Language Usage in Media

Owen Bernstein and Lindsey Greenhill

12/14/2020

Project Introduction

Amid rising coronavirus cases, nationwide protests calling for racial justice and an end to police brutality, and a contentious election, more Americans than ever before are tuning in to watch the news. Despite covering the same events, however, newscasters have never told such differing stories. The purpose of this project is to understand the differences in this news coverage and examine what factors influence how the news is reported. We became interested in this topic after hearing and reading about increasing media polarization between conservatives and liberals. Per work from the Pew Research Center, Americans are increasingly watching news channels that are seen as in line with their own political views, and avoiding those that are not. The result is a media industry that is, for the most part, divided by political ideology. With this in mind, we wanted to see if news channels that are seen as leaning left or right present meaningfully different news coverage. To explore this question, we are specifically measuring the usage of five types of language – populist, environmental, progressive, conservative, and immigration related – across three prominent cable news channels (Fox News Channel, MSNBC, and CNN) in the weeks just before and after the presidential election.

Initial Hypothesis

Our initial hypothesis is that Fox News, the most conservative of the news channels according to a 2017 Stanford University study will use more conservative, populist, and immigration related language while the more liberal news sources, CNN and MSNBC, will use more progressive and environmental related language. We anticipate these results because Donald Trump, and to a lesser extent the Republican party in its entirety, has frequently been portrayed as populist, has a strong association with immigration and immigration policy, and is relatively conservative. Therefore we expect that Fox News will use more language language relating to these topics. On the other hand, Joe Biden and the Democratic party have a stronger association with environmental policy and are relatively progressive. For this reason we expect CNN and MSNBC to use more language relating to these content categories.

Data and Project Design

In order to assess potential differences in news coverage among cable news channels, we used a data set created by the Internet Archive's Third Eye Project. This data set contains news chyrons - the scrolling captions at the bottom of broadcast images - for each minute of broadcast for four different news channels (MSNBC, CNN, Fox News, and BBC). Although it would be ideal to work with full transcripts of news broadcasts, we did not have access to that data, and we believe chyrons are reasonably representative of each channel's coverage. We only included data from MSNBC, CNN, and Fox News, because the data for BBC was problematic in its transcription. The data includes coverage from October 17, 2020 to November 14, 2020. This research design is cross-sectional meaning that the independent and dependent variables are measured at the same time. Inclusion in this project is based on the data in the main data set.

We performed textual analysis to classify the language each channel used into five categories: populist, environmental, progressive, conservative, and immigration Related. We classified the language using preexisting dictionaries taken from a 2011 study of partisan language in Belgium. It would have been better to use dictionaries created for American politics, but we still think that the ones in the study are relevant and therefore useful. See the word baskets section for more details on the dictionaries.

We ran a linear regression to see if there was a statistically significant relationship between channel and language usage. We also included binary variables of whether or not the coverage was before the election and whether or not the coverage was in prime time in the regression. Every chyron after November 3, 2020 is considered to be post election. Every chyron between 8PM and 11PM is considered to be prime time.

Word Baskets

As mentioned above, we used word baskets developed in a different study to classify the language in the chyrons as populist, conservative, progressive, immigration related, or environmental. If the word matched any of the words in the word basket, it was counted an an instance of that type of language. The word baskets are as follows:

Populist Basket: deceit, treason, betray, absurd, arrogant, promise, corrupt, direct, elite, establishment, ruling, caste, class, mafia, undemocratic, politics, political, politicize, politician, progaganda, referendum, regime, shame, admit, tradition, people

Conservative Basket: belief, family, church, norm, porn, sex, values, conservative, conservatism, custom

Progressive Basket: progress, right, freedom, self-disposition, handicap, poverty, protection, honest, equal, education, pension, social, weak

Environment Basket: green, climate, environment, heating, durable

Immigration Basket: asylum, halal, scarf, illegal, immigrant, immigration, immigrate, Islam, Koran, Muslim, foreign

Again, these dictionaries were originally created for politics in Belgium, and therefore we see some words that might not be as applicable in the U.S. For example, immigration related language in the US might in reality focus relatively more on immigration from Mexico rather than using words such as "Islam" or "Koran".

The table below shows the **total** language counts for each category above.

Language Sentiment Across Channels
October 17 to November 14

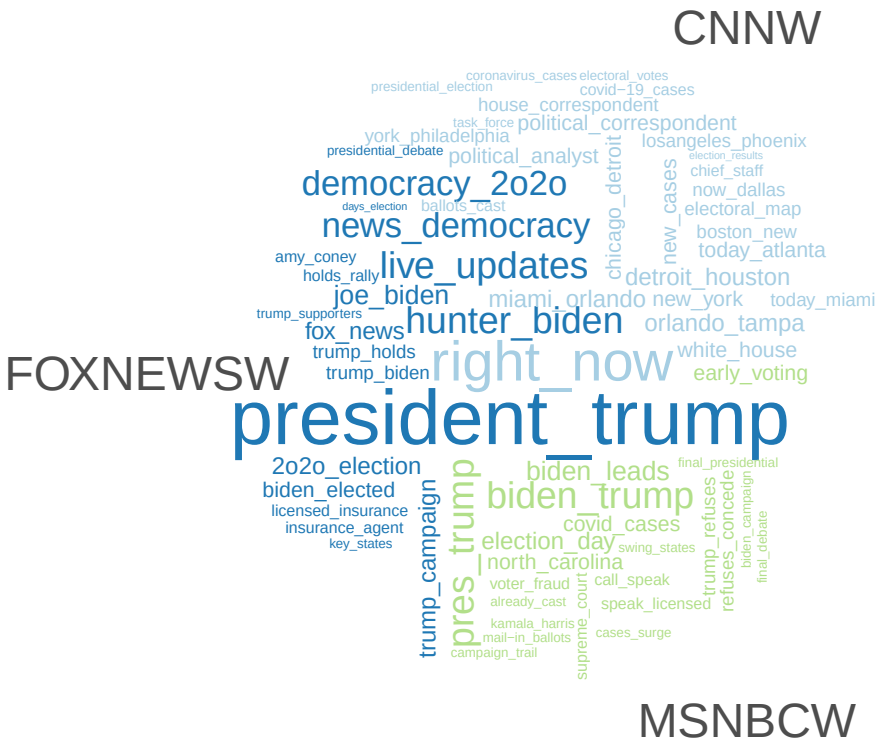
	Word Count for Each Content Category				
Channel	Populism	Environment	Immigration	Progressive	Conservative
CNNW	1022	29	98	940	80
FOXNEWSW	163	15	74	107	37
MSNBCW	219	24	77	80	94

Results - Exploratory Analysis

We used the quanteda package to perform initial textual analysis of the news chyrons. We also looked at the distributions of language usage across channels. Overall, our findings suggest that our initial hypothesis is incorrect.

Word Cloud Analysis

The graphic below shows a word cloud graphic that compares the language usage between Fox News, CNN, and MSNBC. In our analysis, we excluded filler words, punctuation, etc. and looked for two word phrases. Fox News is dark blue, CNN is light blue, and MSNBC is green.



Discussion

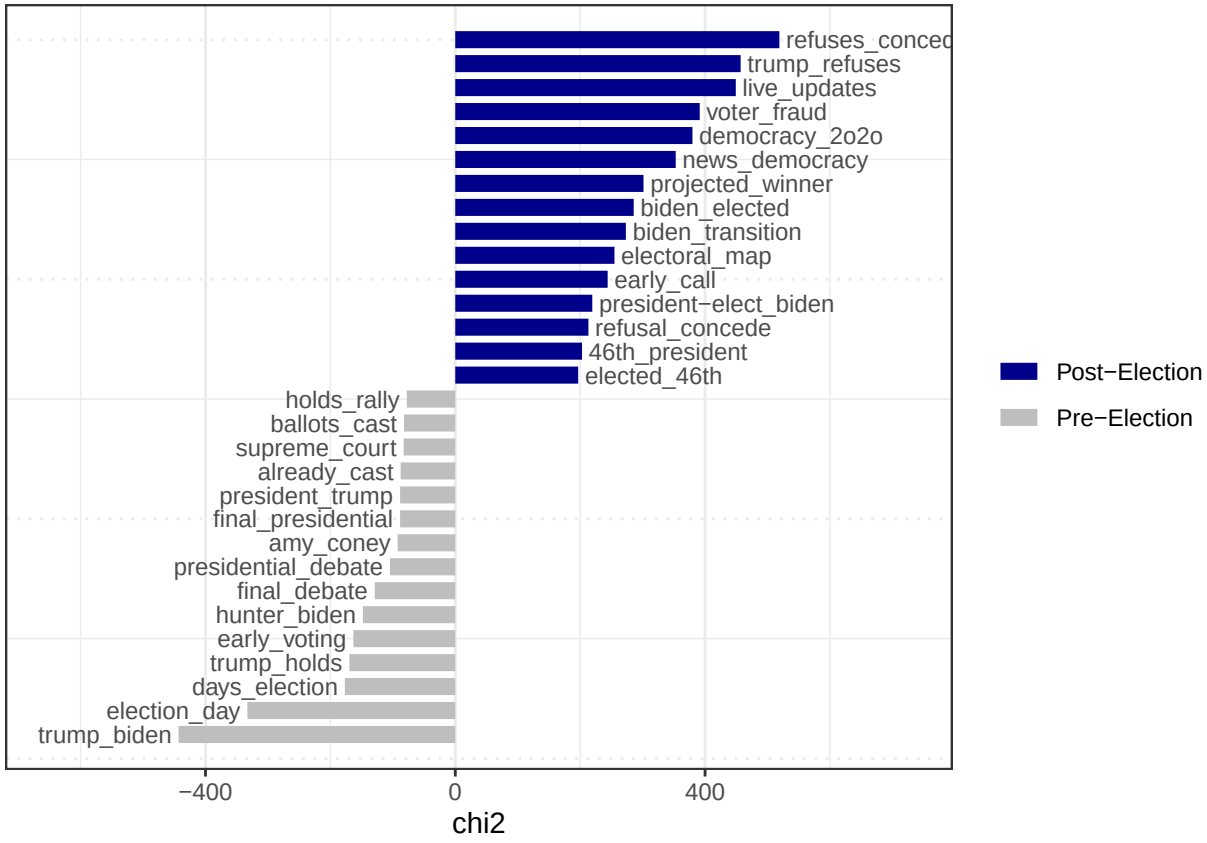
The most mentioned phrase among all the networks is President Trump (from Fox), which makes sense given how close our time frame is to the 2020 election. Fox also seemed to focus on Hunter Biden, Joe Biden, and democracy. CNN's most used phrase was "right now," which is perhaps an indication of the tone of their news coverage. CNN also mentions cities such as Chicago, Detroit, Houston, Miami, and a few more. MSNBC also mentions Trump often. It also used the phrases ""biden leads," "north carolina," "election day," and "covid cases" frequently.

Keyness Plot Analysis

In addition to word cloud analysis, we also looked at keyness plots of the news chyrons. A keyness plot is a plot that compares the usage of words between two different data sets. We wanted to look at the potential impact of our post_election and prime_time variables. We created a keyness plot for each variable. The first keyness plot looks at the difference in language use from all channels between pre and post election coverage. The second keyness plot looks at the difference in language use from all channels between prime time and non prime time coverage.

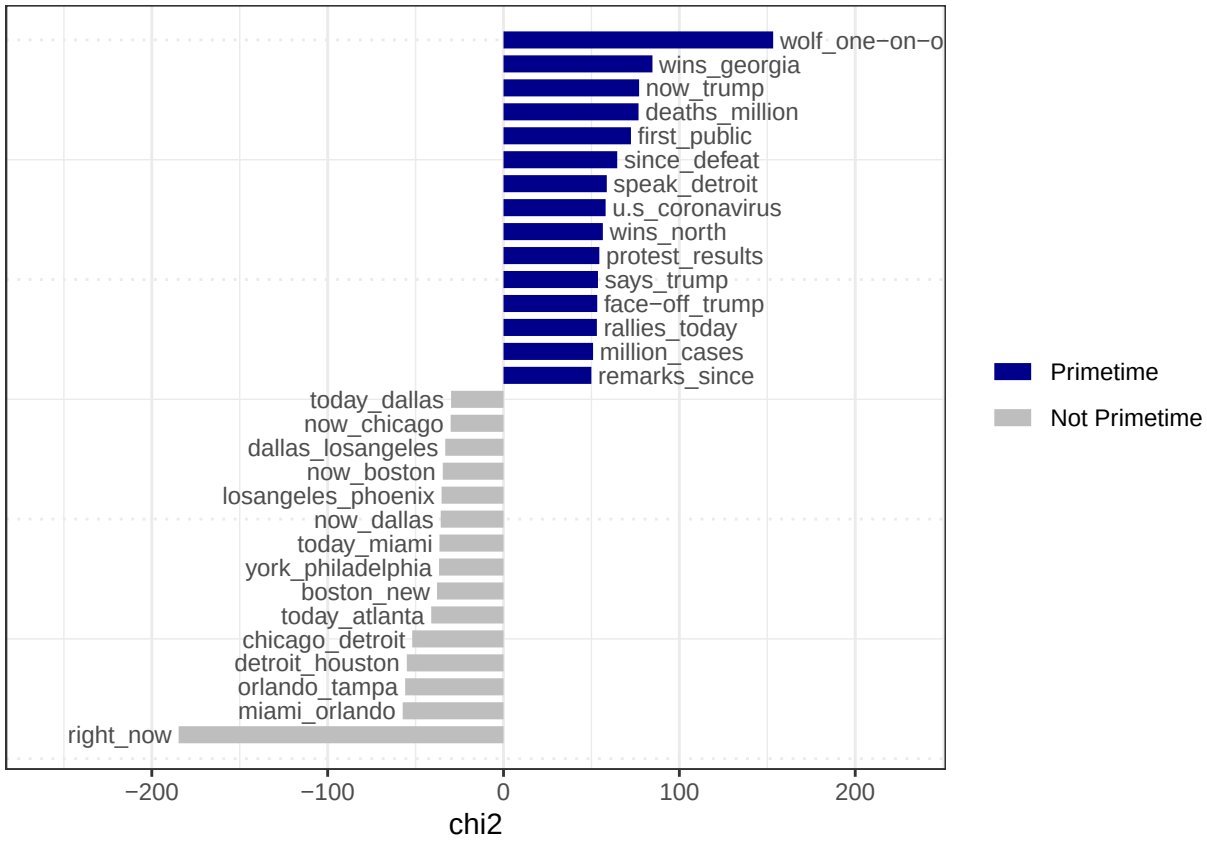
Election Keyness Plot

The differences in language usage pre and post election are intuitive. Before the election, news coverage focused much more on campaigns and voting processes. After the election, news coverage focused more on calling the election for Biden and the white house transition from Trump to Biden. Overall, there seems to be a noticeable shift in topics between pre election and post election coverage.



Prime Time Keyness Plot

We thought it could be interesting to look at prime time coverage vs non prime time coverage because of the salience of prime time coverage. Additionally, the respective hosts for each channel's prime time news shows tend to be particularly divisive politically which we thought might have an impact on the language used during their shows. Interestingly, non prime time coverage seems to focus more on cities compared to prime time coverage. This may be because non prime time coverage discusses specific events and locations while prime time coverage tends towards national issues. Prime time coverage seems to focus on the election compared to non prime time coverage.

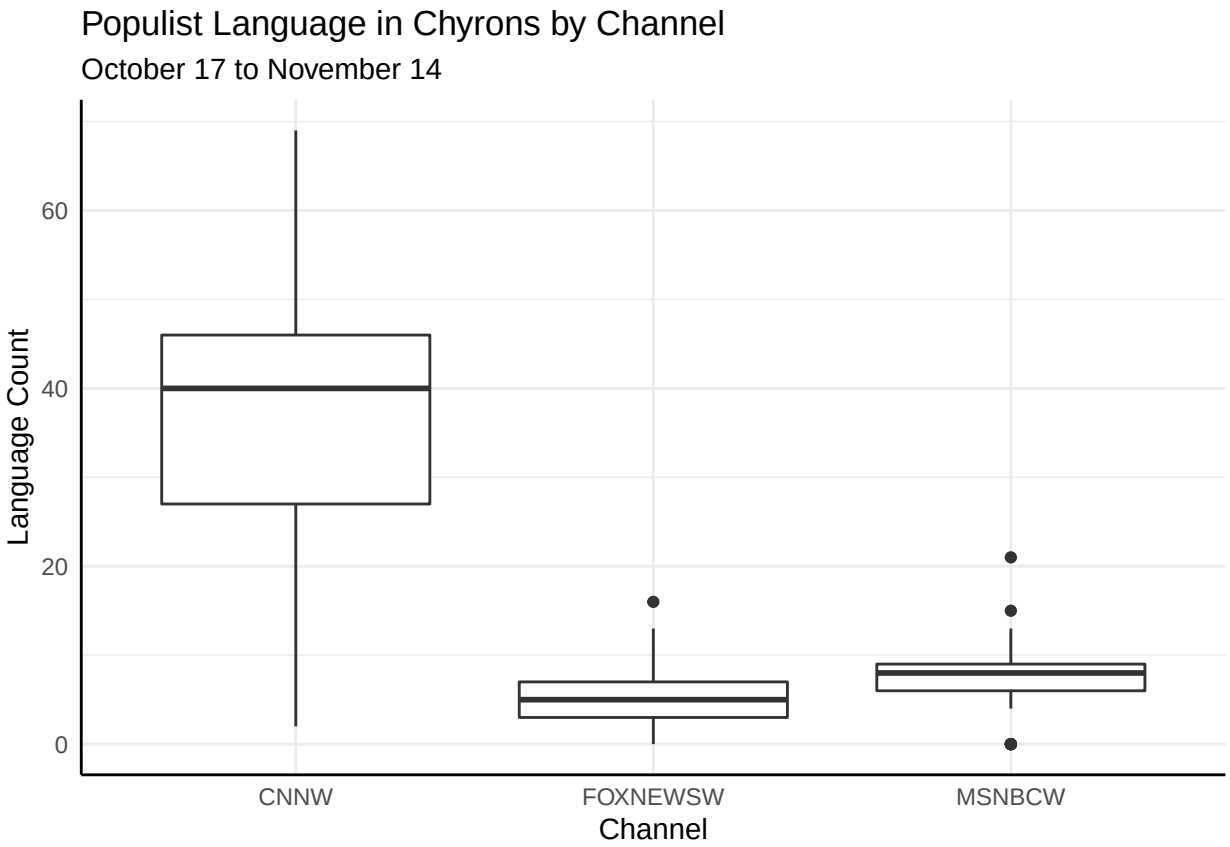


Boxplot Analysis

Before running a regression, we wanted to visualize the distribution of language usage across channels to get an initial sense of the trends.

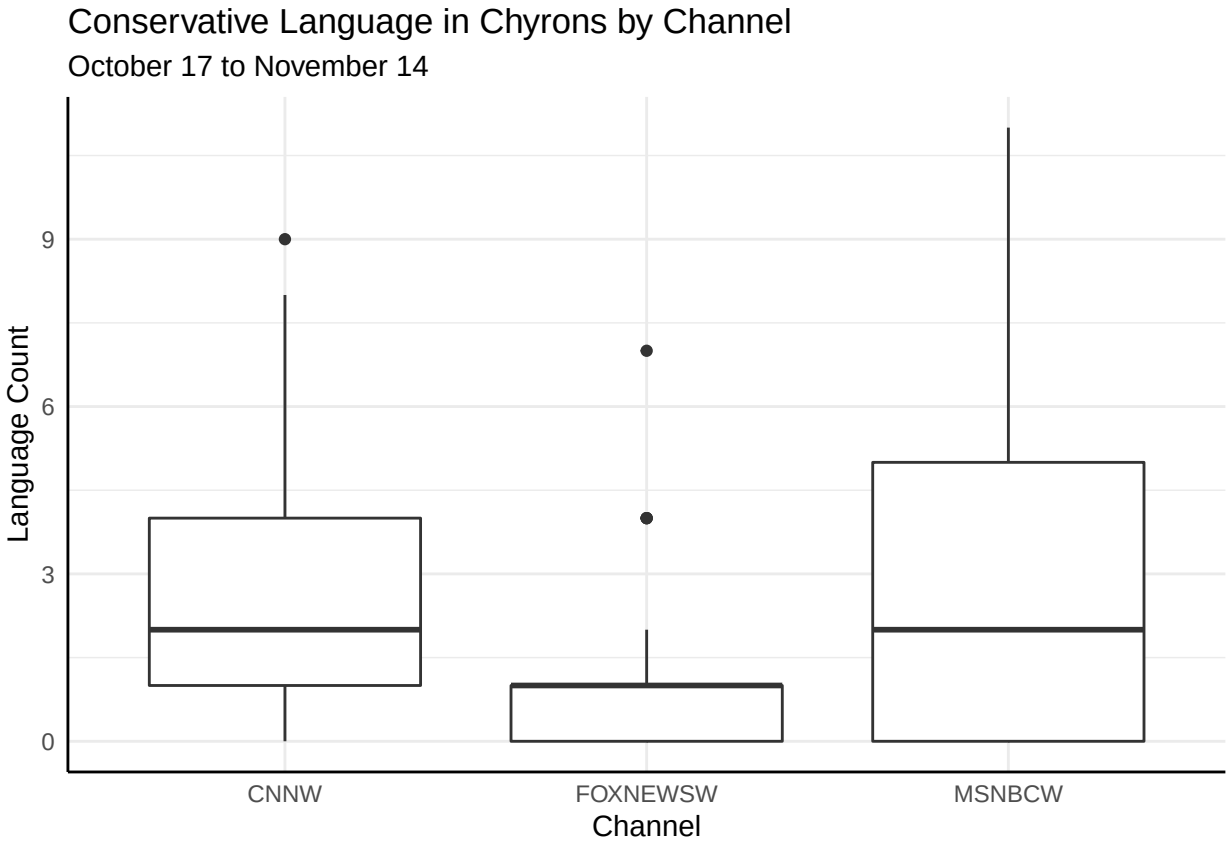
Populism Boxplot

The boxplot below shows the distribution of **daily** populist language counts for each channel. From this graphic, it appears that CNN uses more populist language on average compared to Fox News or MSNBC. This finding is the opposite of what we expected in our initial hypothesis.



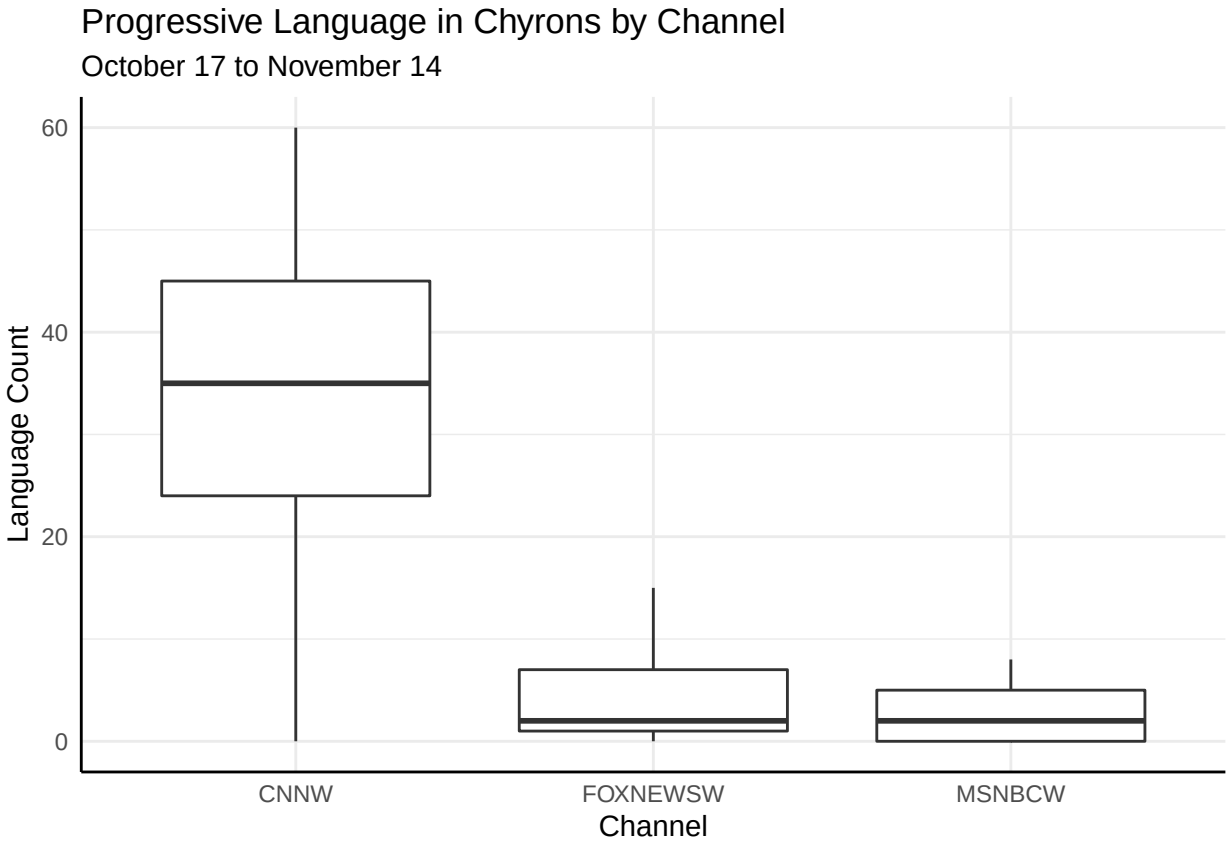
Conservative Boxplot

The boxplot below shows the distribution of **daily** conservative language counts for each channel. From this graphic, it appears that CNN and MSNBC uses more conservative language on average compared to Fox News. This finding is the opposite of what we expected in our initial hypothesis.



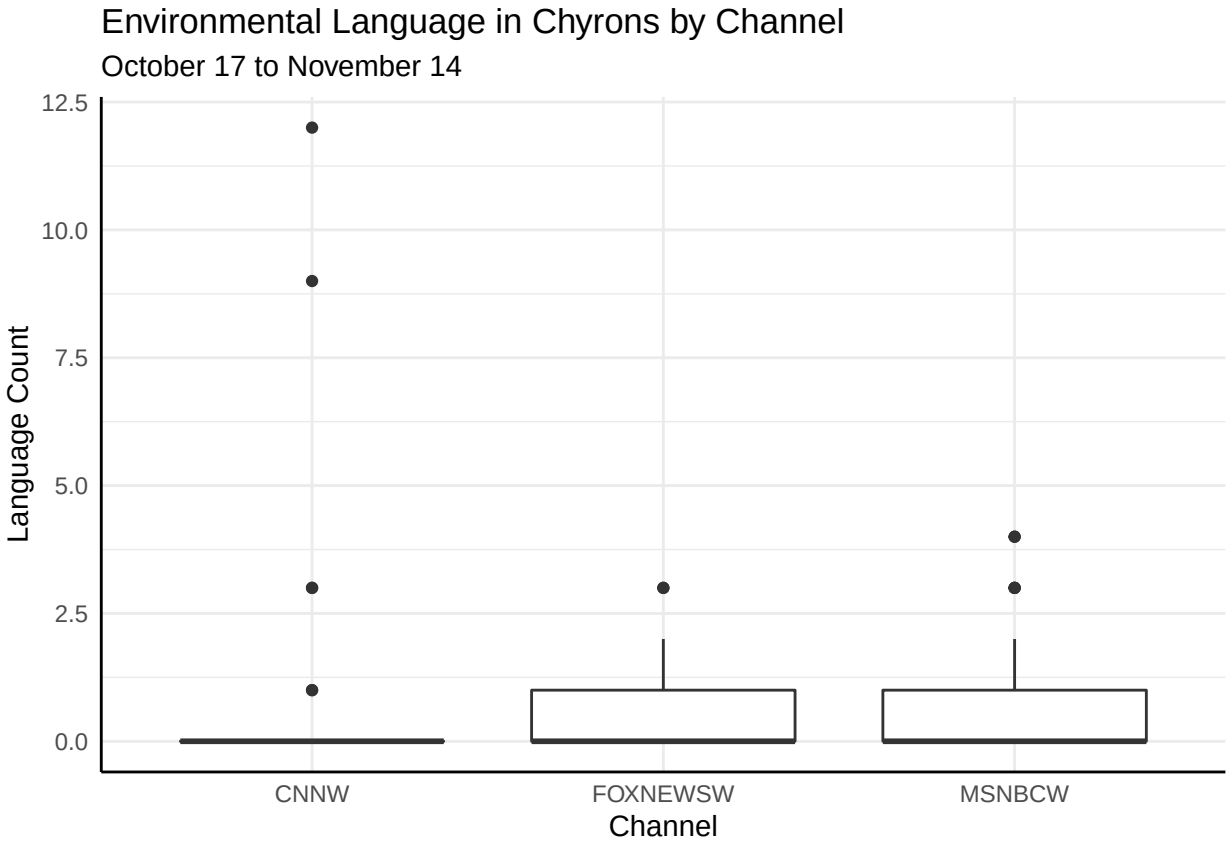
Progressive Boxplot

The boxplot below shows the distribution of **daily** progressive language counts for each channel. From this graphic, it appears that CNN uses more progressive language on average compared to Fox News or MSNBC. This finding is in line with what we expected in our initial hypothesis.



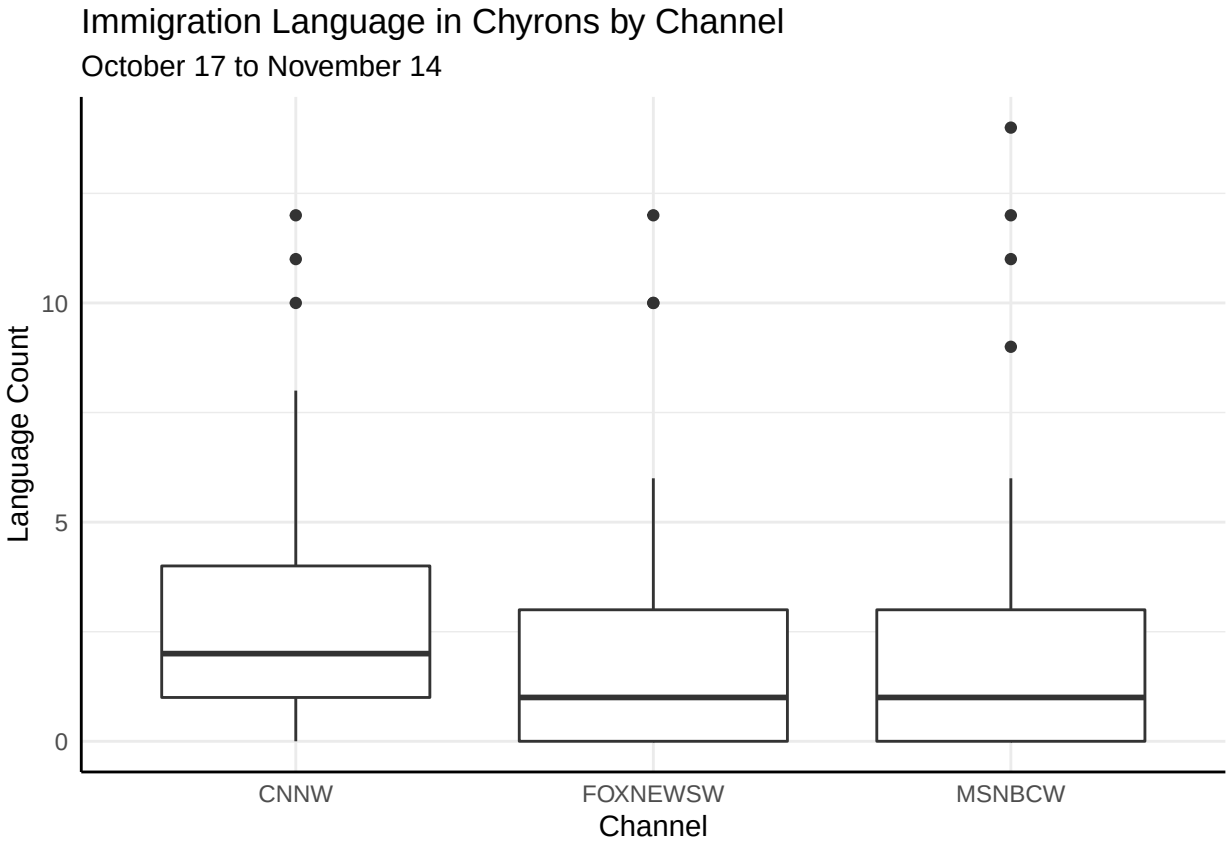
Environment Boxplat

The boxplot below shows the distribution of **daily** environmental language counts for each channel. From this graphic, there doesn't appear to be a news channel that uses environmental language more than other news channels. Moreover, it appears that the use of environmental language is in general quite infrequent.



Immigration Boxplot

The boxplot below shows the distribution of **daily** immigration related language counts for each channel. From this graphic, it appears that CNN uses slightly more immigration related language on average compared to Fox News or MSNBC. This finding is not in line with what we expected in our initial hypothesis.

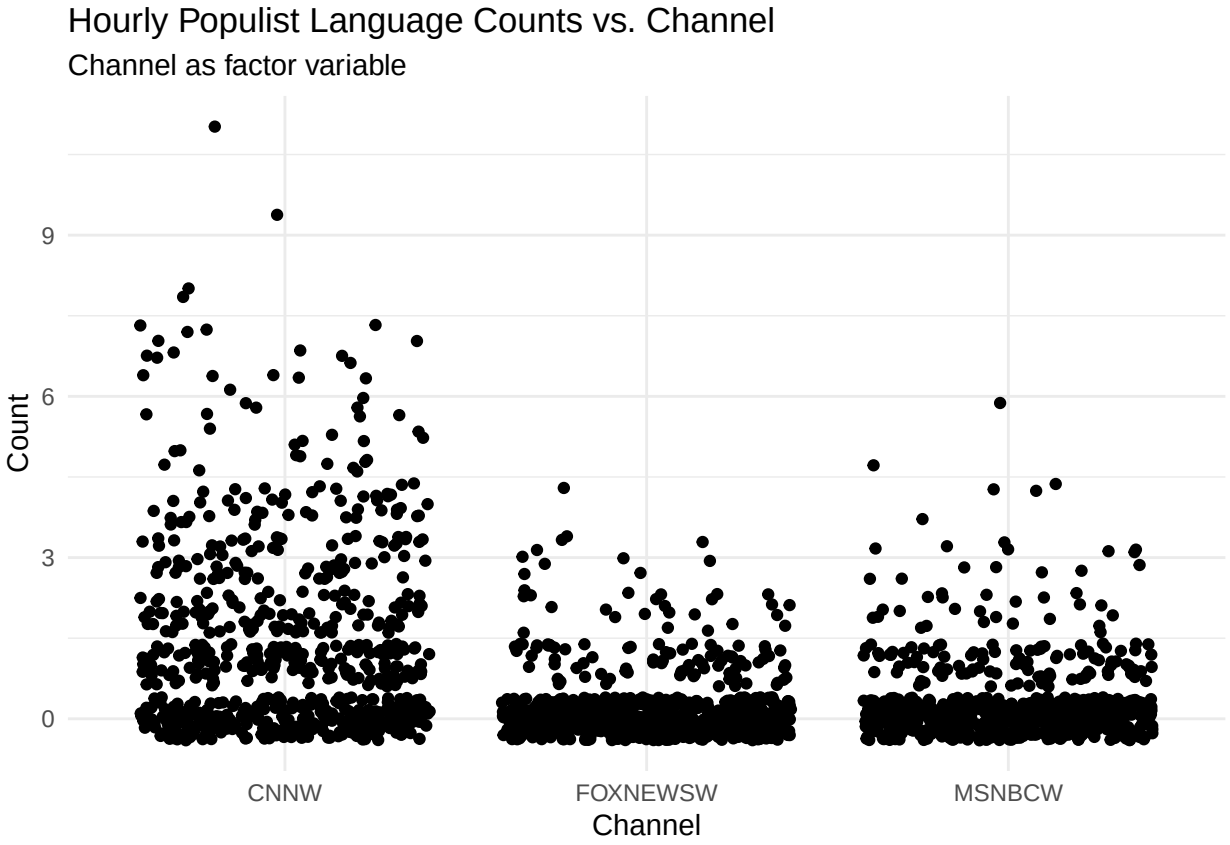


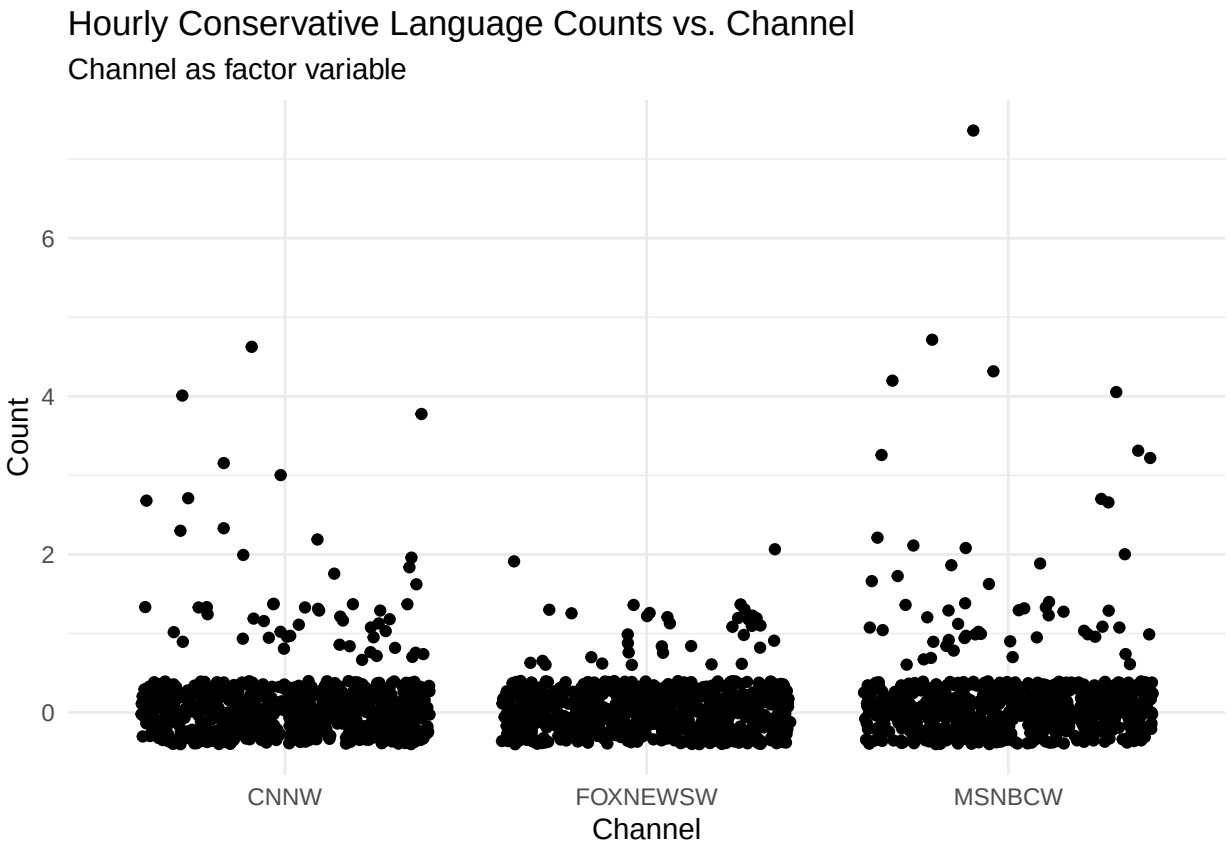
Results - Regression Analysis

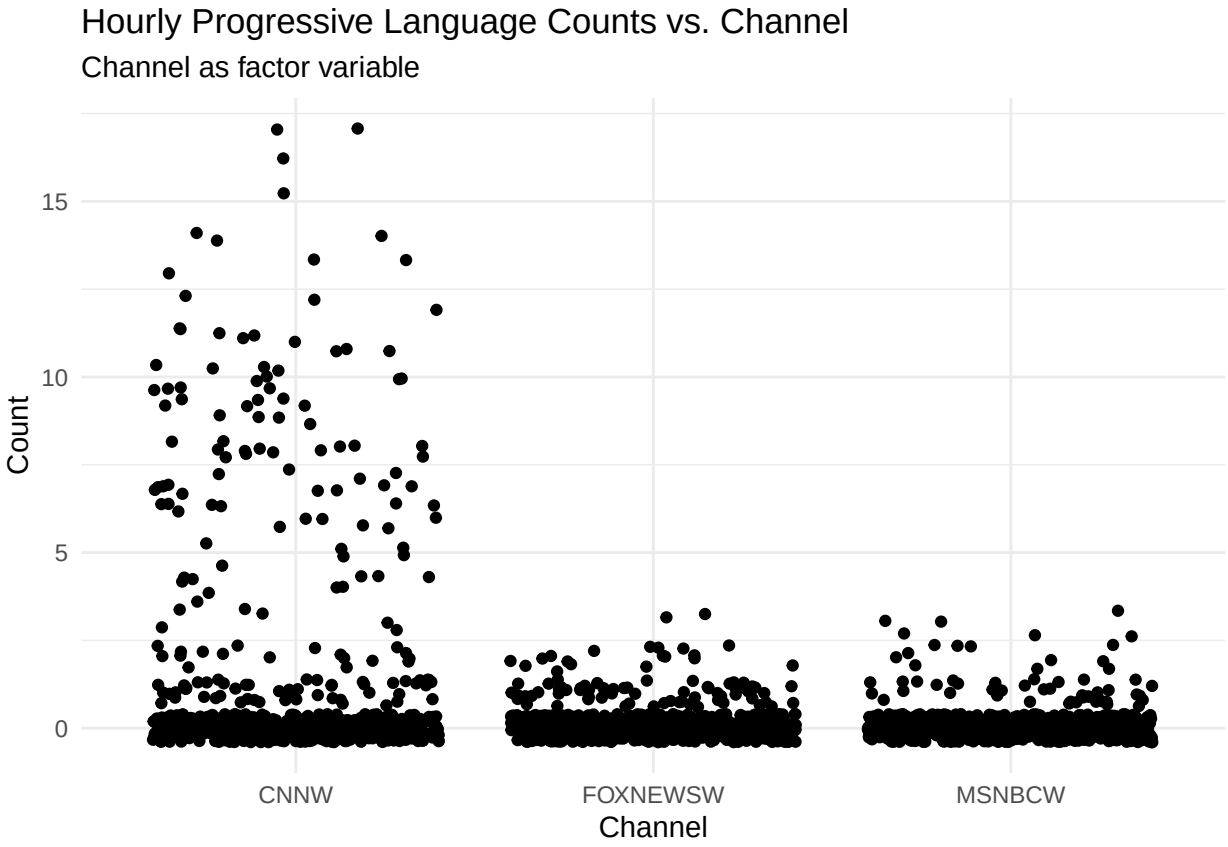
We ran five regressions for the five language classifications (populist, conservative, progressive, environmental, immigration) to look at the relationship between language usage and channel. The dependent variable in our regression is hourly language counts. The independent variables are channel (as a factor variable), prime time (as a binary variable), and pre election (as a binary variable).

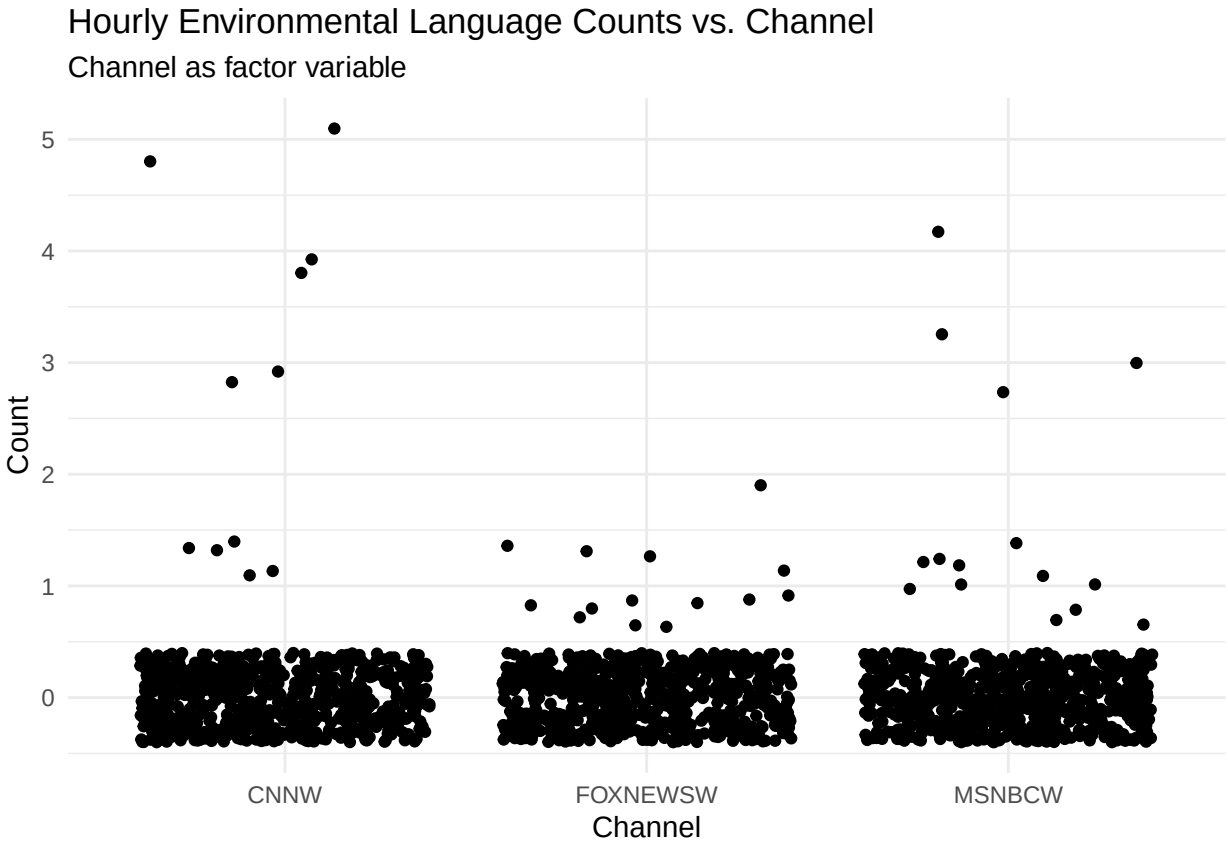
Scatterplots

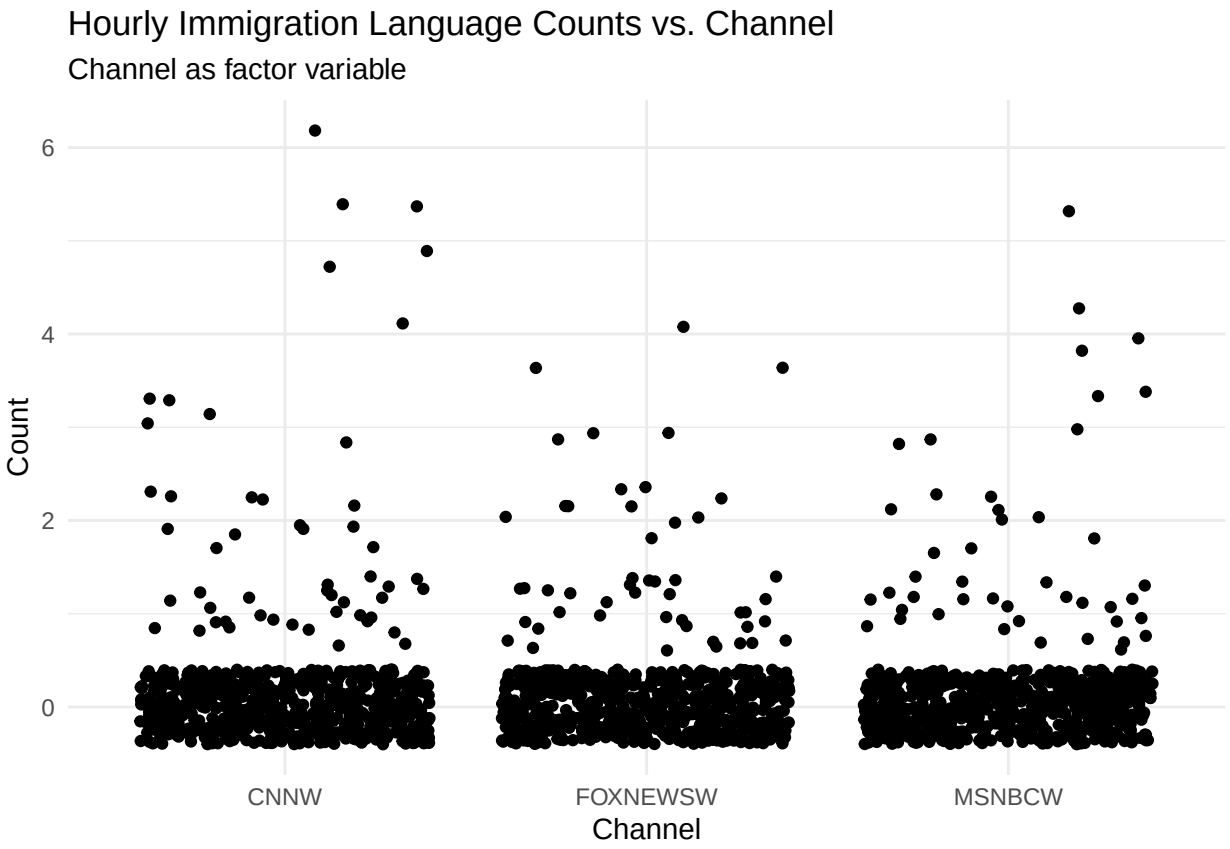
The following scatterplots visualize the relationship between language usage and channel.




Hourly Conservative Language Counts vs. Channel
Channel as factor variable
Count
6
4
2
0
CNNW
FOXNEWSW
MSNBCW
Channel


Hourly Progressive Language Counts vs. Channel
Channel as factor variable
Count
15
10
5
0
CNNW
FOXNEWSW
MSNBCW
Channel


Hourly Environmental Language Counts vs. Channel
Channel as factor variable
Count
5
4
3
2
1
0
CNNW
FOXNEWSW
MSNBCW
Channel


Hourly Immigration Language Counts vs. Channel
Channel as factor variable
Count
6
4
2
0
CNNW
FOXNEWSW
MSNBCW
Channel

Regression Results

The table below shows the results of five different regressions.

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Mon, Dec 14, 2020 - 14:07:04

	Dependent variable:				
	populism	immigration	environment	progressive	conservative
	(1)	(2)	(3)	(4)	(5)
as.factor(channel)FOXNEWSW	−1.240*** (0.063)	−0.035 (0.028)	−0.020 (0.016)	−1.202*** (0.099)	−0.062*** (0.024)
as.factor(channel)MSNBCW	−1.152*** (0.063)	−0.029 (0.028)	−0.007 (0.016)	−1.240*** (0.099)	0.023 (0.024)
as.factor(election)Pre-Election	0.341*** (0.053)	0.086*** (0.024)	0.006 (0.013)	0.214** (0.083)	0.077*** (0.020)
as.factor(primetime)Primetime	0.199*** (0.069)	−0.029 (0.031)	−0.011 (0.017)	−0.426*** (0.109)	0.091*** (0.026)
Constant	1.230*** (0.056)	0.093*** (0.025)	0.040*** (0.014)	1.296*** (0.089)	0.052** (0.022)
Observations	2,071	2,071	2,071	2,071	2,071
R^2	0.206	0.008	0.001	0.098	0.019
Adjusted R^2	0.205	0.006	−0.001	0.096	0.017
Residual Std. Error (df = 2066)	1.168	0.525	0.290	1.844	0.448
F Statistic (df = 4; 2066)	134.162***	3.963***	0.605	56.010***	9.848***
Note:					*p<0.1; **p<0.05; ***p<0.01

Discussion

Populist Regression - The results suggest that Fox News and MSNBC use less populist language than CNN (a result supported by the former boxplots). Both coefficients on the news channels are statistically significant at a .01 level. The pre-election coefficient is positive, suggesting that news coverage on average uses more populist language pre-election holding channel and primetime constant. This coefficient is also statistically significant at a .01 level. The primetime coefficient is positive, suggesting that news coverage on average uses more populist language during prime time holding channel and pre-election constant. This coefficient is statistically significant at the .01 level.

Immigration Regression - The results suggest that Fox News and MSNBC use less immigration related language than CNN. However, neither of these coefficients are statistically significant. The pre-election coefficient is positive, suggesting that news coverage on average uses more immigration related language pre election holding channel and primetime constant. This coefficient is statistically significant at a .01 level. The primetime coefficient is negative, suggesting that news coverage on average uses less immigration related language during prime time holding channel and pre-election constant. However, this coefficient is not statistically significant.

Environment Regression - The results suggest that Fox News and MSNBC use less environmental language than CNN. However, neither of these coefficients are statistically significant. The pre-election coefficient is positive, suggesting that news coverage on average uses more environment related language pre election holding channel and primetime constant. This coefficient is not statistically significantl. The primetime coefficient is negative, suggesting that news coverage on average uses less environment related language during prime time holding channel and pre-election constant. However, this coefficient is not statistically significant.

Progressive Regression - The results suggest that Fox News and MSNBC use less progressive language than CNN. Both of these coefficients are statistically significant at the .01 level. The pre-election coefficient is positive, suggesting that news coverage on average uses more progressive language pre election holding channel and primetime constant. This coefficient is statistically significant at a .05 level. The primetime coefficient is negative, suggesting that news coverage on average uses less progressive language during prime time holding channel and pre-election constant. This coefficient is statistically significant at the .01 level.

Conservative Regression - The results suggest that Fox News uses less conservative language than CNN while MSNBC uses more conservative language than CNN on average. Only the Fox News coefficient is statistically significant. The pre-election coefficient is positive, suggesting that news coverage on average uses more conservative language pre election holding channel and primetime constant. This coefficient is also statistically significant at a .01 level. The primetime coefficient is positive, suggesting that news coverage on average uses more conservative language during prime time holding channel and pre-election constant. This coefficient is statistically significant at the .01 level.

Causality: These results should not be interpreted as causal for a few reasons. First, the main relationship of interest is between channel and language usage. The channel variable by nature cannot be changed, and therefore, it cannot be considered as a treatment of any kind. Second, there are a variety of confounding variables that could have an effect of the results such as viewership, news host, etc.

Conclusion

Ultimately, this project has shown us that there does appear to be significant differences in the language used by cable news channels, but that these differences are not predicted by ideology. In four of five regressions predicting hourly language usage by channel there was a statistically significant difference. This shows that the channels do in fact use these specific types of language in varying amounts. Furthermore, in four of five regressions there was a significant difference in hourly language usage based on whether coverage was before or after the election and in three of five regressions there was a significant difference based on whether or not the coverage was during prime time. However, the relationships between channel and language usage by content category did not align with our initial hypothesis. In three of five regressions our initial hypothesis incorrectly predicted the direction of the relationship between channel and language usage. This suggests that while there is a difference in language usage by channel, it is not determined by the channel's ideology. Analysis of this study is limited by the research design. Specifically, the basket of words used in the study is designed for application in Belgian politics rather than American politics. Additionally, the basket of words approach only counts the instances of language usage rather than the context in which the word was used which can be important. This analysis could be improved with more complete transcripts of news coverage and a basket of words specifically designed for use in American politics. These improvements would make the data used more representative of the actual content of news coverage and thus would allow our analysis to be more accurate.