if the similarity scores of these subreddits confirm our idea that similar subreddits make similar lexical choices. According to the online overlap tool, the overlap of r/anime with r/teenagers is 1.09, while the overlap of r/anime with r/debatereligion is 0.41. These charts and figures confirm not only the specific age and formality predictions we made according to the contraction variable, it also confirms the overarching hypothesis.

British Spellings

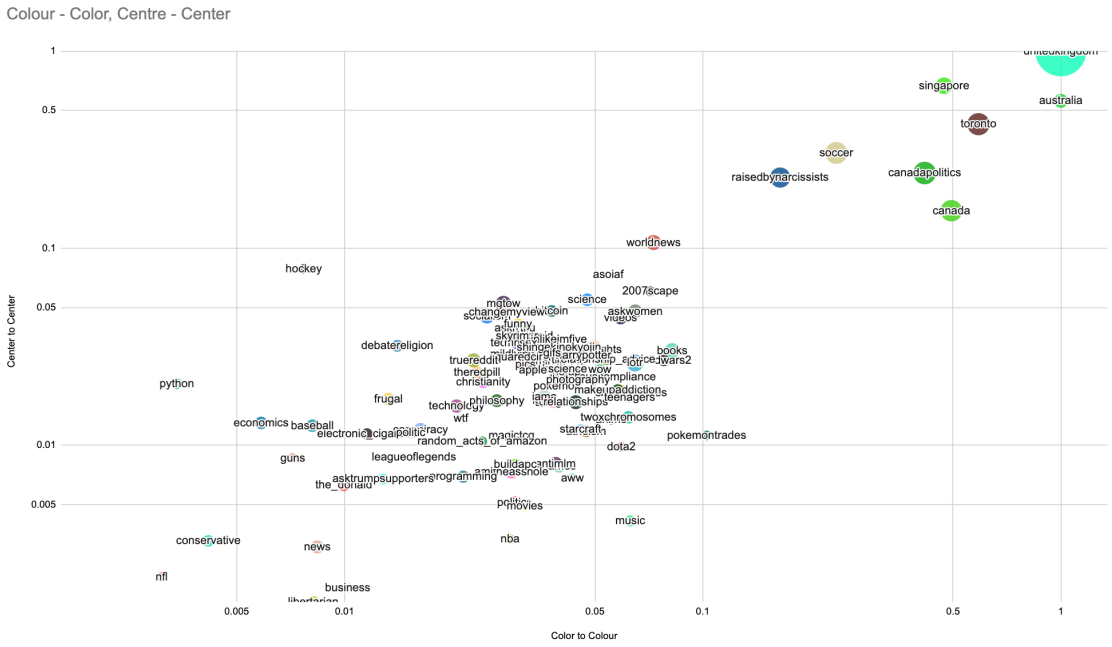

Fig 3: A bubble chart where the X axis is a scale of “Color” to “Colour” and the Y axis is a scale of “Center” to “Centre”. The size of the bubble is a scale of “Labor” to “Labour”.

Now onto the analysis of occurrence of british spellings. In measuring British spellings, we are implicitly gathering a metric for the presence of Americans in a subreddit, as British spellings are more often the standard variety around the world.

Unsurprisingly, r/unitedkingdom showed by far the greatest incidence of British spellings, closely followed by other british-related communities, like r/singapore and r/australia. The subreddits on the opposite end of the spectrum are definitely communities that are traditionally associated with American culture and speech. Particularly pronounced lacks of British spellings occurred in r/nfl, r/libertarian, and r/the_donald. It is worth noting that the extremely high occurrence of the "labour" spelling in British subreddits may be due to the UK having a major political party called "Labour". A closer look at these subreddits reveals their disparity in british spellings.

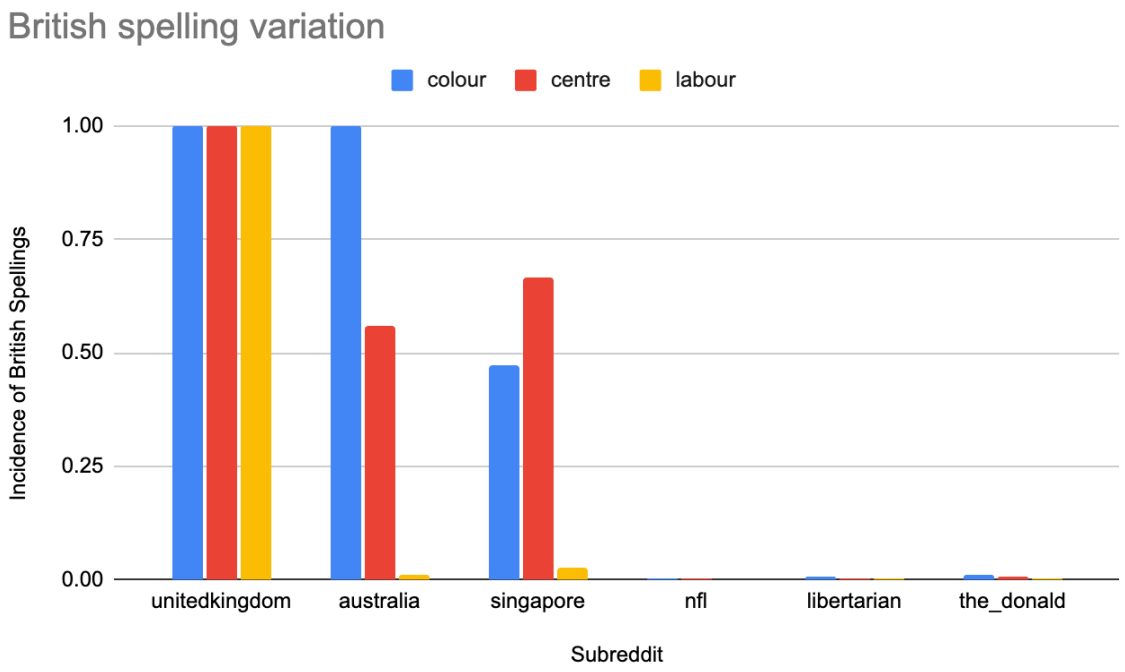

Fig 4: Bar chart illustrating extreme similarities and differences in British spellings among selected subreddits.

This data once again confirms our predictions on the association of British spellings to british users, and american spellings to american users. We can once again confirm this using the online similarity score tool. The subreddit r/unitedkingdom and r/australia have an overlap score of

1.64, while the overlap or r/unitedkingdom with r/nfl was only 0.39. Returning back to our overarching hypothesis, this confirms that similar users make similar linguistic choices.

Acronyms

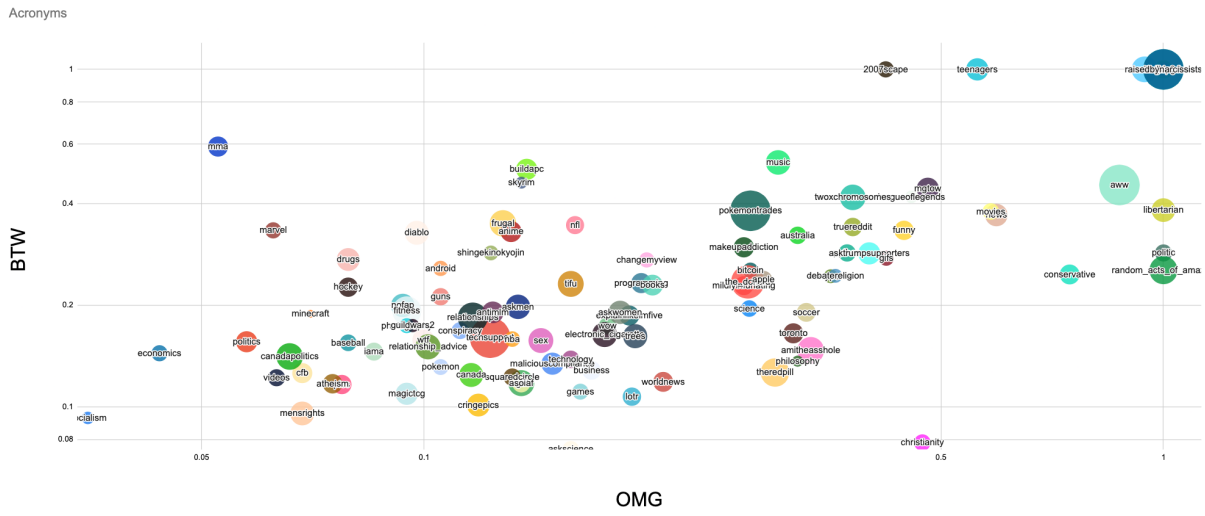

Figure 5: A bubble chart where the X axis is a scale of “oh my god” to “OMG” and the Y axis is a scale of “by the way” to “BTW”. The size of the bubble is a scale of “in real life” to “irl”.

Similar to contractions, we are looking at the incidence of acronyms as marker for informality and a younger demographic. Figure 5 strongly confirms this. Subreddits on the lower end of the scale included r/socialism, r/economics, and r/politics, while those on the higher end included r/raisedbynarcissists, r/starcraft, and r/teenagers. The higher end of the spectrum in this case consisted of a more formal political discussion, while the higher end was informal and hobby based. In addition to code switching, this is potentially indicative of jargon. Many online communities centered around common interests create their own jargon. For example, Bitcoin enthusiasts commonly use the acronym HODL - hold on for dear life. As previously, we will observe selected reddits from the bubble chart and compare their overlap scores.

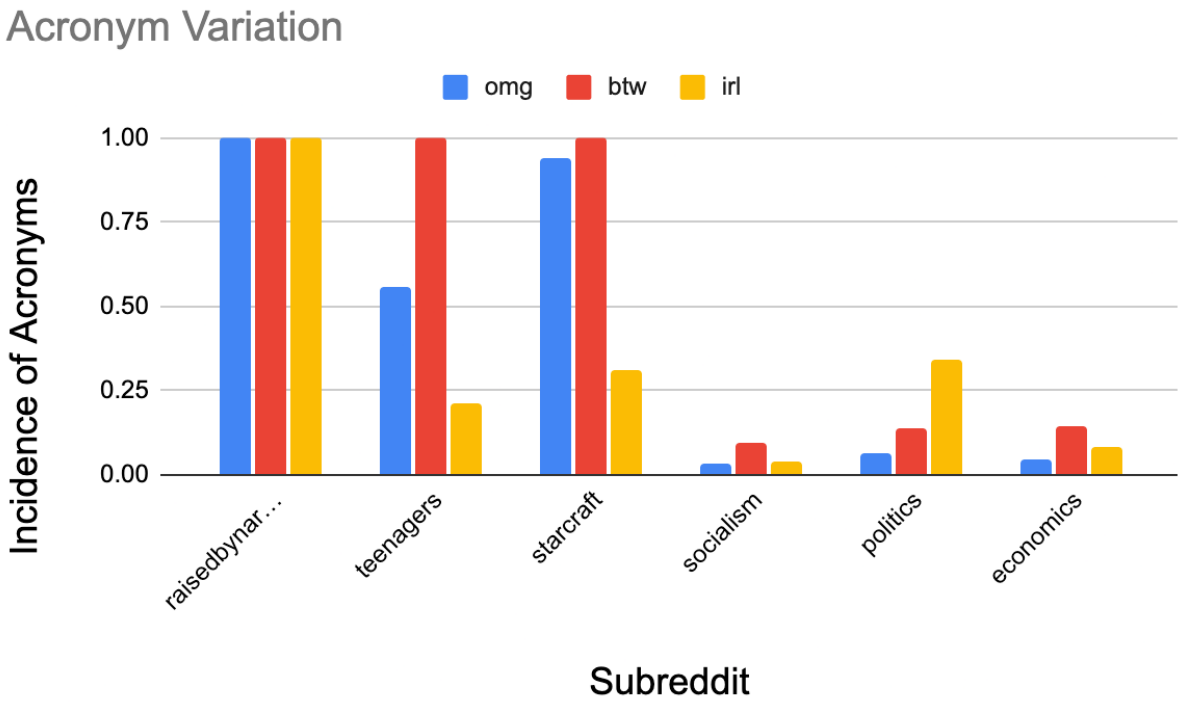

Figure 6: Bar chart representing extreme variation in incidence of acronyms among subreddits.

This figure more clearly demonstrates the differences in incidence of acronyms between informal subreddits aimed at a younger audience and formal discussion subreddits not explicitly targeted at younger people. This variable, however, did not show the same tendency to align with user base similarity scores. The overlap of r/raisedbynarcissists and r/teenagers is only 0.4, while the overlap of r/raisedbynarcissists is 0.8. This shows us that acronyms are not as useful when predicting a subreddit's demographic, or that we simply did not test the right acronyms, leading to somewhat uninformative data.

Regional Variation

We attempted to observe the frequencies of the regional word variants that were put forward in the Data analysis assignment. We calculated their frequencies across all our subreddits, and found that it was not the best way to go about studying regional variation. We

found that a lot of the words were nouns, and nouns occurred so infrequently in our data that when converted to a ratio the vast majority were either the maximum occurrence or, more often, the minimum occurrence. We decided to analyze this data through the lens of a few topic aligned binary comparisons. This allows us to observe more interesting results in the data as we can look at two communities at once, which are related by user base and therefore similar in lexical choices.

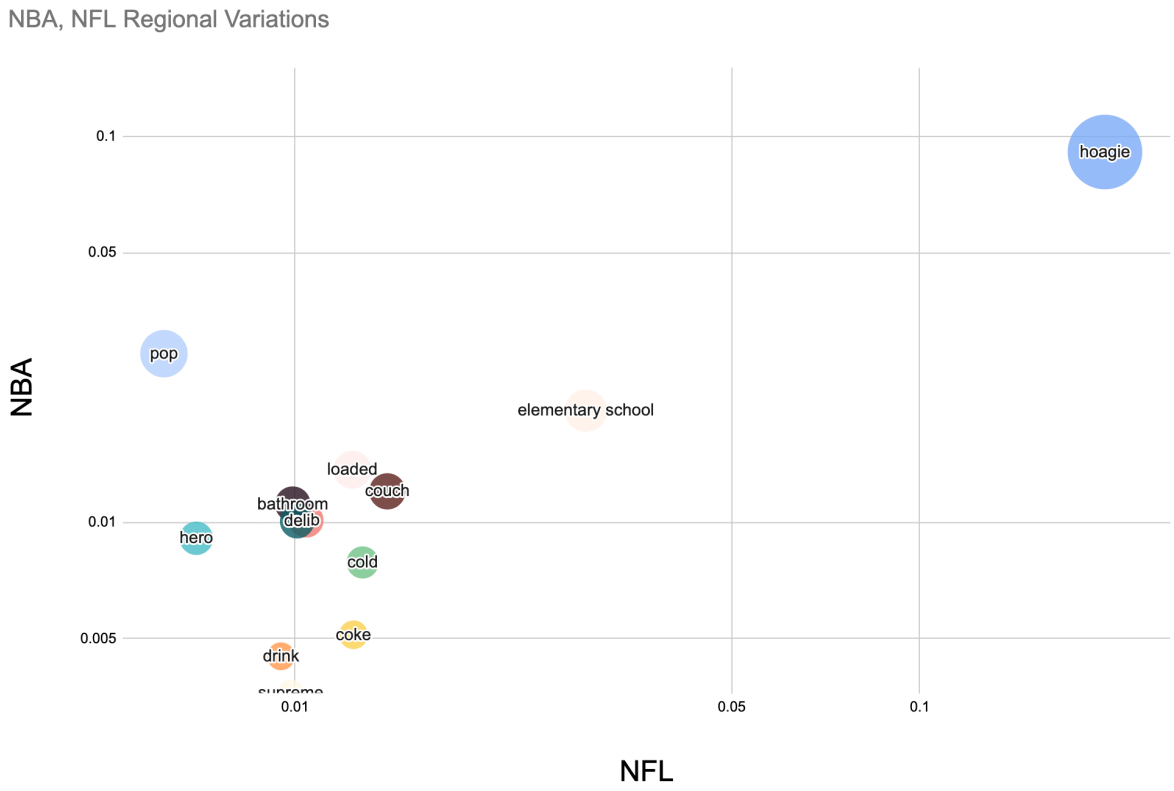

Figure 7: Bubble chart observing regional variation in NFL/NBA

The most salient word in the NFL/NBA chart appears to be hoagie, a term specific to Philadelphia and the surrounding area. This makes sense, as Philadelphia is one of the biggest sports communities in the US, with major teams in every league. It follows that the regional variants specific to that area show up more in sports related subreddits. From this, we could

predict that if we tested more words related to large sports fan bases like Boston, New York, and Chicago, we may have seen similar variation.

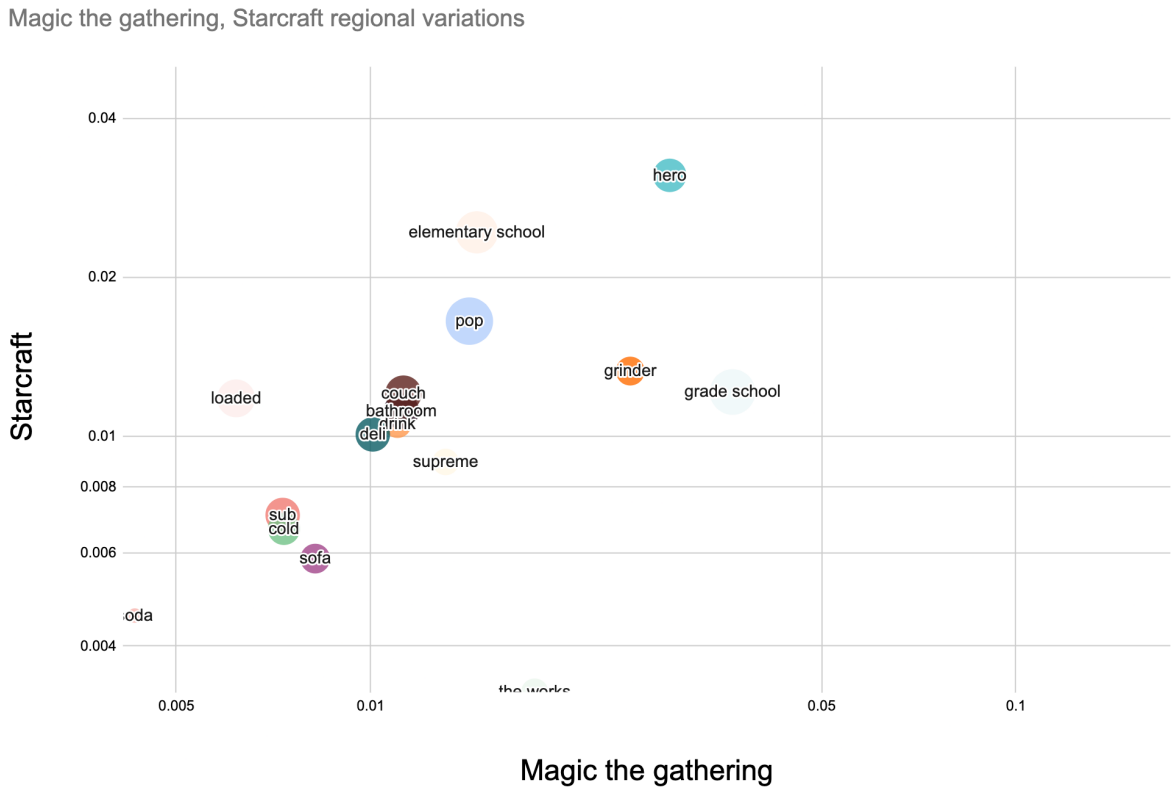

Figure 8: Bubble chart observing regional variation in Magictg/Starcraft

Another interesting comparison is between r/magictg, made for the popular game Magic the Gathering, and r/starcraft, another videogame. We see the word “hero” show up frequently in both communities, as well as “grade school” and “elementary school”. “Hero” is a word often used in video games and in this case is an example of gamer jargon. The presence of “grade school” and “elementary school” confirms that videogames have a younger user base.

Similarity Network

We used the subreddit similarity scores that we mentioned earlier to create a network of similarity. This idea was inspired by the network graph shown in "Conflicts in Distributed Systems." (Kumar et al.), it shows different reddit communities plotted on an imaginary geographical plane. To replicate this for our purposes, we used the low-dimensional reddit-embeddings file provided by the same study, and found the similarity scores of each reddit to every other reddit in our set. After plotting[1], each node was colored based on the frequency of salient British spellings (specifically colour, centre, and labour). This graph provides further insight into where these spellings are being used.

Generally, this data shows that British spellings correlate with reddit similarity. We can see the cluster of American conservatives has a very low use of British spellings. While the closely related group of serious-topic-discussion subreddits (r/science, r/worldnews) are much more likely to use British speech. Across the moat of serious talk are the subreddits of the British commonwealth.

On the sports front, the more American sports (r/nfl, r/nba) are less likely to use British spellings than (r/soccer). These areas of hobby talk are where the correlation between similarity and British frequency fades. We can use the information we know about British spellings to make insights about the users in these subreddits. For example, if we know r/leaugeoflegends and r/starcraft are traditionally similar reddits, but they exhibit differences in their frequency of british spellings, we can make the assumption that there are more american speakers that are interested in League of Legends than StarCraft. This isn't a perfect correlation, but it's a really interesting example of the type of sociolinguistic observations that can be made using just lexical variation.

We did this same type of graph with contractions, and similar observations can be made[2]. Contractions correlated even more with similarity in that case.

Conclusions

[1] See appendix A
[2] See appendix B

Overall, each variable we investigated aligned with our general predictions. Acronyms and contractions were correlated with topics that are associated with younger audiences and informal discussions. Informal language could be found in informal topics, as well as in informal mannerisms. Someone could be speaking informally about a formal subject and vice versa, but our metrics track both. British spellings were an accurate metric for how related a subreddit was to the UK, and could tell us a lot about the proportion of the community that was American as well.

Our analysis of regional variation showed some interesting parallels in lexical choices of related communities, however, given the data we had and the regional variation markers we were testing, we could not use the regional variation data as confidently to make predictions. Our results were not decisive enough to give insights on covert prestige, or significant differences in the geographic backgrounds of the user bases.

To widen the scope of this investigation, it could be interesting to observe how these factors correlate with upvote count which could provide valuable insights on how communities reward these linguistic choices.

Our observations concerning contractions, acronyms, British spellings, and regional variations all support our overarching hypothesis. Online communities with similar user bases tend to exhibit similar lexical choices, and it is possible to make predictions of a subreddit's demographic from the language used on it.

Bibliography

Byrne, D. (1971) The Attraction Paradigm. Academic Press, New York.
Turner, J. C., Hogg, M. A., Oakes, P. J., Reicher, S. D., & Wetherell, M. S. (1987). Rediscovering the social group: A self-categorization theory. Basil Blackwell.

Kumar, Srijan, Jure Leskovec, and William L. Hamilton. "Conflicts in Distributed Systems." Proceedings of the 2018 World Wide Web Conference, 2018, pp. 745-754.

Trudgill, Peter. "Sex, covert prestige and linguistic change in the urban British English of Norwich." Language in Society, vol. 7, no. 2, 1978, pp. 179-195.

Appendices
Appendix A - similarity network graph - British
Appendix B - similarity network graph - Contractions
Appendix C - Code