Exploring Implicit Biases and Bias Propagation within Social and Food-Ideological Groups

Vivian Ngo
Department of Statistics
University of Toronto
Toronto, Ontario, Canada
viv.ngo@mail.utoronto.ca

Abstract

Online ratings have the potential to impact customers' decisions when they decide between which restaurant, hotel, or attraction to visit. In this study, we search for biases pertaining to different types of restaurants and quantify their ability to propagate through social networks. The results indicate that there are statistically significant, albeit small-scale, biases that exists towards different restaurant types. In addition, these biases can propagate through social networks as friends are more likely to visit similar restaurants. Related code for this study is available on GitHub at CSC2552_project.

1 Introduction

Online ratings have the potential to affect people's decisions when they decide what food to eat, which hotel to stay at, and more. These decisions directly impact a business's profit and the livelihoods of business owners. However, studies have shown that online ratings can be subject to biases based on the preferences and perceptions of individual customers (S and Balaji, 2017). For example, for customers who frequently visit high end restaurants, their rating of a restaurant could significantly decrease due to even the slightest reduction in ambience or food quality.

In this study, we will use the 2019 Yelp Dataset to examine rating trends of customers within different social groups and with different restaurant preferences. That is, are people within the same social group more likely to visit similar restaurants, and do their restaurant preferences affect their ratings of restaurants? These questions are interesting because the combination of them would provide insight into users' innate biases and their possible propagation through social networks.

Since they can help to illustrate and quantify the innate biases that individuals have, they also show how these biases can affect a restaurant's overall rating, causing restaurants of equal quality to have unequal ratings. According to previous studies, a one star increase in Yelp rating can increase revenue by 5 to 9 percent so it would be useful for Yelp administrators to be able to understand, quantify and control for biases that disadvantage certain businesses (Hajas et al., 2014).

These questions have not been commonly asked or answered due to the difficulties associated with it, including the challenge of differentiating between customers and between restaurants in a simple manner. Moreover, because websites that host reviews are usually not considered social media websites and lack the abundance of user interaction, they are commonly not used to extrapolate information about social networks, as compared to sites such as Facebook and Twitter. In this paper, we will use information from the Yelp dataset as well as a technique inspired by semantic density analysis to extract this information. This paper will also introduce a novel method of measurement which will be termed the Food Ideology score, the principle of which can be applied to many different ideology measurements.

2 Related Works

Many projects have been conducted to study restaurant review data, prediction, and analysis. The methodology used in this study was also inspired by several papers written about other applications.

One related paper is *Analysis of Yelp Reviews* which analyzed businesses in and around college towns to study the cyclical patterns of data as well

as biases that can occur in online review data (Hajas et al., 2014). The results of the paper suggest that the cumulative rating of restaurants as a whole converge, whereas the reviews of individual restaurants tend to fluctuate. While the paper focused heavily on the spatial and temporal aspects of the ratings of restaurants, it did not determine the causes for these differences. For example, the paper suggests that the fluctuating ratings of individual businesses could be caused by food preferences but no analysis was done to test this hypothesis. In this study, we will attempt to extend the analysis of Hajas et al. (2014) and answer this question. Whereas *Analysis of Yelp Reviews* focuses on patterns amongst review ratings, this study will focus on determining factors affecting the ratings and the results of both studies should be complementary.

Another paper that is useful for this study is *Exposure to ideologically diverse news and opinion on Facebook* which studies the exposure to cross-ideological content on Facebook by comparing the components of friendship networks and interaction with cross-ideological content (Bakshy et al., 2015). In this paper, the authors create a measure of content alignment of Facebook content by averaging the self-reported political affiliations of Facebook users that are exposed to that content. Analysis is then done on content of varying alignment scores.

The measures used in this paper were intuitive, the sample size was large, and the study benefitted from the data being nonreactive. However, there is an issue of a cyclical analysis since the goal is to measure users' exposures to cross-content but the alignment scores are calculated using users' ideologies to begin with. In this study, we utilize the idea of an ideology score, but apply it specifically on food preferences. This food ideology score will then be employed in models that predict ratings that customers give restaurants to determine if food preference is a strong predictor. Although the ideology score will be built using restaurant information, it will not depend on restaurant *ratings*, so there will be no issues of a cyclical analysis. The importance of an ideology score in this study is that it provides a way of differentiating customers analogously to restaurants in order to detect biases when holding other variables constant.

3 Data

The data used in this study is provided by Yelp as part of the 2019 Yelp Open Dataset that is publicly shared for personal and educational purposes. The provided datasets include information about businesses, reviews, users, checkins, tips, and photos. For the purposes of this project, we will utilize the business, reviews, and users data. In addition, we will focus on the subset of restaurants in Ontario. Ontario was chosen because of its large quantity of restaurants as well as its diverse range of restaurant types.

Whereas the original dataset contains close to 6 million reviews of businesses in several countries, the data used for this analysis contains 399,072 reviews, 10,485 restaurants, and 42,096 user accounts, spanning from June 2008 to November 2018.

Some key variables of interest are restaurant categories, restaurant ratings and review star ratings. Restaurant categories indicate the type of cuisine that a restaurant serves, such as bar food, pub food, pizza, or ice cream while restaurant ratings are the star ratings that are given to restaurants as an aggregate rating based on all of the user ratings. Information about friendships is also extracted from the user dataset, where every user has a list of associated friends. More details about the usage of this data will be described below.

4 Methods

The first research question is whether users from the same social group tend to visit more similar restaurants and the method used to answer this question is introduced in *Semantic density analysis: Comparing word meaning across time and phonetic space* (Sagi et al., 2009). This paper discusses a method to analyze word meanings using the density of semantic vector clusters. For every word in the corpus, the researchers find a cluster of context vectors associated with the word and the density of these vectors is then measured as the average cosine similarity between pairs of context vectors. The interpretation of a dense cluster is that the contexts of the word are similar, whereas a less dense cluster could mean that the word has multiple meanings and is used in a larger variety of contexts.

In this study, we analyse the semantic density of word vectors for users within the same friend group to determine if people in the same friend

group tend to visit more similar restaurants than random groups of people. In particular, every user is assigned a word vector which reflects the types of restaurants that they have visited and the vectors within a social group are then compared to the vectors within a random group of people. This is repeated 1000 times and the final differences of the distributions are quantified through statistical tests. The specific steps taken are described below.

First, a restaurant-category matrix is constructed so that every restaurant has a vector associated with it which contains all of the information about its categories. Entry (r, c) of this matrix will be 1 if category c appears in restaurant r, and 0 otherwise. This is analogous to a word-context matrix in vector semantics. This matrix is also adjusted using term frequency–inverse document frequency (TF-IDF) so that extremely common categories, such as "food" and "restaurant", are given less weight. For every user, their personal word vector is defined as the sum of the vectors in the restaurant-category matrix that correspond to the restaurants that they have reviewed.

Next, we utilize Yelp's friend data. Every user in the Yelp dataset has a list of other users which are their "friends". This list of friends is used as a proxy for the user's social network. In order to see if people from the same social group frequent more similar restaurants, 1000 random users were sampled and 20 friends were sampled randomly again for each of these users. For every pair of friends in the sample of 20 friends, we calculated the cosine similarity between their two personal word vectors. The average cosine similarity of these 20 word vectors were then recorded for every user. An important note is that every friend of a user also needs to have an associated word vector in order for a cosine similarity score to be calculated. However, in the dataset, not every friend is listed as a user. To overcome this issue, friends of users that were not listed as users themselves, were removed from the friends list. This removal did not greatly decrease the sample size and a total of 20 friends was still able to be collected from each of those 1000 users.

For comparison, another 1000 samples with 20 random users each were also collected. Similarly to the method above, the average cosine similarity of every sample was recorded. The 1000 average cosine similarity scores of these groups of random people are then compared to the average cosine similarity scores of the 1000 friend groups.

A two sample t-test of equal means with no assumption of equal variance was first used to compare the average cosine similarities of the friend groups to the average cosine similarities of randomly simulated Yelp users. In addition, the Kolmogorov Smirnov (KS) test of equal distributions was used to determine if the underlying distributions of the two groups were the same, statistically. A higher cosine similarity within the friend groups would indicate that they visit more similar restaurants, relative to groups of random people.

The next research objective is to determine what factors of a restaurant experience play a larger role in determining the amount of stars that different people give. For this analysis, an ideology weighting similar to the measurement mentioned in Bakshy et al. (2015) is constructed. Using k-means clustering, restaurants are classified into two different groups, Category 0 and Category 1, using their associated vector in the restaurant-category matrix. Next, every user is assigned a weighted ideology score:

$$S_i = \frac{(R_{1i} + 1) * \frac{1}{N_1}}{(R_{1i} + 1) * \frac{1}{N_1} + R_{0i} * \frac{1}{N_0}} * 100 \qquad (1)$$

where:
S_i = Ideology Score for User i
R_{1i} = The number of restaurants that User i went to that are in Category 1
R_{0i} = The number of restaurants that User i went to that are in Category 0
N_1 = The number of restaurants in Category 1
N_0 = The number of restaurants in Category 0

This ideology score is between 0 and 100 and can be loosely interpreted as the preference of an individual to eat at and review a restaurant of Category 1.

The next step is to use the ideology score as a predictor and conduct regression analysis to determine what factors strongly affect the star rating that a user gives a restaurant. Models with different combinations of predictors are explored and compared using the Akaike information criterion (AIC), Bayesian information criterion (BIC), and likelihood ratio test (LRT) to determine the best model that provides a good balance of predictive power and simplicity. This model will thus

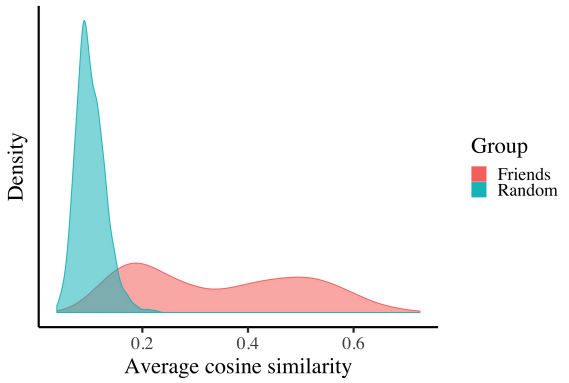

Figure 1: Density of Average Cosine Similarity for Friends vs Random Users. As shown, the average cosine similarities of random users are concentrated in the lower range with a mode of around 0.10. On the other hand, the average cosine similarities of friends are much more spaced out and reach larger ranges almost triple of those of random users.

provide information about what factors are important in predicting the star rating given by a customer. Further, the estimated coefficient and p-value of the ideology score predictor are indicative of whether people of different food preferences/ideologies tend to score restaurants differently.

Finally, because restaurants are very diverse and can be represented using more than two clusters, an additional analysis is done on a higher number of restaurant clusters. The number of appropriate clusters is approximated using the elbow method and regression methods will be used again to determine if different clusters of restaurants are rated differently on average.

5 Results

According to the results of the semantic density analysis, the average cosine similarity amongst friends averaged around 0.35 whereas the average cosine similarity amongst random Yelp users averaged around 0.10.

As can be seen in Figure 1, the average cosine similarity of people who share a common friend is more than triple that of random Yelp users. A two sample t-test used to test for equal means also produced a p-value of $2.73 * 10^{-265}$ (test statistic = 47.61, degree of freedom = 1058.03). This provides strong evidence that the mean of the average cosine similarities within groups of

Cluster 0	Cluster 1
chinese	pizza
bar	italian
fast	wings
burger	chicken
sandwiches	services

Table 1: Keywords of Restaurant Groups

friends and random Yelp users are not the same. A Kolmogorov-Smirnov test for equal distribution also provided strong evidence against the hypothesis of equal distributions (p-value $< 2.2 * 10^{-16}$). Recall that the average cosine similarity is used as a proxy for the similarity of restaurants that users visit. Thus, these tests provide evidence that users within the same friend group visit more similar restaurants than random users.

Using k-means clustering, the restaurants were then clustered into two main groups. The groups are associated with the key categories shown in Table 1. Every user was then given an ideology score using these classifications (Equation 1).

In addition to the restaurant classifications, several other predictors were used to explore and determine the best model choice. In particular, the predictors that were experimented with are the average stars given by a user, the rating of a restaurant, and a user's ideology score. User ID was also used as a random effect.

In Figure 2 are the p-values associated with the predictors as well as the model diagnostics associated with every model, including the AIC, BIC, and log likelihood. In every model, the average stars from the user and the restaurant's rating are useful predictors for the given review stars. However, the usefulness of the other predictors are not as clear. To select the best model, we compare the model diagnostics.

The AIC and BIC are measures that quantify the balance of a model's fit to the data and the model's complexity. Smaller values of these measures indicate a more preferred model. On the other hand, the log-likelihood quantifies the model's fit to the existing data and is preferred to be high.

Judging by the model diagnostics, the two best models are models 6 and 7 which include all the predictors mentioned above as well as User ID as a random effect. The two models only differ in terms of whether they include the interaction between user's ideology score and restaurant clas-

Model	Predictor P-values						Model Diagnostics		
	Average stars from the user (a)	Restaurant rating (b)	User's ideology score (c)	Classification of restaurant (d)	Interaction Term (c x d)	User ID (e)	AIC	BIC	Log Likelihood
1	0	0	-	-	-	-	1096729	1096772	-548360.5
2	0	0	0.9491	-	-	-	1096748	1096802	-548368.9
3	0	0	-	0.5828	-	-	1096739	1096793	-548364.4
4	0	0	0.9647	0.5841	-	-	1096758	1096823	-548372.8
5	0	0	0.5661	0.0179	0.0209	-	1096768	1096844	-548377.2
6	0	0	0.3712	0.494	-	Random	1096464	1096540	-548224.9
7	0	0	0.8321	0.0093	0.0121	Random	1096474	1096560	-548228.8

Figure 2: P-values for predictors and model diagnostics for 7 models used to predict the stars of a given review. Model diagnostics are color-coded from red to green where green is the most favorable value and red is the least. Predictor p-values are color coded where green indicates greater statistical significance of a predictor in a model and red indicates less statistical significance.

sification. Even though the model diagnostics of model 6 are more preferable over those of model 7, the difference is very small and does not strongly justify picking model 6 over model 7. Thus, a conceptual approach can be used to choose between them. Because the research question of this study revolves around implicit biases, the interaction between a restaurant type and a user's preference is a key predictor to include in the model, regardless of whether it is (statistically) useful or not. Therefore, model 7 is chosen over model 6.

The final regression model is thus:

$$\begin{aligned}\text{Review.Stars}_{ij} = &-2.0120874 \\ &+ 0.7885 * \text{ARU}_i \\ &+ 0.7754 * \text{RR}_j \\ &+ 0.0000 * \text{S}_i \\ &+ 0.0601 * \text{RC}_j \\ &- 0.0009 * (\text{S}_i * \text{RC}_j) \\ &+ \beta_i * \text{User.random.effect}_i \\ &+ \epsilon \end{aligned} \quad (2)$$

where:
ARU_i = Average rating from User i
RR_j = Rating for Restaurant j
S_i = Ideology Score for User i
RC_j = Classification of Restaurant j (0 or 1)
β_i = Coefficient for User i's random effect
ϵ is random Gaussian noise

As we can see, a user's average rating of restaurants and a restaurant's rating are useful predictors when predicting the review stars. More specifically, they are positively associated with review stars. This indicates that every one unit increase in a user's average rating corresponds to an average increase of 0.79 in their review stars. A higher restaurant rating is also associated with a higher review rating, which is not surprising since a restaurant with a higher rating probably earned the rating by providing better quality of food and service, and continues to do so, leaving customers more satisfied. On the contrary, the effects of a user's ideology score and a restaurant's classification are more complicated.

If a restaurant is classified into Group 0, then the interaction term has no effect on the review stars, and thus the user ideology has no effect on the review stars. In fact, in addition to having a coefficient of zero, ideology score is not a statistically significant predictor in the model. However, if a restaurant is classified into Group 1, then the review stars is expected to decrease by 0.0009 on average, for every increase one point increase in a user's ideology score ($p - value = 0.0121$). This could be a result of users developing higher standards for cuisines that they visit often. However, because this effect only applies to restaurants of Category 1, it is difficult to determine the cause of this phenomenon without more in-depth causality analysis or experimentation. Also, one should

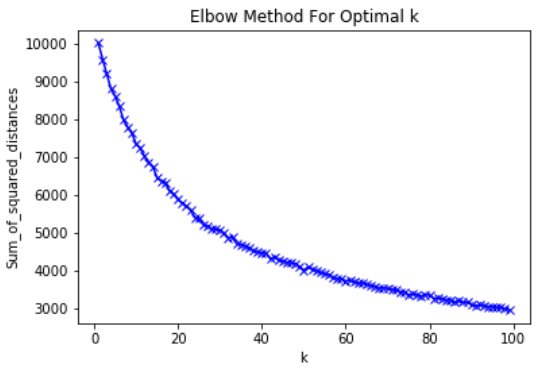

Figure 3: Sum of squared distances (SSD) by number of clusters in the clustering of restaurants. SSD decreases quickly from 0 to 20 clusters but begins to decrease at a slower pace from roughly 20 clusters.

note that the coefficient of -0.0009 is incredibly small considering that review ratings range from 1 to 5. Therefore, this effect could be statistically detectable but not practically significant.

Because the impact of ideology score is small but a restaurant's category has a more statistically significant effect on a review's star rating, we then focused on differences between restaurant classifications. According to Figure 3, the sum of squared distances using the k-means method starts to decrease at a slower rate at roughly 20 clusters.

Running a regression model with average stars from a user, a restaurant's rating, restaurant classification, and User ID, we get that all these predictors are statistically significant. The coefficients for the restaurant classifications are listed in Table 2. Holding other variables constant, different groups of restaurants are rated significantly differently. However, once again, the scale of these differences is small. This can be seen in Figure 4 as well.

Figure 4 shows a boxplot of the review stars of restaurants from varying restaurant groups. Even though the boxplots have been ordered from lowest to highest regression coefficient (indicating negative to positive effect on review stars), there is not a clear trend in the distribution of the review stars. Therefore, even though different restaurant groups have statistically different review stars on average, these differences are very small and may not be significant in practice.

Group Keywords	Coefficient
breakfast, brunch	-0.046
mexican, fast	-0.041
tea, coffee	-0.030
bars, nightlife	-0.029
italian, pizza	-0.018
burgers, fast	-0.009
asian, fusion	-0.006
korean, seafood	0
thai, vietnamese	0.002
sandwiches, fast	0.003
japanese, sushi	0.009
event, planning	0.011
eastern, middle	0.015
indian, vegetarian	0.020
chicken, wings	0.020
greek, meditteranean	0.021
chinese, ethnic	0.021
pizza, fast	0.025
pakistani, indian	0.055
caribbean, chinese	0.059

Table 2: Key Categories of 20 Restaurant Clusters

6 Discussion

Overall, the results show that groups of friends on Yelp tend to visit more similar types of restaurants than groups of random users. However, biases due to preference of different restaurant types are small in effect.

The regression analysis showed that ideology score impacts a user's review stars depending on restaurant category but this effect is in the order of tens of thousandths. A deeper dive into more restaurant clusters also shows that review stars are statistically significantly affected by restaurant types, but the effects of these are also in the order of tenths or hundredths, whereas review stars range from 1 to 5. Therefore, there does appear to be statistically detectable implicit bias due to different restaurant types, albeit this bias is small. Additionally, it is possible for this bias to propagate through social groups because users in the same friend group tend to visit similar restaurants. In a broader context, this shows that biases can exist and propagate through social groups on online review platforms and be statistically detectable but could appear at such a small scale that it is not considered to be significant in practice.

More analysis can be done to further explore

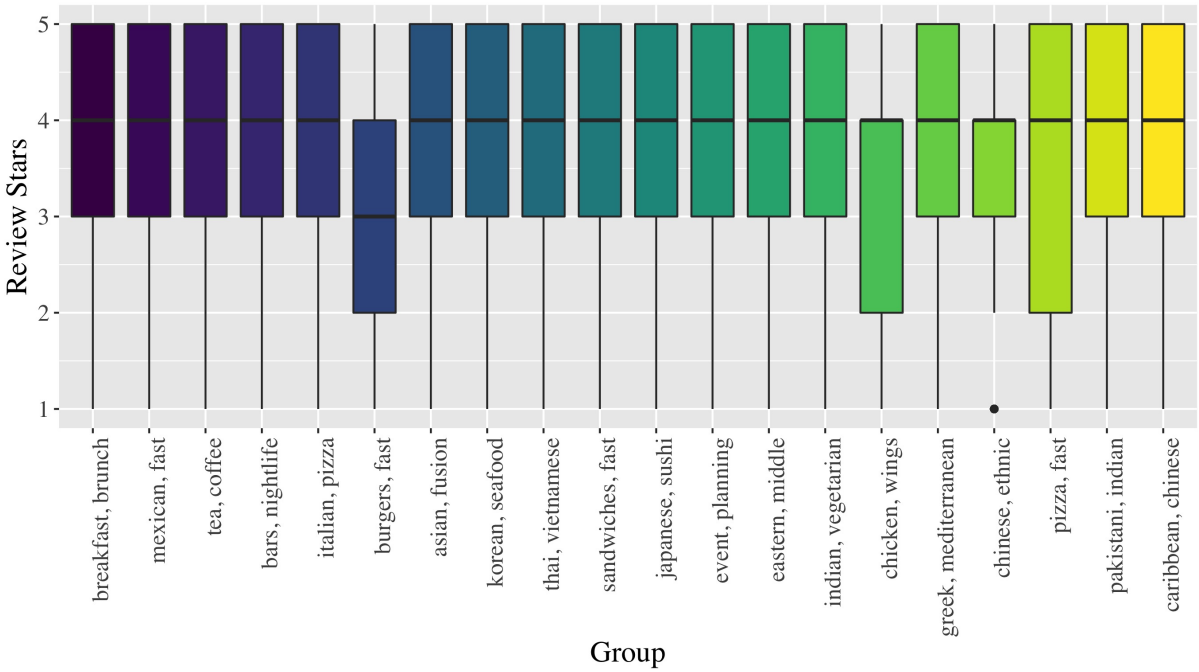

Figure 4: Boxplot of Review Stars by Restaurant Group/Cluster, ordered from lowest regression coefficient to highest. There is no clear increasing trend from left to right, as one would expect from the regression coefficients. However, there are clear differences in the distributions of review stars, by group.

bias propagation through social networks and it is important to note the potential limitations of this study. To begin, one limitation is that restaurant reviews were used as a proxy for restaurant visits, and thus restaurant preference. However, this can be inaccurate since users are unlikely to review a restaurant multiple times even if they enjoy it and visit the restaurant often. In fact, some users might never leave a review for their favorite restaurants. This assumption could affect the calculation and interpretation of the ideology score. Future studies should search for a better measurement of restaurant preference or visitation count.

Moreover, this study used food/cuisine categories to differentiate between restaurants but this may not be the best differentiating factor to use. For example, restaurants that serve coffee are likely to be grouped together under this classification method but they can vary greatly in price and target customers. Future studies should utilize additional distinguishing factors such as price range, ambience, location, and business hours to provide more distinctions between restaurants and create more defined groupings.

Future studies should also include the following steps to further the analysis in this study. To begin, it is important to note that there are many factors that can cause restaurants to have different ratings, such as seasonal changes or change in management. To mitigate the effects of confounding variables and to pinpoint sources of bias, an experimental study would be a stronger option than an observational study.

It is important to note that restaurant food quality, ambience, and accommodations are difficult to control for in an experiment. Therefore, it may be appropriate to conduct more general studies on biases around attributes related to a restaurant, such as price and location, rather than specific restaurants themselves. These studies should not require in-person restaurant visits and would directly test for biases that people have before visiting a restaurant, which is the question of interest. The results of these studies can also provide a baseline for more complicated studies involving specific restaurants and in-person experiences.

In a broader context, the results of this study demonstrate that online platforms, such as Yelp, provide opportunities for users to share opinions and influence one another. This sharing can further lead to the spread of preferences and even biases, causing certain establishments to be treated unfairly. Although the biases detected in this study were small, there could be other platforms in

which larger biases exist or the population of users is large enough for a small effect to be impactful. The methods from this study can be extended and applied to other applications to detect biases and how likely they are to spread.

References

Eytan Bakshy, Solomon Messing, and Lada Adamic. 2015. Political science. exposure to ideologically diverse news and opinion on facebook. *Science (New York, N.Y.)*, 348.

Peter Hajas, Louis Gutierrez, and Mukkai S. Krishnamoorthy. 2014. Analysis of yelp reviews.

Dhanasekar S and Balaji. 2017. Rating the online review rating system using yelp.

Eyal Sagi, Stefan Kaufmann, and Brady Clark. 2009. Semantic density analysis: Comparing word meaning across time and phonetic space. *Proceedings of the EACL 2009 Workshop on GEMS: Geometrical Models of Natural Language Semantics*, pages 104–111.