Final Project (PSYC6136: Categorical Data Analysis): Exploring Free Associations

Riya Trikha
215571821
May 1st, 2026

Human beings have uniquely evolved to communicate complex and abstract thoughts through storytelling. Storytelling has served several purposes. For instance, communities have used stories to disseminate important information throughout generations. In Inuit culture, the tale of the Qalupalik, a scary monster that snatches children into the sea, is passed down by Elders through generations, with the goal of keeping them off thin ice during winters.

Interestingly, there seems to be a bidirectional effect: empirical research investigating stories show that the level of engagement for stories we read have a subsequent impact on our thoughts and behaviours. For instance, Green and Brock (2000) introduced the concept of transportation, which measures the extent to which people immerse themselves into a particular narrative world. Through this process, individuals who find themselves transported by a narrative may be more likely to modify their beliefs. Indeed, the researchers found that higher transportation into a narrative led to increased story-consistent beliefs, suggesting that engagement with the story world can affect beliefs formed in the real world. Moreover, Honey and colleagues (2023) argued that mental content related to a narrative persist in our minds without effort or external cues long after engaging it, a concept referred to as "lingering".

In this line of thinking, the current analyses are exploring whether we can capture lingering and investigate how it may differ over several variables. In order to achieve this, we used a free association task, where we ask participants to "type any word that comes to mind for 5 minutes" both before (pre-chain) and after (post-chain) reading a short story called *"Paper Menagerie"* by Ken Liu[1]. Using a small subset of the existing data, my goal with the free association chains is to investigate whether the words in the post-chain are more semantically related to the theme words of the story (theme similarity) than the pre-chain. If so, it would suggest evidence for lingering. Additionally, I investigated whether theme similarity is associated with two additional factors: age and levels of transportation.

The following analyses will include a multiple correspondence analysis (MCA) to examine the relationship between the variables of interest, as well as a principal component analysis (PCA) to visualize theme similarity. The raw data/variables are continuous (age in years, transportation levels were on a 7-point Likert scale and theme similarity measures

[1] The story is also attached in the folder if you are interested in reading.

range from 0 to 1). In order to conduct the categorical analyses, all variables were categorized.

Methods: Data Prepping, Cleaning and Wrangling

First, I defined some helpful functions I use throughout my analyses. A simple function is the “Negate” function, which was useful when subsetting data during cleaning and wrangling.

Negate function:

```
'%ni%' <- Negate('%in%')
```

Two functions were created to aid in calculating theme similarity. The “cos.sim” function allows for the calculation of two matrices (ma and mb) that contain word embedding vectors. Moreover, the “glowca_sim” function calculates theme similarity; the function takes the word embeddings, which were collected using dictionaries from the “semdistflow” package (Reilly & Zukerman, 2026), for the free association chains and calculate the cosine similarity with the word embeddings for the theme words. The function requires the data that contains the embeddings (dat), specifying the column name that contains the free association words (words), and the column range that does not contain the embeddings (othercol). The output contains the theme similarity measures for the MCA and the raw word embeddings for the PCA.

Cosine/Theme similarity functions:

```
cos.sim <- function(ma, mb){
  mat=tcrossprod(ma, mb)
  t1=sqrt(apply(ma, 1, crossprod))
  t2=sqrt(apply(mb, 1, crossprod))
  mat / outer(t1,t2)
}
# Theme similarity function
glowca_sim <- function(dat, words, othercol) {
  #lemmatizes target dataframe on column labeled 'lemma1'
  dat2 <- dat %>% mutate(lemma1 = textstem::lemmatize_words(words))
  #join data to lookup databases specifying common join name explictly
  joindf_glowca <- dplyr::left_join(dat2, glowca, by=c("lemma1"="word"))
  #Select numeric columns
  dat_glo_onlynumeric <- joindf_glowca[-c(othercol)]
  #convert join dataframes to matrices
  mat_glo <- data.matrix(dat_glo_onlynumeric)
  # save embeddings for PCA
  mat_glo_temp <- cbind(dat2, mat_glo)
  embeddings <- rbind(embeddings, mat_glo_temp)
  #compute theme similarity
  vals_glo <- as.data.frame(cos.sim(as.matrix(mat_glo),
as.matrix(theme_glowca_mat_glo)))
  vals_glo <- data.frame(vals_glo)
```

```
  #Rename first column of theme similarity values
  colnames(vals_glo) <- theme$themewords
  # Bind dataframe with free association and theme similarity
  done <- cbind(dat2, vals_glo)
  finalsheets <- list(embeddings, done)
  return(finalsheets)
}
```

The last function was created to calculate the threshold for low, medium and high theme similarity. Low theme similarity will be any value below or equal to the mean of all theme similarity values minus one standard deviation, and high theme similarity will be any value above or equal to the mean of all theme similarity values plus one standard deviation. Any value in between will be considered medium theme similarity. This will be helpful for when the continuous form of theme similarity is categorized.

```
threshold <- function(var) {
  mean <- mean(var)
  standev <-sd(var)
  mean1bsd <- mean - standev
  mean1asd <- mean + standev
  threshlist <- list(mean = mean,
                     sd = standev,
                     mean_sdbelow = mean1bsd,
                     mean_sdabove = mean1asd)
  print(threshlist)
}
```

Next, word embeddings for theme words and free association words were computed using the functions created.

Word embeddings for theme words:

```
# Theme words
theme <- as.data.frame(c("origami", "immigrant", "mother", "tiger", "magic"))
colnames(theme) <- "themewords"
theme <- theme %>% mutate(lemma1 = textstem::lemmatize_words(themewords))
#join data to lookup databases specifying common join name explictly
theme_glowca <- dplyr::left_join(theme, glowca, by=c("lemma1"="word"))
#Select numeric columns dplyr fn operates on tibbles
theme_glowca_onlynumeric <- theme_glowca[-c(1:2)]
#convert join dataframes containing hyperparameter values (glowca and sd15)
to matrices
theme_glowca_mat_glo <- data.matrix(theme_glowca_onlynumeric)
```

Word embeddings for free association words/chains:

```
for (i in 1:length(fa_list)){
  dat <- read.csv(paste0(data.dir, fa_list[i]), header = TRUE)
  dat_long <- melt(dat, id.vars=c("ID", "age", "transportation"))
  colnames(dat_long) <- c("ID", "age", "transportation", "fa.cond", "words")
```

```
  dat_long <- dat_long[dat_long$words %ni% "",]
  dat_long$words <- tolower(dat_long$words)
  # Function - return theme similarity measures and embeddings
  done <- glowca_sim(dat_long, words, c(1:6))[[2]]
  embeddings <- glowca_sim(dat_long, words, c(1:6))[[1]]
  # Bind into metasheet
  glowca_fa_final <- rbind(glowca_fa_final, done)
}
```

To prepare the data for MCA, mean theme similarity was calculated for pre- and post-story free association chain and for each theme word. With this, each participant has five values for theme similarity (one for each theme word).

```
# Convert to long format
glowca_fa_final_long <- melt(glowca_fa_final, id.vars=c("ID", "age",
"transportation", "fa.cond", "words", "lemma1"))
names(glowca_fa_final_long)[c(7,8)] <- c("themewords", "cossim")

# Calculate mean
glowca_final_mean <- glowca_fa_final_long %>% group_by(ID, age,
transportation, fa.cond, themewords) %>% dplyr::summarize(
  mean = mean(cossim, na.rm = TRUE))
names(glowca_final_mean)[6] <- "cossim_mean"
glowca_final_mean$cossim_mean <- as.numeric(glowca_final_mean$cossim_mean)
```

Next, continuous variables were recoded into three categories.

Age was categorized into “Adolescent” (participants 16-years-old and younger), “Young Adult” (participants 17- to 21-years-old) and “Adult” (participants 22-years-old and older).

Transportation was categorized into “Low Transportation” (scores lower than or equal to 2), “Medium Transportation” (scores from 3-5), “High Transportation” (scores higher than or equal to 6).

```
## Categorizing variables
# Age
glowca_final_mean <- glowca_final_mean %>% mutate(agecat = case_when(age <=
16 ~ "Adolescent",
                                                                     age >=
17 & age <= 21 ~ "Young Adult",
                                                                     age >=
22 ~ "Adult"))
# Transportation
glowca_final_mean <- glowca_final_mean %>% mutate(transcat =
case_when(transportation >= 1 & transportation <= 2 ~ "Low Transportation",

transportation >= 3 & transportation <= 5 ~ "Medium Transportation",

transportation >= 6 & transportation <= 7 ~ "High Transportation"))
```

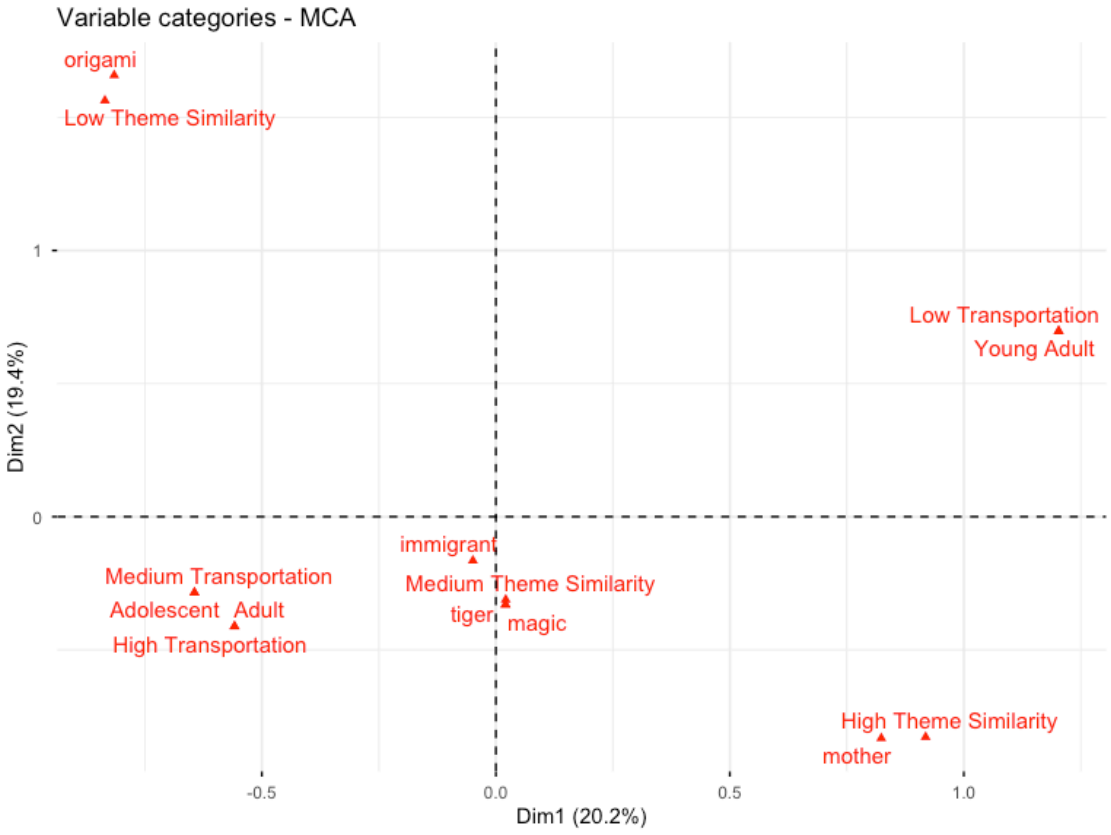
```
# Theme Similarity
th <- threshold(glowca_final_mean$cossim_mean) # Threshold function
glowca_final_mean <- glowca_final_mean %>% mutate(cossimcat =
case_when(cossim_mean < th$mean_sdbelow  ~ "Low Theme Similarity",

cossim_mean > th$mean_sdbelow & cossim_mean < th$mean_sdabove ~ "Medium Theme
Similarity",

cossim_mean > th$mean_sdabove ~ "High Theme Similarity"))
```

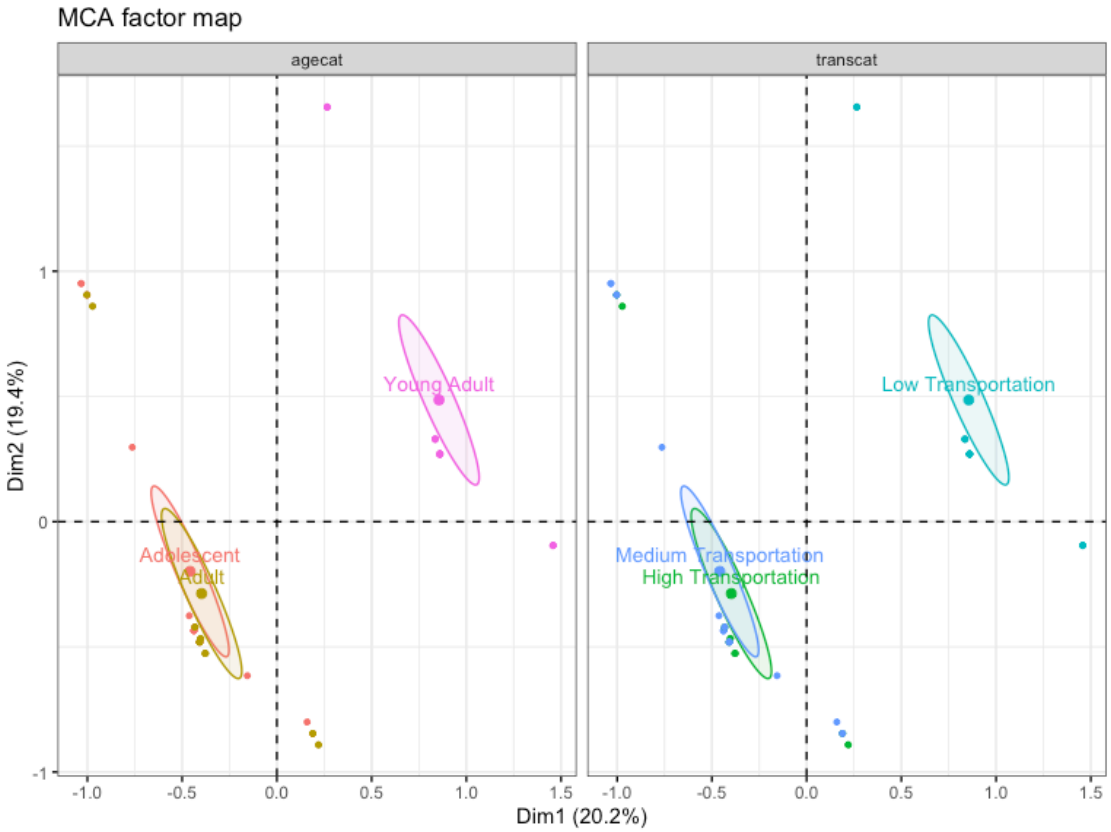
Analyses: Multiple Correspondence Analysis

The multiple correspondence analysis is an extension of the correspondence analysis, as it allows for analyzing and visualizing the relationship/association between more than two categorical variables. MCA was computed separately for the pre- and post-story free association chain, and once for both chains together.

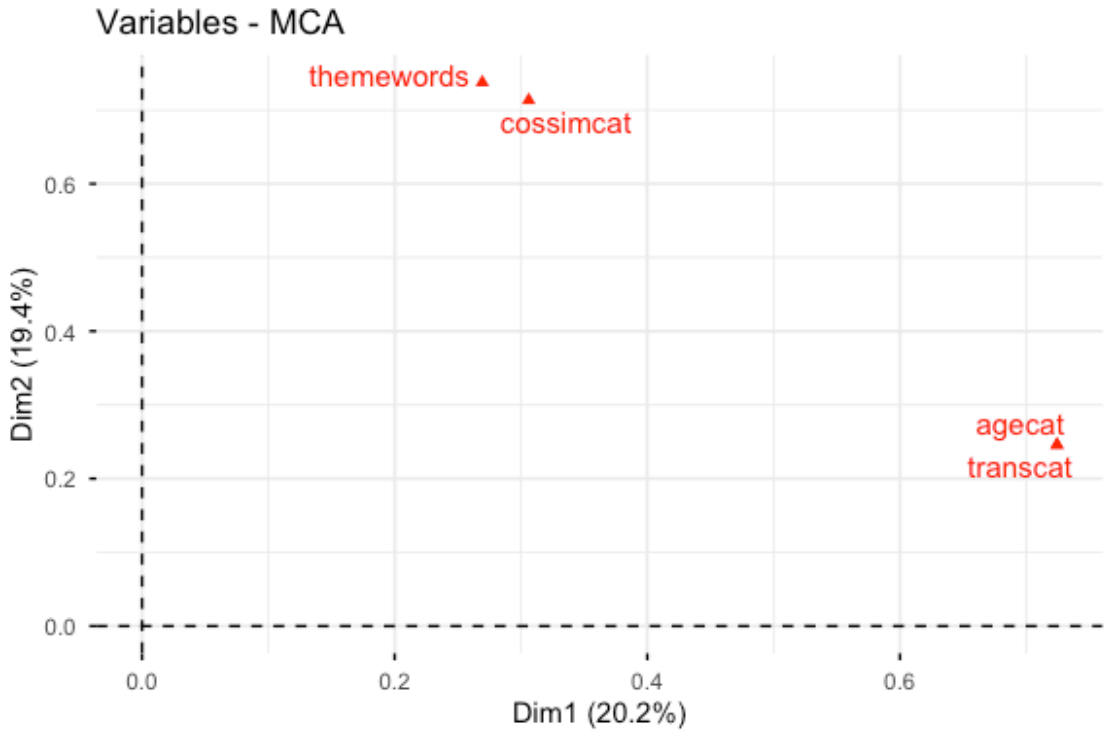
Once MCA is computed, “fviz_mca_var()” and “fvis_ellipses()” from the “factoextra” package (Kassambara & Mundt, 2020) to visualize how each variable of interest lie within the two dimensions individually. The ellipses in the “fvis_ellipses()” allows for a confidence ring around the center/mean point of each level of the variable. These plots will aid in understanding how well age, transportation and theme similarity associate with each other, and whether that differs before and after reading the story.

First, I was interested in visualizing a relationship between all variables.


Variable categories - MCA
origami
Low Theme Similarity
Low Transportation
Young Adult
immigrant
Medium Theme Similarity
tiger
magic
Medium Transportation
Adolescent
Adult
High Transportation
High Theme Similarity
mother
Dim2 (19.4%)
Dim1 (20.2%)
1
0
-0.5
0.0
0.5
1.0

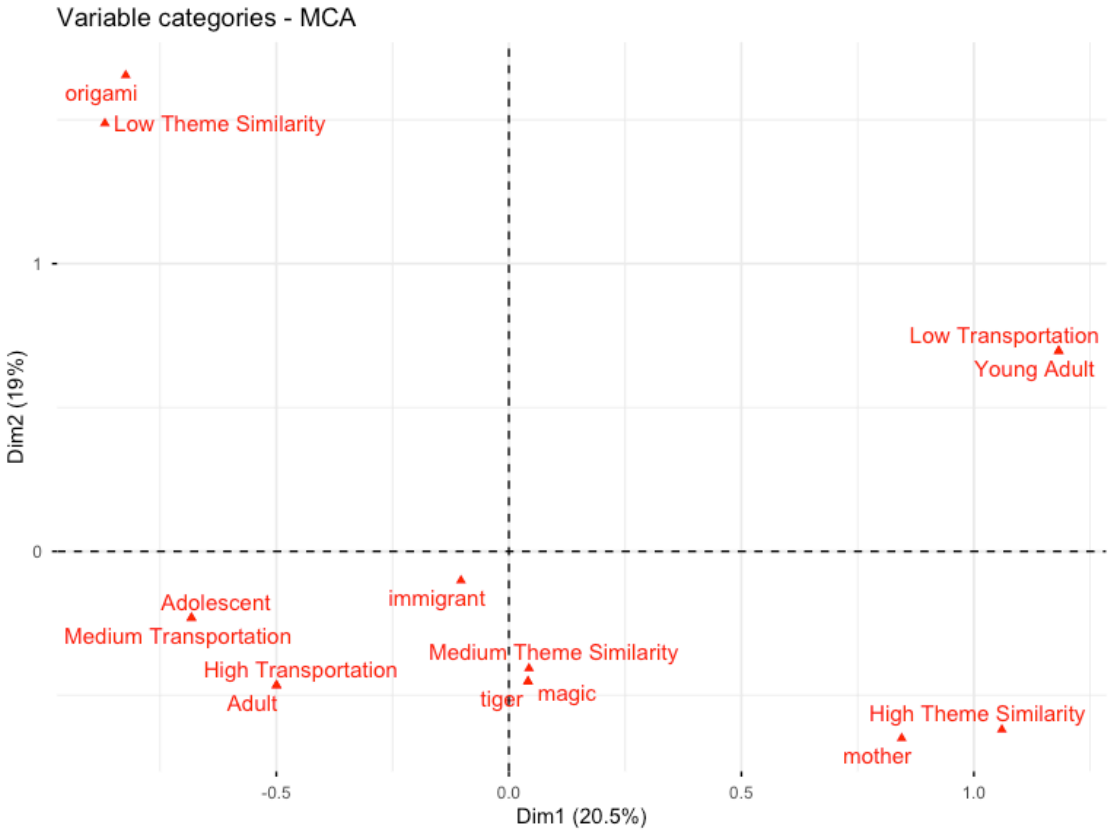

MCA factor map
agecat
transcat
Young Adult
Adolescent
Adult
Low Transportation
Medium Transportation
High Transportation
Dim2 (19.4%)
Dim1 (20.2%)
1
0
-1
-1.0
-0.5
0.0
0.5
1.0
1.5

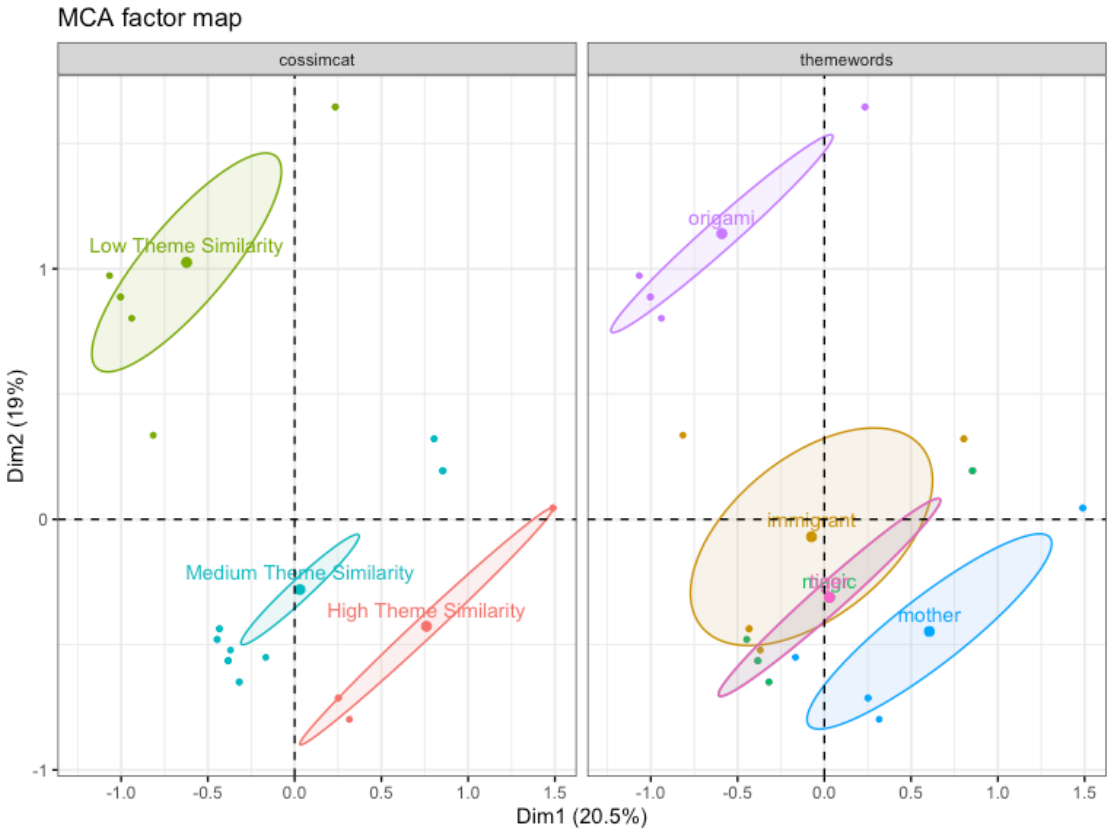
We can see several relationships being shown. First, theme similarity and transportation seems to be reflected by dimension 2 (y-axis), going from lower theme similarity and transportation at the top, to higher theme similarity and transportation at the bottom. Indeed, this is confirmed by the "fvis_mca_var()" plot, which visualizes the correlation between variables and the principal dimensions in the MCA.



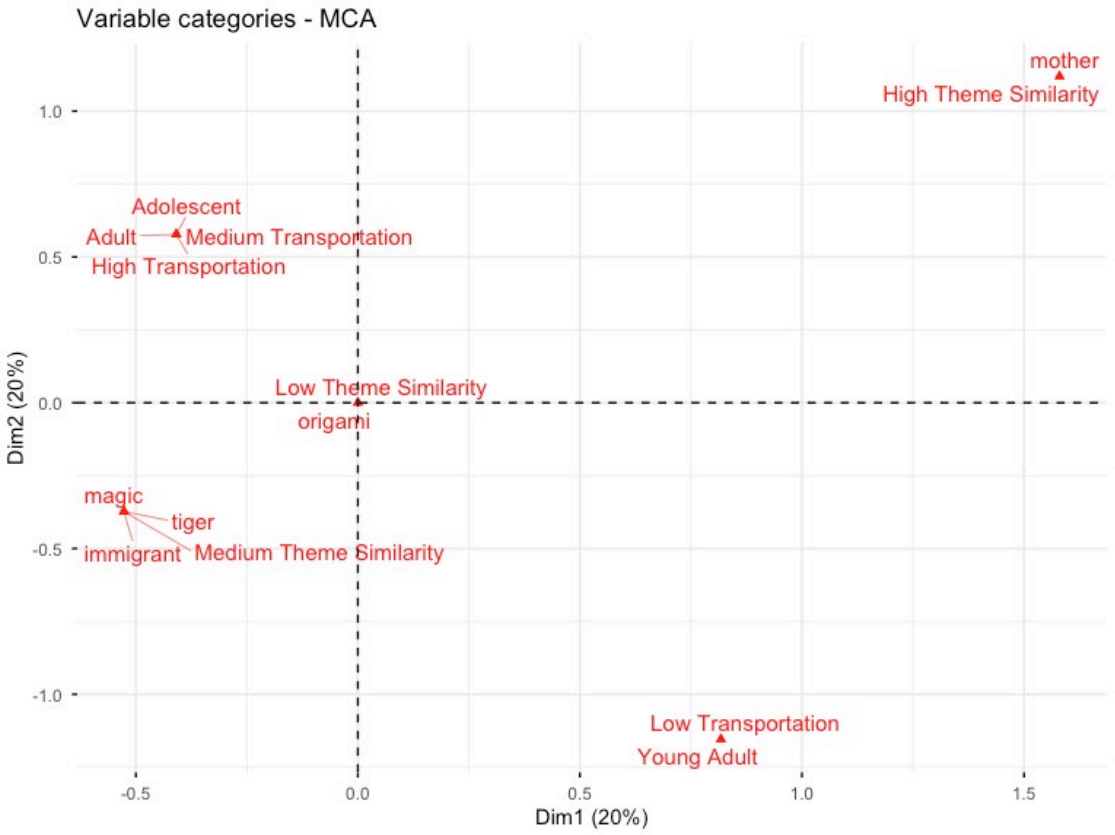
It also seems that young adults reported lower transportation than adolescents and adults, who reported medium and high transportation, respectively. Interestingly, the theme word "origami" was associated with lower theme similarity, whereas "mother" was associated with higher theme similarity. Since this is across both pre- and post-story word chains, the next step is to see if this association is driven by the pre- or post-story condition.

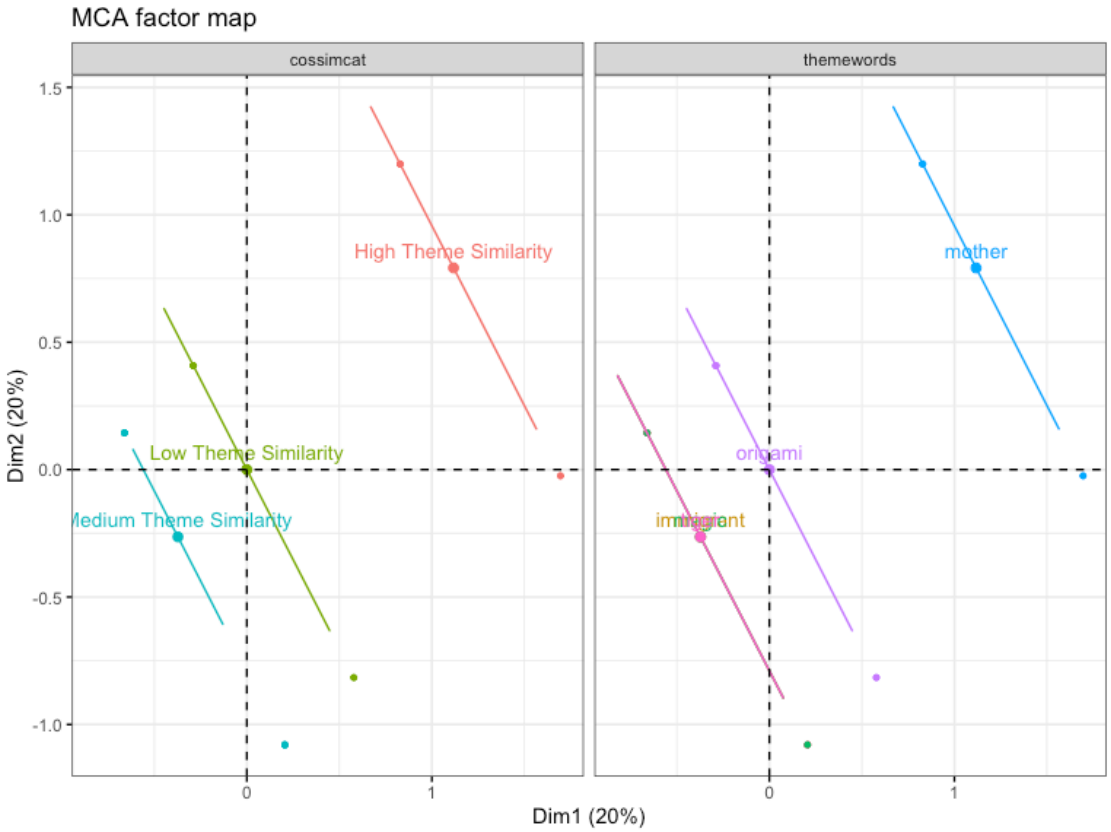
MCA: Pre-story free association chain


Variable categories - MCA
origami
Low Theme Similarity
Low Transportation
Young Adult
Adolescent
immigrant
Medium Transportation
Medium Theme Similarity
High Transportation
tiger
magic
Adult
High Theme Similarity
mother
Dim2 (19%)
Dim1 (20.5%)
1
0
-0.5
0.0
0.5
1.0

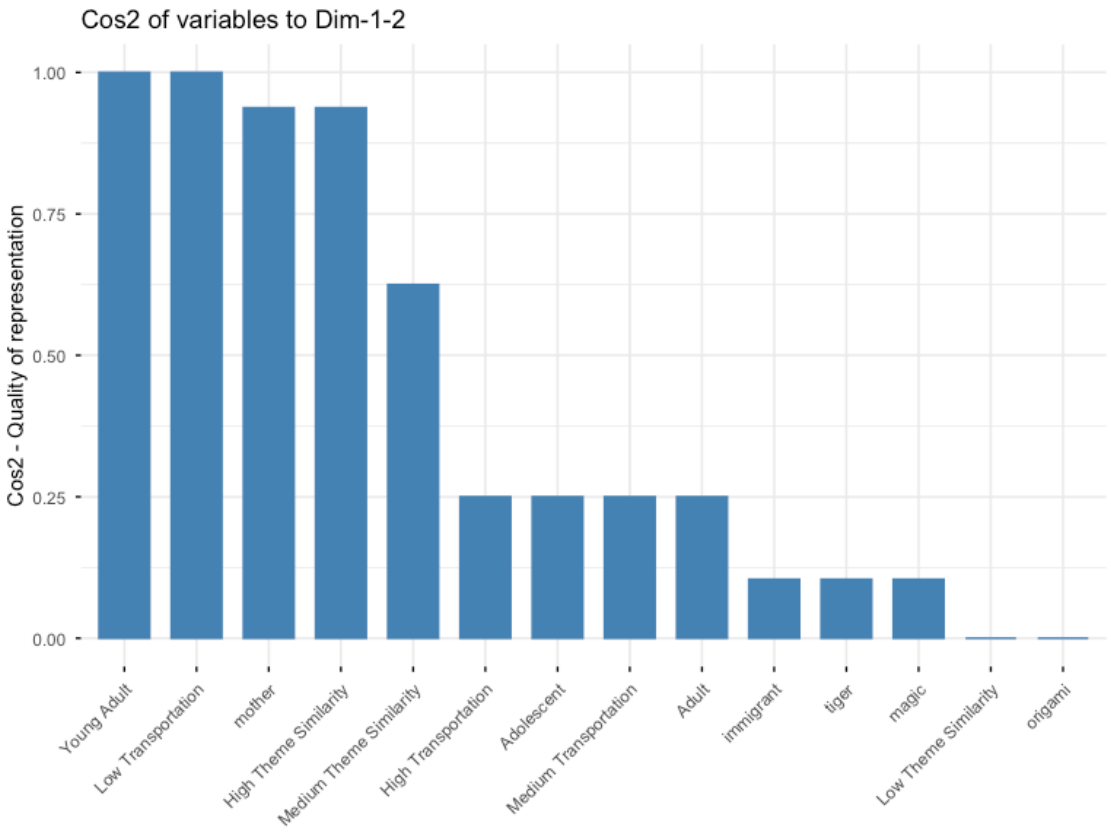

MCA factor map
cossimcat
themewords
Low Theme Similarity
origami
Medium Theme Similarity
High Theme Similarity
immigrant
magic
mother
Dim2 (19%)
Dim1 (20.5%)
1
0
-1
-1.0
-0.5
0.0
0.5
1.0
1.5

MCA: Post-story free association chain





Unexpectedly, the patterns for both pre- and post-story free association chains are similar: “mother” is associated with higher theme similarity and “origami” is associated with lower theme similarity. Importantly, in the post-story chain MCA, the association between “origami” and low theme similarity are directly on the origin, suggesting it may not be well represented on the factor map (at least by the first two dimensions). This is confirmed by plotting “cos2”, or the quality of representation for each variable.



Analyses: Principal Component Analysis

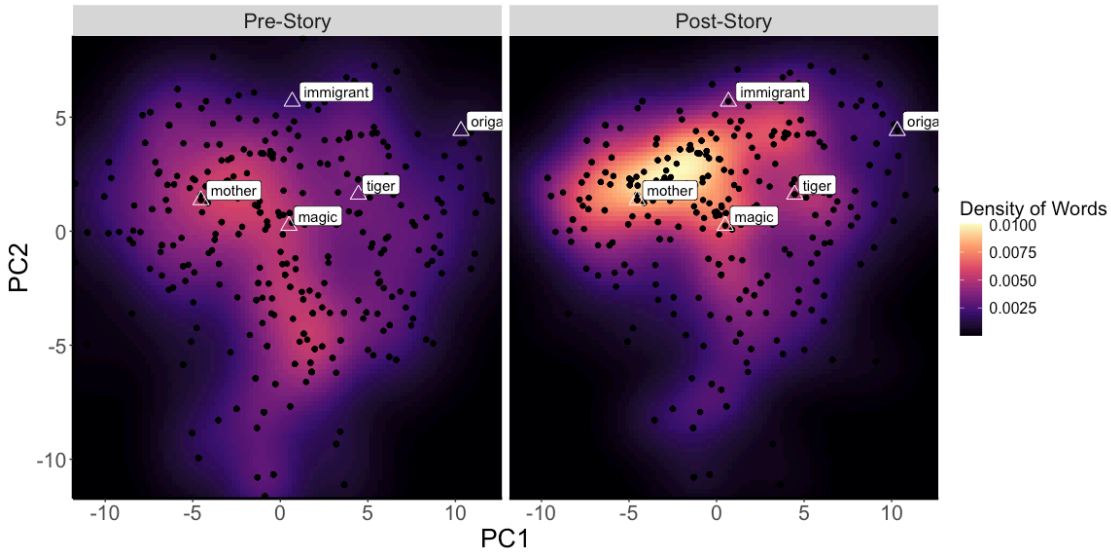
Though the MCA provided insight into general relationships between variables, I was also interested in examining how one's semantic "space" might differ pre- and post-story. That is, we can visualize where each free association word lies in semantic space using the embeddings calculated earlier. Since the word embedding vectors are 300 dimensions long, a PCA was conducted to reduce the number of dimensions, and have the first few principal components represents most of the variability in the data.

```
embed_pca <- embeddings_final[-c(1:6, 307)]
embed_pca <- rbind(embed_pca, theme_glowca_mat_glo)

# Conduct PCA on all embeddings together
res.pca.all <- prcomp(embed_pca, scale = TRUE)

pca_df.all <- as.data.frame(res.pca.all$x)

# Separate PCs for theme words and chain words
pca_df <- pca_df.all[c(1:829),]
pca_df <- cbind(embeddings_final[c(1:6, 307)], pca_df)

pca_df.theme <- pca_df.all[-c(1:829),]
pca_df.theme$themewords <- theme$themewords
```

```
# Recategorize
pca_df <- pca_df %>% mutate(agecat = case_when(age >= 10 & age <= 16 ~
"Adolescent",
                                                                    age >=
17 & age <= 21 ~ "Young Adult",
                                                                    age >=
22 & age <= 29 ~ "Adult"))
pca_df <- pca_df %>% mutate(transcat = case_when(transportation >= 1 &
transportation <= 2 ~ "Low Transportation",

transportation >= 3 & transportation <= 5 ~ "Medium Transportation",

transportation >= 6 & transportation <= 7 ~ "High Transportation"))

# For plotting
pca_df$transcat <- factor(pca_df$transcat, levels=c("Low Transportation",
"Medium Transportation", "High Transportation"))
```

The first density plot shows the distribution of the free associates pre- and post-story, along with the theme words (in a different colour and represented by a different point shape). This plot provides deeper insight into what was found in the MCA. Though we continue to see that "origami" has a lower number of words that are semantically similar to it (represented by the darker purple shading in both pre- and post-story chains), we also see the density of words semantically similar to "mother" increases post-story. This might suggest that family, or specifically mother, is an important theme in the story, and may "linger" more than other themes.

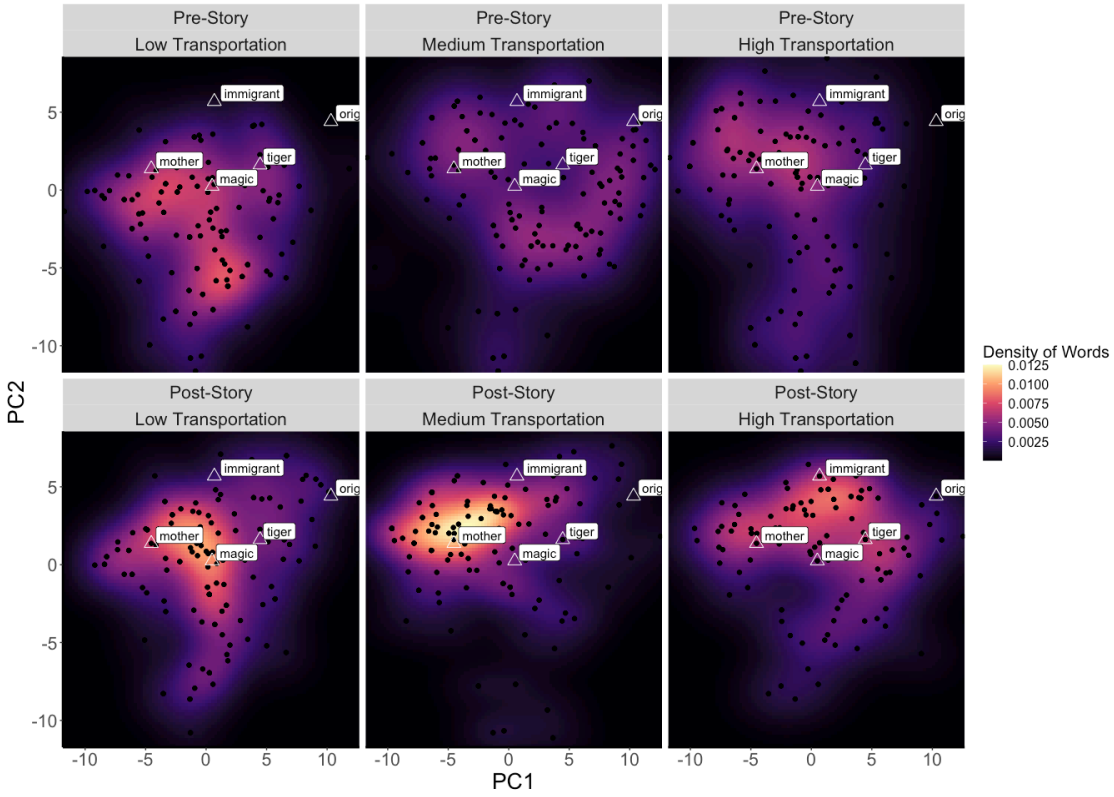
Density plot showing semantic similarity of free associates pre- and post-story.



The second density plot shows the distribution of the free associates as in Figure 1, but now faceted across different levels of transportation. This plot shows that before reading the story, the free associates seem to be very random/unrelated to the theme words

across transportation levels (which is expected). After reading the story, however, we see that for density of words semantically similar to “mother” increases for lower levels of transportation, but density seems to evenly increase over all theme words for high transportation. This not only supports the idea that “mother” is a salient theme, in that it “lingers” for those who reported lower transportation into the story, but also shows that higher transportation results in free associates being more similar to theme words post-story.

Density plot showing semantic similarity of free associates across different levels of transportation.



All in all, there are interesting patterns being observed in this very small subset of data. I hope to continue exploring these trends with the full dataset, and include other factors such as trait curiosity and rumination tendencies.