BIRLA INSTITUTE OF TECHNOLOGY AND SCIENCE PILANI

K.K. BIRLA GOA CAMPUS

UNDERGRADUATE THESIS

Why are people essentialists?

A Computational Investigation

Author:

Samarth MEHROTRA

Supervisor:

A/Prof Amy PERFORS

A thesis submitted in partial fulfillment of the requirements for the degree of B.E. (Hons.) Computer Science

November 30, 2018

Certificate

This is to certify that the thesis entitled, *Why are People Essentialists? A Computational Investigation*, submitted by Samarth Mehrotra (ID No. 2015A7TS0062G) in partial fulfillment of the requirements of BITS F421T, embodies the work done by him under my supervision.

_____

Supervisor

A/Prof Amy PERFORS

Date:

Abstract

Samarth MEHROTRA (2015A7TS0062G)

Why are people essentialists?
A Computational Investigation

Essentialism is the view that certain categories have an underlying reality or true nature that cannot be observed directly. A number of developmental studies in psychology have shown that people, especially young children, are essentialists. However, the question of why people are essentialists has remained unanswered. In this thesis, we try to answer this question by exploring two possibilities using a computational approach. First, we try to answer the question whether perceptual features for natural kinds support the inference of a hidden causal variable. We compare the likelihood of various causal structures to answer this question. Second, we study the role that language plays in the development of essentialist beliefs in children. We implement a classifier for the automated identification of generic noun phrases and use this to study how the frequency of generic noun phrases varies across lexical categories. Our analysis indicates that the environment offers perceptual and linguistic input for people to develop essentialist beliefs. However, the input is available for both natural and artefact categories and fails to explain why people develop essentialist beliefs only for natural kinds.

Keywords: Psychological Essentialism, Bayesian Networks, Generic Noun Phrases

Acknowledgements

First and foremost, I would like to thank my supervisor, Dr. Amy Perfors. From our first few interactions via email to the last day of my thesis, she has provided me with the best possible guidance that a student could ask for. She has given me the space to work independently but at the same time ensured that I was never getting lost and always moving in the correct direction. Most importantly, she has helped me understand what cognitive science is all about and how computational modelling fits into cognitive science research. The last five months have been a great learning opportunity and this would not have been possible without her support and direction.

I owe special thanks to Dr. Charles Kemp, who is the examiner of this thesis. His advice and constructive comments have been very useful in shaping this thesis. I have been able to learn a great amount by merely observing his way of thinking and questioning during the weekly lab meetings. Observing his depth and breadth of knowledge, has inspired me to expand my technical knowledge from computer science to a number of other disciplines.

I would like to thank my on-campus supervisor, Dr. Neena Goveas, for all the help in ensuring that I met the deadlines and requirements for the thesis. This thesis would not have been possible without her support.

Finally, I would like to thank everyone who is associated with the Computational Cognitive Science Lab and the Memory and Language Lab at the University of Melbourne, for making it a very easy and cooperative environment to work in. Through lab meetings and lunch discussions, I have been

exposed to a number of research areas in cognitive science and have been able to understand where computer science fits in.

Contents

List of Figures

List of Tables

List of Abbreviations

CNN	**C**onvolutional **N**eural **N**etwork
FCI	**F**ast **C**ausal **I**nference
RFCI	**R**eally **F**ast **C**ausal **I**nference
BN	**B**ayesian **N**etwork
BF	**B**ayes **F**actor
BIC	**B**ayesian **I**nformation **C**riterion
DAG	**D**irected **A**cyclic **G**raph

Chapter 1

Introduction

1.1 What is Essentialism?

According to Gelman (2003), essentialism is the idea that certain natural and social categories have an underlying reality which might not be observed directly. An essence is a hidden, unobservable property of a category which is responsible for various surface features. For example, the essence of a lion causes it to have a mane, four legs, an ability to roar, etc. Biologically, the essence could be attributed to the DNA of a living organism. In chemistry, properties of water (odourless, colourless, etc.) can be attributed to it's chemical composition i.e. H_2O. Medin and Ortony (1989) further claim that essentialism is a placeholder concept. Adults do not necessarily know what the essence is, however, they believe that certain categories have an underlying reality.

In this thesis we try to answer the question: why are people essentialists? We study whether perceptual features of essential categories support the inference of hidden causes. We also study the role that language plays in the development of essentialist beliefs.

We begin by presenting a summary of studies which suggest that people are essentialists.

1.2 Evidence of essentialist beliefs in children

The existence of these essences in the real world is a philosophical question, often termed as metaphysical essentialism. However, the question of whether these essences are present in people's representations of categories is a psychological question called psychological essentialism. Medin and Ortony (1989) suggest that psychological essentialism is the psychologically plausible analog of the implausible theory of metaphysical essentialism. Gelman (2003) refers to this idea that people's representation of objects reflect the belief of an essence as representational essentialism.

Given that essentialism is a placeholder concept and people might not know what the internal essence exactly is, direct evidence that people are essentialists is difficult to obtain (Gelman, 2003). Gelman (2003) argues that the existence of essences would imply that people's categories are structured and have an inductive potential, extend beyond surface features, include non-obvious properties and hidden causal features which are stable over physical transformations. In this thesis, we try to answer why representations of essentialised categories consists of hidden causal features and allow for category based inferences. For now, we present evidence which shows that people's categories and conceptual representations show all the properties listed above.

For instance, Gelman and Wellman (1991) conducted a series of experiments to test children's understanding of features beyond the surface level. Children by three years of age were able to distinguish between internal (eg. bones, blood) and surface (eg. skin) features. Although children's responses were often inaccurate or uninformative (lemons have lemony stuff inside, stuffed dogs have blood inside but no bones), they were able differentiate

between inside and outside. Children of age four considered internal and intrinsic features to be important for an object's identity and functioning (Gelman and Wellman, 1991). Further, children showed the belief that members of a category share certain intrinsic properties which are unaffected by the environment in which the object is brought up in. They anticipated that certain physical properties will develop irrespective of the environment.

In addition to the evidence showing that children infer an underlying structure with various categories, there is evidence that children's categories have a strong inductive potential. For instance, Gelman and Markman (1986) conducted an experiment to test children's ability to make inferences. Children were shown an image of a dolphin and told that the dolphin pops out of water to breathe. Children were able to infer that a second dolphin would also pop out to breathe, indicating that children can make simple inferences. The experiment further tested on what basis were children making these inferences. Children were shown an image of a tropical fish, told that it was a fish and that it breathes underwater. They were then shown an image of a dolphin, told that it was a dolphin and that it pops out of water to breathe. Children had to decide how a third fish (they were told it was a fish) which looked like a dolphin, breathes. Children of age four based their inferences on common category membership even when there was conflicting perceptual evidence available to them. Another finding of the set of experiments was that children did not base all their inductions on category membership and were able to differentiate between properties which can be projected based on category membership and those which rely on perceptual features. These results suggest that children are able to make inferences based on category membership.

There is evidence which suggests that children differentiate between natural kinds and artefacts and develop the notion of an essence only for natural

kinds. For instance, children of age five were able to recognize that an animal cannot be transformed into another animal. For example, a raccoon cannot be transformed into a skunk (Keil, 1989). Children were shown before and after images in which a raccoon was transformed into a skunk through some physical changes. Deceptive features (for example, 'smelly stuff' associated with skunks) were also included in the pictures. Children claimed that the transformed animal was still a raccoon. However, children were able to recognize that artefacts can change their identity through transformation. For example, a coffee-pot can be transformed into a bird feeder but a lion cannot be transformed into a tiger (Keil, 1989).

Gelman (1998) was able to demonstrate that as children grow older (from preschool to second grade) they draw more inferences within natural kinds (eg. carrots) than within artifacts (eg. balls). In very young children, inferences were based usually on generalizability of the property, where as older children took domain specific information into account. Gelman (1998) suggests that children were able to make more inferences for the natural kinds because children perceived the categories as more homogeneous as compared to artefact categories.

In this thesis, we are concerned with the causal aspect of essentialism. We examine whether perceptual features support the inference of a hidden cause, which might be the essence. The next section provides evidence that people do in fact infer hidden causes and causes are important in categorization and representation.

1.3 Causality in Categorization, Explanations and Essentialism

Causes are central to people's representations of categories. The causal status effect (Ahn et al., 2000) states that the position of a feature in a causal structure determines its centrality in representation and categorization. For example, Fish DNA is more important than 'having gills' since Fish DNA is the cause for the feature 'having gills'. In a category membership task, removal of a causal feature lowered category membership likelihood more than the removal of effect features. In free sorting tasks, people preferred to create categories which shared a common cause instead of a common effect. Further, goodness of exemplar judgements were affected by the presence of causally central features - deeper the missing cause, worse was the rating of the exemplar (Ahn et al., 2000).

A number of experiments show that even young children are able to identify causal features and that this information influences categorization and induction. Gopnik and Sobel (2000) were able to demonstrate that by two-and-a-half years of age, children use causal information to guide categorization and induction. Children were shown an object called 'blicket' which turned a machine off. In the categorization task, children were shown other objects, some of which could turn the machine off and some of which could not. Children had to label which objects were 'blickets'. In the induction task, children were shown objects, a few of which were labelled as 'blickets'. Children had to predict whether the object had the causal power to turn off the machine. Children were effectively able to use causal information even if there was conflicting perceptual information involved.

Gelman and Kremer (1991) studied causal explanations given by children for various properties of categories. Specifically, they studied whether children

recognize that causes can be inborn and internal. Children were shown a picture of a category (for example, a rabbit) and were told a particular property (for example, a rabbit has long ears) or behaviour (for example, a rabbit hops) which was relevant to the image. They were then asked a reason for the property/behaviour (Why does the Rabbit have long ears?) and were given two options: 'Did a person make long ears?' or 'Is there anything inside that made long ears?'. The list of behaviours were from two classes (as rated by adults): self-generated/self-sustained and other-generated/other-sustained. Children showed the belief that self-generated activities are more likely to have an inherent cause. Children as young as four realized that natural causes exist apart from human causes. Children were also able to identify different types of natural causes: inborn, intrinsic or growth.

Gelman (2003) introduces the concept of a causal essence: an entity which causes other category typical features. For example, the Y-chromosome is possibly the essence which is responsible for various surface features of men (moustache, beard, etc.), i.e. the Y-chromosome is a common causal variable which is responsible for a number of observable properties. People's representations of categories consist of a causal essence even if they are unaware of what the actual essence is (or even if an actual essence does not exist), i.e the causal essence is a placeholder.

Causal, representational essentialism suggests that people's representations of categories consist of a causal variable (the essence) which is responsible for various surface features. According to essentialist theories, the essence should be causally responsible for a number of properties rather than the essence being an effect or outcome of the properties. As Medin and Ortony (1989) suggest, twins are not twins because they are similar but rather they are similar because they are twins. This would imply that people's representations of categories should support the common-cause structure (Figure

1.1) over the common-effect structure (Figure 1.2), at least for the categories which tend to be essentialised. Ahn et al. (2001) suggest that people's representations of natural categories consist of a common cause structure. Keil (1989) suggests that the common effect structure is the underlying structure for artefact categories. Keil (1989) further suggests that natural kinds consists of richer and denser clusters of features as compared to artifact categories. In Chapter 3, we compare causal structures across categories. We also compare the correlation between pairs of features to understand if natural kinds have a larger number of highly correlated features.

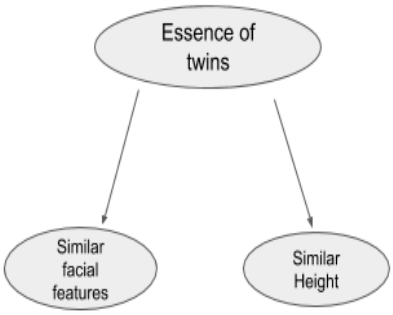


FIGURE 1.1: Common Cause

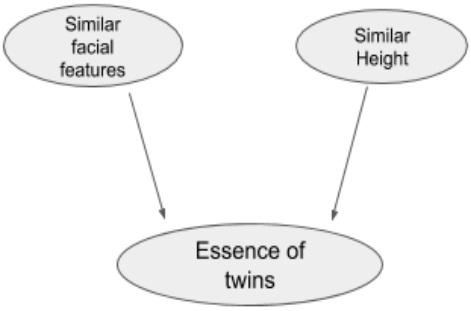


FIGURE 1.2: Common Effect

1.4 Why are people Essentialists?

Essentialist theories claim that people tend to develop the notion of causal, representational essentialism for natural categories (tigers, man etc.) and social categories (race, ethnicity etc.) but not artefact categories (Gelman, 2003). So far we have presented evidence which suggests that people are essentialists and can infer hidden causes. However, what has remained unanswered is why are people essentialist.

One possible reason is that essentialism is an inherent cognitive bias. It is possible that children, from birth, are biased towards thinking about natural kinds in an essentialist sense. Other possibilities are that essentialism is learned from the structure of the perceptual or linguistic environment, or that both the environment and innate biases play a role in the development of essentialist beliefs.

One kind of environmental input that might be relevant is the structure of the categories in the real world. What kind of features do objects have and do those features support the inference of a hidden cause. In this thesis we explore this possibility by comparing the likelihood of different causal structures.

Another possibility is that language plays a role in the development and strengthening of essentialist beliefs. A generic noun phrase[1] is a linguistic form which is used to express essential qualities about a category. Further, generics convey that a category is stable, structured and allows for category based inferences. Gelman and Tardif (1998) and Gelman et al. (2008) show that generic speech in child-directed speech is domain-specific and generic phrases are used more frequently while referring to animals and natural

[1]Generic noun phrases refer to categories or classes of objects rather than an individual object of a particular class; Example: Dogs eat meat. (generic); My dog does not eat meat (non-generic)

kinds. It is possible that adults are producing a larger number of generic phrases for animal categories because they themselves think of these categories in an essential sense. However, children develop essentialist beliefs even in the absence of generic speech. This potentially implies that generic noun phrases are not necessarily the only reason for the development of essentialist beliefs but help in the strengthening of inherent cognitive biases. The results of the study conducted by Gelman et al.(2008) are based on manually tagging generics in eight corpora from the CHILDES dataset. Their results suggest that child-directed speech is domain specific and biased towards animal categories. However, their results are based on a small dataset. In this thesis, we extend the original study conducted by Gelman et al (2008) to a larger corpus of child speech by developing an automated system for the identification of generic noun phrases.

1.5 Aim and Hypothesis

The aim of this thesis is to study why people develop the notion of a causal essence for natural kinds. We examine whether the visual and perceptual features of a category support the inference of a common cause. Since people tend to have stronger essentialist beliefs for natural categories as compared to artefact categories, we expect a greater number of animal categories to support the common cause structure as compared to artefact categories. Further, we study the role that generic noun-phrases play in the development of essentialist beliefs. We develop a computational framework to extend the initial study conducted by Gelman et al.(2008) on the CHILDES dataset. Our goal is to test the hypothesis that both children and adults produce a larger number of generic statements for natural categories as compared to artefact categories.

The rest of the thesis is broken up into two parts. In Part 1 we collect a dataset of perceptual features and compare causal structures across categories. In Part 2, we study the role of generic noun phrases in the development of essentialist beliefs.

Part 1

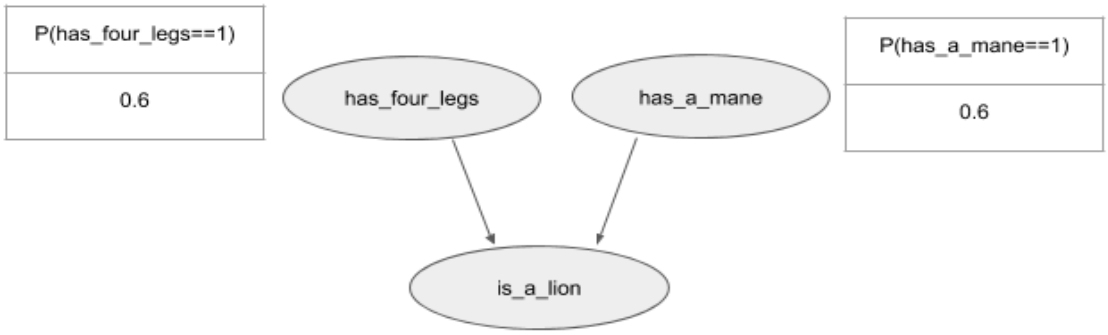
Chapter 2

Part 1: Methodology and Experiment

The goal of the first part of the thesis is to study whether perceptual features are responsible for latent causal variables in people's representations of categories. In order to do so, we collect a dataset of perceptual features and compare the likelihood of various causal structures. We first provide an introduction to Bayesian Networks and Causal Graphical Models, following which we describe the experiment which was conducted to collect the dataset. We compare the causal structures based on their likelihood in Chapter 3.

2.1 Bayesian Networks

The study relies on the use of Bayesian Networks to capture the causal structures present in people's representations. A Bayesian Network is a directed acyclic graph (DAG) which represents the joint probability distribution of a set of variables. Bayesian Networks consists of two parts: (a) the structure of the graph, i.e. nodes and directed edges and (b) conditional probability tables associated with every node (a conditional probability table specifies



has_four_legs	has_a_mane	P(is_a_lion==1| has_four_legs,has_a_mane)
1	1	0.9
1	0	0.2
0	1	0.35
0	0	0.01

FIGURE 2.1: Bayesian Network
A Bayesian Network consisting of three variables. Each node has a corresponding conditional probability table.

the probability of each value of a variable X given each possible combination of values of the Parent Variables of X). Together, the graph and set of conditional probability tables represent the joint probability distribution of the set of variables:

$$P(X_1, X_2, X_3 ... X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i)))$$

Figure 2.1 is an example of a Bayesian Network of three binary variables: $has\_four\_legs, has\_a\_mane, is\_a\_lion$.

A Causal Bayesian Network is a Bayesian Network where the parents of each vertex are it's direct causes. Learning the structure of a causal bayesian network from observed data, comprises of learning: (a) Structure Learning: the structure of the directed acyclic graph and (b) Parameter Learning: the conditional probability tables associated with every node.

has_a_mane	is_a_lion
0	0
1	0
1	1
0	0
0	1
0	1
0	0
1	1
0	0
1	1

TABLE 2.1: Binary Data
Data for two binary variables. We use similar data to estimate causal structures.

2.1.1 Parameter Learning

If we consider the case of only binary variables, each entry of the conditional probability table is equivalent to estimating the probability of a biased coin. Estimating all the parameters of the network is equivalent to estimating the probability of multiple biased coins.

Consider the Bayesian Network shown in Figure 2.2 and the observed data shown in Table 2.1. Estimating the parameters for each cell of the conditional probability table is like estimating the probability of getting a head when a biased coin is tossed. The maximum likelihood estimate of the probability of a head is: Number of Heads in observed data/ Total number of tosses. Therefore for the cell A_1, the Maximum Likelihood Estimate is 4/10. Similarly, the estimates for cell B_1 and B_2 are 3/4 and 2/6, respectively.

Parameter estimates can also be made by using the Maximum Aposteriori approach, i.e. a prior probability is multiplied with the maximum likelihood estimate. The conjugate prior for a binomial distribution is a beta-prior. If the variables (discrete) are not binary variables, the probability estimates are equivalent to estimating the probability of the sides of a dice (multinomial distribution) and the conjugate priors come from the Dirichlet family. In

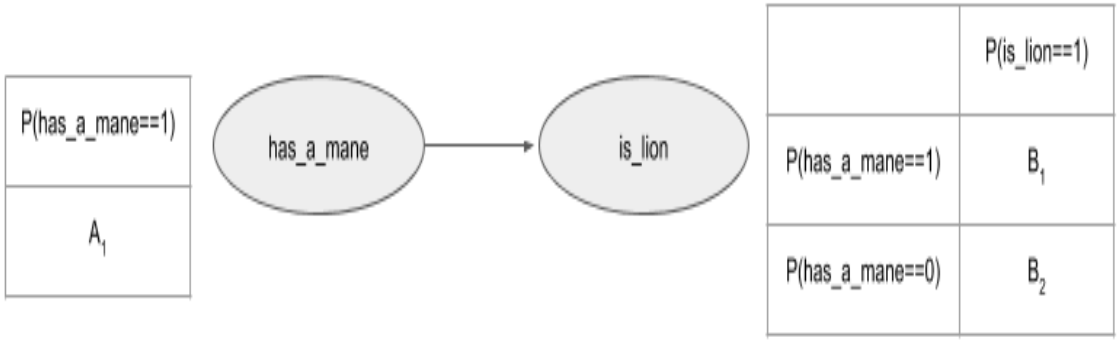


FIGURE 2.2: Bayesian Network
Estimating the conditional probability tables from data for a given graph structure.

Chapter 3, we use a dataset of binary variables and maximum likelihood estimates for parameter estimation.

2.1.2 Structure Learning

There are two broad categories of algorithms for learning the structure of the Bayesian Network.

Score-Based Learning Algorithms

In score-based methods, a score is assigned to each candidate Bayesian Network. For a given graph G and observed data D, the score is calculated as: $P(G|D)$ i.e. the posterior probability of the graph given data. By Bayes' theorem:

$$P(G|D) = \frac{P(D|G)*P(G)}{P(D)} \quad ,$$

where $P(D \mid G) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} \Gamma(a_{ij}, b_{ij}, s_{ij}, t_{ij})$.[1] P(G) is a prior distribution over various graph structures. The Bayesian Information Criterion (BIC) (Schawarz,

[1]
- q_i is the number of different combinations of values of parents of variable X_i
- a_{ij} and b_{ij} are the priors on heads and tails, respectively, of X_i when parents of X_i take their jth instantiation
- s_{ij} and t_{ij} are the observed number of heads and tails, respectively, of X_i when parents of X_i take their jth instantiation
- $\Gamma((a_{ij}, b_{ij}, s_{ij}, t_{ij})) = \frac{\Gamma(a_{ij}+s_{ij}, b_{ij}+t_{ij})}{\Gamma((a_{ij}, b_{ij})}$

1978), is a score-based method which uses a prior to penalize complex structures. $BICScore = log(P(D|G)) - \frac{d*log(N)}{2}$, where N is the size of the data set and d is the number of parameters of the structure. Scoring all possible Bayesian Networks is at least exponential in the number of variables. A number of heuristic/greedy methods exist to search the space of bayesian networks. The K2-algorithm (Cooper and Herskovits, 1992) is an example of a greedy search over the space of Directed Acyclic Graphs.

Constraint-Based Methods

Constraint-based methods use data to estimate whether certain conditional independencies between variables hold. Common constraint-based tests are the Mutual Information test when dealing with binary variables and the t-test for correlation when dealing with Gaussian variables. Given a set of conditional independencies in a probability distribution, constraint-based learning algorithms try to find an equivalence class of Directed Acyclic Graphs for which the Markov condition entails those conditional independencies (multiple DAGs might be compatible with the conditional independence tests). A Parental Ancestral Graph is a graphical structure which represents the features common to all the DAGs.

The PC algorithm (Spirites et al., 1993) is an example of a constraint-based algorithm which follows the causal sufficiency assumptions, i.e. there are no latent or selection variables. The PC algorithm finds the Markov equivalent class of DAGs using the conditional independence tests and creates a Parental Ancestral Graph. The parental ancestral graph has two types of arrows: $A \rightarrow B$ and $A - B$. The directed edge means that the edge is present in all DAGs of the Markov equivalent class. The undirected edge means that there is at least one DAG in the equivalence class having the edge $A \rightarrow B$ and at least one DAG with the edge $B \rightarrow A$.

The Fast Causal Inference algorithm (Spirtes et al., 2000) is a generalization of the PC algorithm that allows for arbitrary number of latent and selection variables. Parental ancestral graphs learnt using the FCI algorithm have the following types of edges: $\circ - \circ, \circ \longrightarrow, \longleftrightarrow, \longrightarrow, -$. Bi-directed edges are from hidden variables and undirected edges are from selection variables. A tail on an edge means that the tail is present in all DAGs of the Markov equivalence class. Similarly, an arrowhead on an edge means that this arrowhead is present in all graphs in the Markov equivalence class. A $\circ$- edge mark means that there is at least one graph in the Markov equivalence class where the edge mark is a tail and at least one where the edge mark is an arrowhead. The bi-directed edge ($A \longleftrightarrow B$) means that there is a latent causal variable between observed variables, A and B.

The Really Fast Causal Inference (RFCI) algorithm (Colombo et al., 2012) is a modern and faster implementation of the original FCI algorithm.

Most constraint-based learning algorithms follow a three-step procedure to estimate the set equivalence class of DAGs and the final Parental Ancestral Graph. In the first phase, the Markov blanket associated with each node is learnt to reduce the number of potential directed acyclic graphs. In the second phase, a skeleton of the directed acyclic graph is learnt. For each node, it's neighbours are identified, i.e. parents and children are identified but the arcs are still undirected. In the third phase, directions of arcs are identified.

Hybrid Algorithms

Hybrid algorithms use a combination of conditional independence tests and scoring over various graph structures. The Max-Min hill climbing algorithm (Tsamardinos et al., 2006) is an example of a Hybrid algorithm. The Max-Min hill climbing algorithm uses the conditional independence tests to find a skeleton of the graph and then performs a greedy search to find the exact

orientation of the edges.

In this thesis, we use the Really Fast Causal Inference algorithm to identify if there are latent variables present in people's representations of categories. We use the Max-Min hill climbing algorithm to find the best-fit graph, and we compare specific graph structures (common cause, common effect, etc.) based on their likelihood.

2.2 Overview

Our methodology to examine causal structures in representations of categories followed a two step procedure. The first involved collecting a dataset of perceptual features associated with psychological/mental representations of various categories. We collected data for animal and artefact categories, since we expect animal categories to be essentialised and artefact categories to be non-essentialised. The second step involved the use of Bayesian Networks to compare the inferred causal structures for each of the categories.

In sections 2.3 and 2.4 we describe the experiments which were used for data collection. We analyse the data in Chapter 3.

2.3 Experiment 1

A core part of conceptual representation depends on the features associated with a particular concept. For example, cats are associated with features such as 'meows', 'has four legs', 'is furry', etc. (Kiefer et al., 2011). Semantic property norms have been used to capture the features associated with representation of concepts and to explore aspects of semantic property representation (Devereux et al., 2014).

In property norm studies, participants are asked to list features that they associate with a particular concept. People are presented with a concept (for example, the word 'zebra') and are asked to list features that they associate with the concept. We conducted a variant of standard property norm/feature listing studies to capture features associated with representations of concepts based on visual stimulus. To compare causal structures within a category, we needed a dataset of features for multiple examples of a category. For example, we wanted multiple images of zebras and corresponding features for each of those images, as opposed to a single list of features for the concept zebra. The study allows us to capture features for individual exemplars of a category and learn a causal structure based on this data.

In experiment 1, participants were shown different images for each category/concept and were asked to rate the applicability of certain features. These features were chosen from standard property norm data sets. For example, the participant was shown an image of a lion and corresponding features would be 'has a mane', 'can roar', 'has a tail', etc. Participants of the study were expected to rate the applicability of each feature on a scale of 1 to 7 (1: the feature is not relevant to the image/category and 7: the image is highly relevant to the image/category). This dataset of ratings was used to compare the likelihood of different causal structures.

2.3.1 Design

Choice of Categories/Concepts

In order to ensure a wide variety of concepts were included, natural categories comprised of mammals, birds, fish and reptiles. Similarly, artifact categories comprised of furniture, utilities, musical instruments and vehicles. 18 categories, comprising of 9 natural kinds and 9 artefact concepts were identified. The choice of categories was such that the concepts existed in standard

property norm data sets (McRae et al., 2005 and Devereux et al., 2014) and the category was part of the 1000 categories used in the ImageNet challenge (Deng et al., 2009). The categories were:

- Natural: ANT, PIG, PEACOCK, SNAIL, GORILLA, FLAMINGO, IGUANA, LION, GOLDFISH
- Artefacts: BALLOON, BUCKET, CANDLE, DESK, FLUTE, MICROWAVE, NECKLACE, TAXI, UMBRELLA

Features

Features associated with each category were chosen from two property norm data sets: McRae et al., 2005 and Devereux et al., 2014. Eight features were selected from each category using the two studies. To ensure a systematic selection of these eight features for each category, the highest weighted features[2] which were common between the two property norm studies were selected first. If the requirement of eight features was not satisfied, the highest weighted features from the remaining set were chosen. The set of features was restricted to perceptual features and features based on category hierarchy such as 'is a bird' were excluded. The final set of features for each category can be found in Appendix A.

For two categories: LION and TAXI, an additional feature which played the role of an attention check was displayed. This feature was 'is an animal' for LION and 'is a vehicle' for TAXI. Data collected from participants who rated either of the features lesser than 5 (on a scale from 1 to 7) was excluded in the analysis.[3]

Choice of images

120 images were chosen for each category from the ImageNet dataset (Deng

[2]Weight in a property norm study is the number of people who listed the feature.

[3]Data for one participant was not included because of a score less than 5 on the attention check features

et al., 2009).[4] To ensure that images spanned a range of category typicality/representativeness, but at the same time typicality was uniform across categories, we followed the following procedure:

- Every image available on ImageNet of a particular category was passed through the pretrained GoogleNet model (Szegedy et al., 2015) and the final softmax layer value for the class was stored.
- For every category, a random sample of 20 images was selected from the set of images which had a softmax value in the range 0.8-1. A random sample of five images was selected from the range 0.6-0.8 and another random sample of five images was selected from the range 0.4-0.6. (The softmax layer of deep neural networks can accurately predict category typicality ratings for images (Lake et al., 2015).)

The average category typicality of images used for each category can be found in Appendix A.

2.3.2 Participants

366 participants (187 Males, 173 Females, 6 Others) were recruited using Amazon Mechanical Turk. Age of the participants ranged from 18 to 72 years, with a mean age of 38.4 years. Mechanical Turk is an online market which offers a diverse and large subject pool for conducting experiments known as Human Intelligence Tasks (Mason and Suri, 2011). Recruiting 366 participants ensured that we were able to get approximately 3 people to rate each image to perform a smoothing/averaging step.

[4]We wanted a dataset which consists of 120 positive examples and approximately 50 negative examples. Collection of negative examples has been discussed in Experiment 2.

2.3.3 Procedure

The study was conducted online using Amazon Mechanical Turk's Human Intelligence Tasks.[5] Participants were first shown a set of instructions and were quizzed on their understanding of the questions. They were allowed to proceed only if they answered all the questions correctly.

In the experiment, each participant was shown a sequence of 18 images - one for each category. Along with each image, 8 corresponding features were displayed. Participants were expected to rate on a scale from 1 to 7, the applicability of a particular feature to the image (7 indicated that the feature was highly relevant and 1 indicated that the feature was not relevant to the image). For example, a participant was shown the image of a snail (Figure 2.3) and the corresponding features were leaves a trail, is found in gardens, is slimy etc. The order of the categories and the order of the features within each category was random for each participant.

Henceforth, this data set has been referred to as Dataset 1. Experiment 1 allowed us to capture people's ratings of features for a category along the relevant dimensions. For example, the features for taxi were is_black, has_a_meter, etc. To capture people's ratings of images along non-relevant dimensions, we conducted an additional experiment.

2.4 Experiment 2

Data collected using Experiment 1, allowed us to collect applicability ratings for various features for 120 images of each category. In order to capture people's ratings of features which might not be relevant to the category, a variant of the experiment was conducted. This was done so that we could

[5]The study can be accessed at: featurescoring.appspot.com. Code for the experiment was written in JavaScript and can be accessed at: github.com/samarth1397.

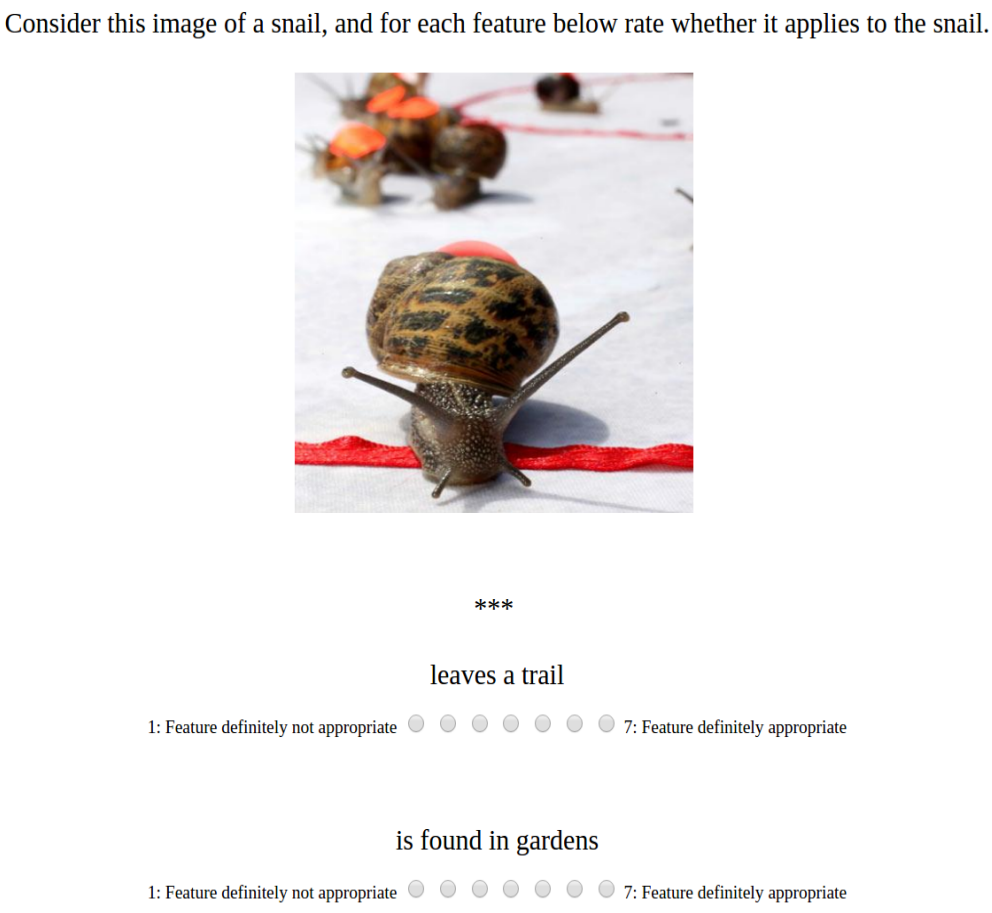


FIGURE 2.3: Experiment 1 Task

Participants saw an image and a list of eight features as shown in the image. Features were from the property norm studies conducted by Mcrae et al. (2005) and Devereux et al. (2014). Participants were expected to rate the applicability of each feature on a scale of 1 to 7.

learn a causal structure using both positive and negative examples which have been rated along the same set of features. The difference was that participants were not shown 8 features corresponding to the category. Instead, participants were shown a random sample of 8 features from the pool of features of other categories. For example, a participant was shown the image of a lion and the corresponding features were: 'leaves a trail', 'has a shell', 'eats bananas', 'is large', etc.

2.4.1 Design

The 3 most representative images of each category , i.e. the images which had the highest softmax scores out of all the images in ImageNet for the particular category, were selected. Eight features were sampled randomly from the pool of features and were displayed along with the image.

For five categories: LION, TAXI, CANDLE, ANT and FLUTE, an additional feature which played the role of an attention check was displayed. Data collected from participants who rated either of the features lesser than 5 (on a scale from 1 to 7) was excluded in the analysis.[6]

2.4.2 Participants

141 participants (86 Males, 54 Females, 1 Other) were recruited using Amazon Mechanical Turk. This number ensured that every feature was rated by approximately 3 people for a particular image. Age of the participants ranged from 24 to 75, with a mean age of 38.6 years.

[6]None of the participants rated any of the features lesser than 5 and so no data was discarded.

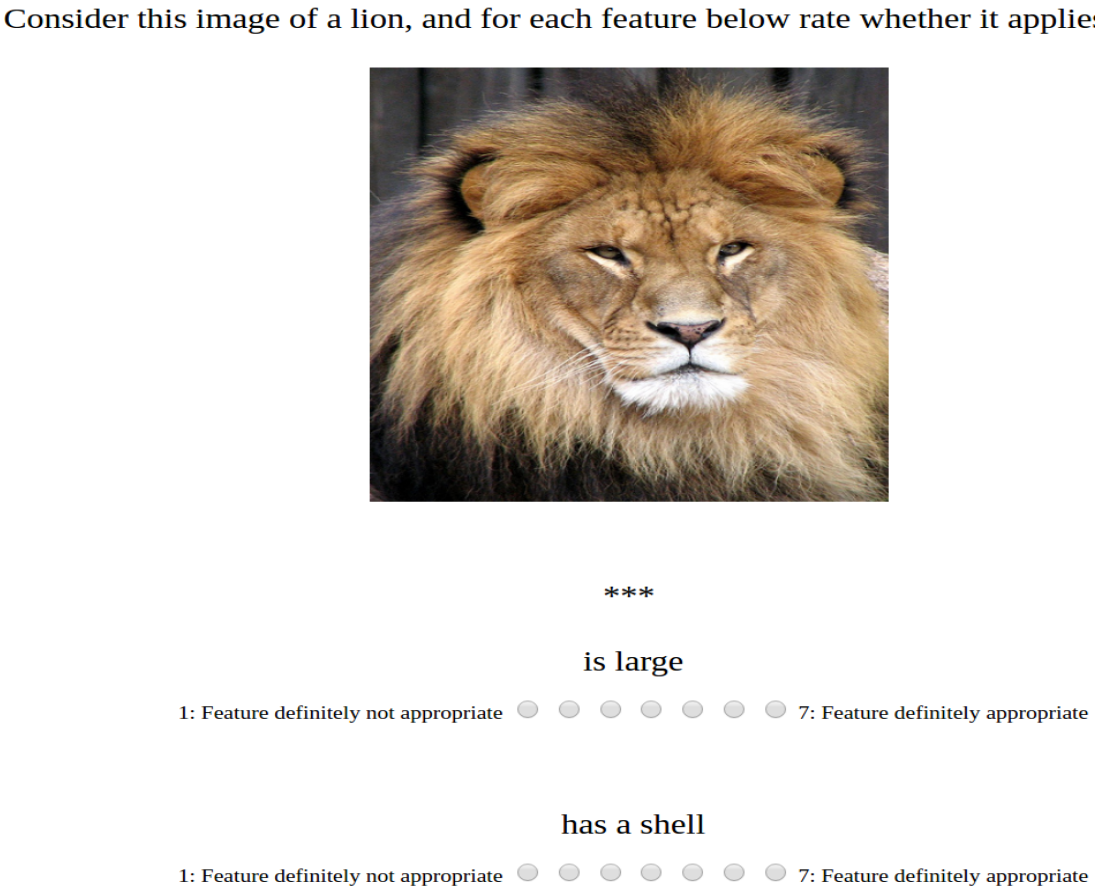


FIGURE 2.4: Example Image

Participants saw an image and a list of eight features from a pool of features for other categories (Two features have been shown in this image.). This experiment allowed us to collect a data set where examples from categories were rated on a set of features from different categories. We merge Dataset 1 and Dataset 2 into Dataset 3, which we use to compare the likelihood of different causal structures.

2.4.3 Procedure

The experiment setup was similar to Experiment 1 and was conducted online using Amazon Mechanical Turk's Human Intelligence Tasks[7].

Similar to experiment 1, each participant was shown a sequence of 18 images and rated applicability for the corresponding features. Figure 2.4 shows an example of an image and a list of features which the participant saw. Henceforth, data collected from this experiment has been referred to as Dataset 2.

[7]The study can be accessed at: otherfeaturescoring.appspot.com. Code for the experiment was written in JavaScript and can be accessed at: github.com/samarth1397.

In the next chapter we use Dataset 1 and Dataset 2 to compare various causal structures which are best supported by the data.

Chapter 3

Part 1: Analysis

In this chapter, we analyse the data collected from the two experiments. We compare various causal graphical models to understand which models are supported by the data and whether the data supports the inference of hidden causal variables. In the rest of this chapter, we use the words essential and animal interchangeably. Similarly, we use the words non-essential and artefact interchangeably.

3.1 Data

Dataset 1 and Dataset 2 were merged to create Dataset 3. Dataset 3 consisted of 171 images and ratings of features for each category. 120 of those images were from Dataset 1. The remaining 51 images and ratings were from Dataset 2. For example, for the category GORILLA, the data consisted of 120 images (from Dataset 1) which were actually images of gorillas (positive examples) and 51 images of other categories (from Dataset 2 - negative examples) and the corresponding ratings for features of the concept Gorilla (from Experiment 2). The data for gorilla, for example, is shown in Table 3.1:

image	eats bananas	is black	is dangerous	has fur/hair	is large
gorilla-1	7	7	7	2	2
gorilla-2	7	7	4	6	7
gorilla-3	4	7	4	6	3
gorilla-4	4	6	4	7	5
gorilla-5	5	6	4	5	7
gorilla-6	5	7	4	5	7
.....	....	....	....	....	....
lion-1	1	1	4	5	7
lion-2	1	1	6	2	4
lion-3	1	1	7	7	3
goldfish-1	1	1	4	4	4
.....	....	.....	....	....	....

TABLE 3.1: Data for gorilla (example)
Data for a category consists of positive and negative examples which have been rated on a list of eight features. Positive examples and ratings were from Experiment 1 and negative examples and ratings were from Experiment 2.

To get a sense of how features within a category might be related, we calculated the correlation[1] between feature pairs. Figure 3.1 shows the average correlation between pairs of features for each category, along with the distribution of correlations within a category. We find that the average pairwise correlation for essential categories is 0.66 which is higher than the average pairwise correlation for non-essential categories (0.57). However, in both groups there are bundles of features with high correlation (>0.8). Individual heat maps of correlation between features for each category can be found in Appendix B. This difference in correlation between pairs of features for essential and non-essential categories served as motivation to analyse the causal structures across categories.

We used Dataset 3 to compare the likelihood scores of different causal structures. Two different strategies were used to analyse the causal structures which are best supported by the data. In the first approach, we introduced a variable which plays the role of an ideal causal essence which is shared by all members of a category. In the second, we use the Really Fast Causal Inference

[1]Correlation was calculated using the Kendall Rank Correlation Coefficient.

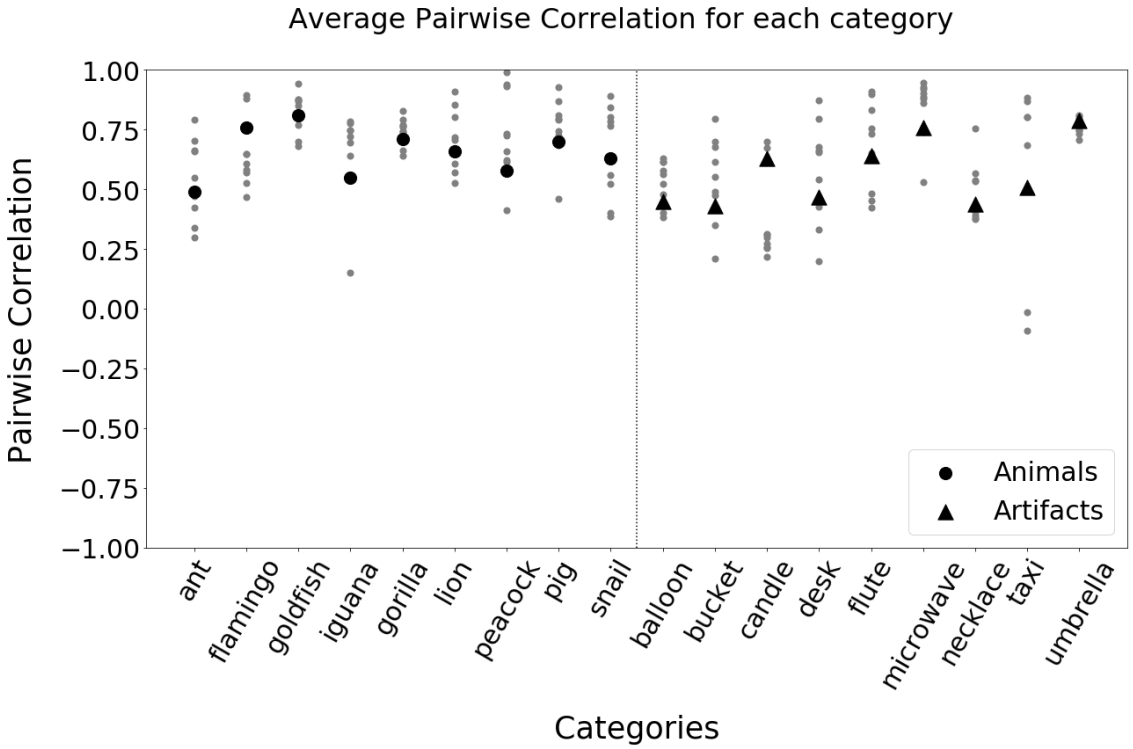


FIGURE 3.1: Average Pairwise Correlations
Correlation between pairs of features of a category. The smaller dots indicate correlation for each pair and the larger marker indicates the average of all pairs. This analysis was useful in understanding if natural categories consist of pairs of highly correlated features. We find that for both artefacts and animals, there are pairs of highly correlated features, although, the average correlation for animal categories is slightly higher.

algorithm to identify if hidden causes are present between pairs of variables.

These two strategies are discussed in detail in the next two sections.

3.2 Explicitly defining the essence variable

A binary variable which served as a dummy variable for the ideal *causal essence* was defined for each category. As described above, Dataset 3 comprises of 171 images for each category. The essence variable was set to 1 for all images which were actually members of the category, i.e. 120 images, and was set to 0 for all images which are not members of the category, i.e. 51 images. Using this data, the likelihood of the common cause structure

was compared against the common-effect structure and the independent features and essence structure. In the rest of this chapter, common cause refers to the causal structure shown in Figure 3.2. Similarly, common effect refers to the structure shown in Figure 3.3 and independent refers to the structure shown in Figure 3.4. Best-fit refers to the best-fit Bayesian Network as returned by the Max-Min Hill Climbing algorithm. Best-fit-independent refers to the best-fit Bayesian Network as returned by the Max-Min Hill Climbing algorithm, with the constraint that the essence variable is independent of (not connected to) any of the other variables.

The aim of this analysis is to understand whether the common cause model has a higher likelihood, indicating that it is better supported by the data, as compared to causal structures which are not aligned with essentialism theories (for example, common effect, independent, best-fit-independent, etc.). Support for the common cause model for animal categories would suggest that people acquire essentialist beliefs based on perceptual input.[2]

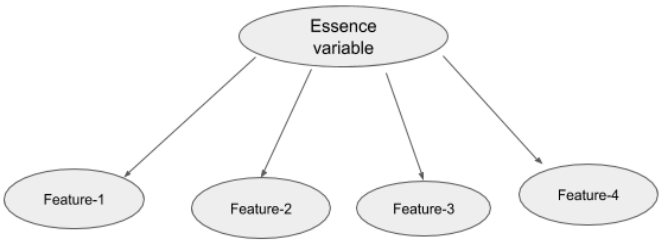


FIGURE 3.2: Common Cause

[2]Due to a small sample size of 171 images per category, the feature ratings in Dataset 3 were converted to binary variables using binning. Additional analysis is performed using 3 Bins instead of 2 Bins. The results can be found in Appendix B.

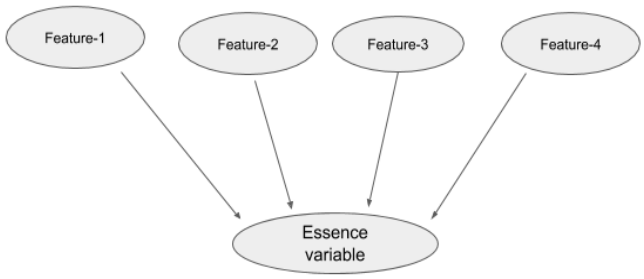


FIGURE 3.3: Common Effect

FIGURE 3.4: Independent

3.2.1 Top 4 Features

We first compared the common-cause, common effect, independent and best fit models using the four highest-rated features.[3] We restricted ourselves to the top 4 features due to a small sample size of 171 data points per category.

Common-Cause vs Independent

We compared the likelihood of the common cause structure with the independent structure by calculating the Bayes Factor.[4] Figure 3.5 shows that for both essential as well as non-essential categories, the common cause structure is strongly supported by the data as compared to the independent structure.

[3] By highest rated, we mean the features which had the maximum weight in the property norm studies conducted by McRae et al. (2005) and Devereux (2014) Weight refers to the number of people which listed the property in the task.

[4] $likelihood of Hypothesis1 / likelihood of Hypothesis2$; Bayes Factor $> 10^2$ indicates very strong evidence for Hypothesis 1 as compared to Hypothesis 2. We report ratios of likelihoods throughout the thesis. However, ratios of BIC returns similar results since the graphs do not differ substantially in the number of parameters.

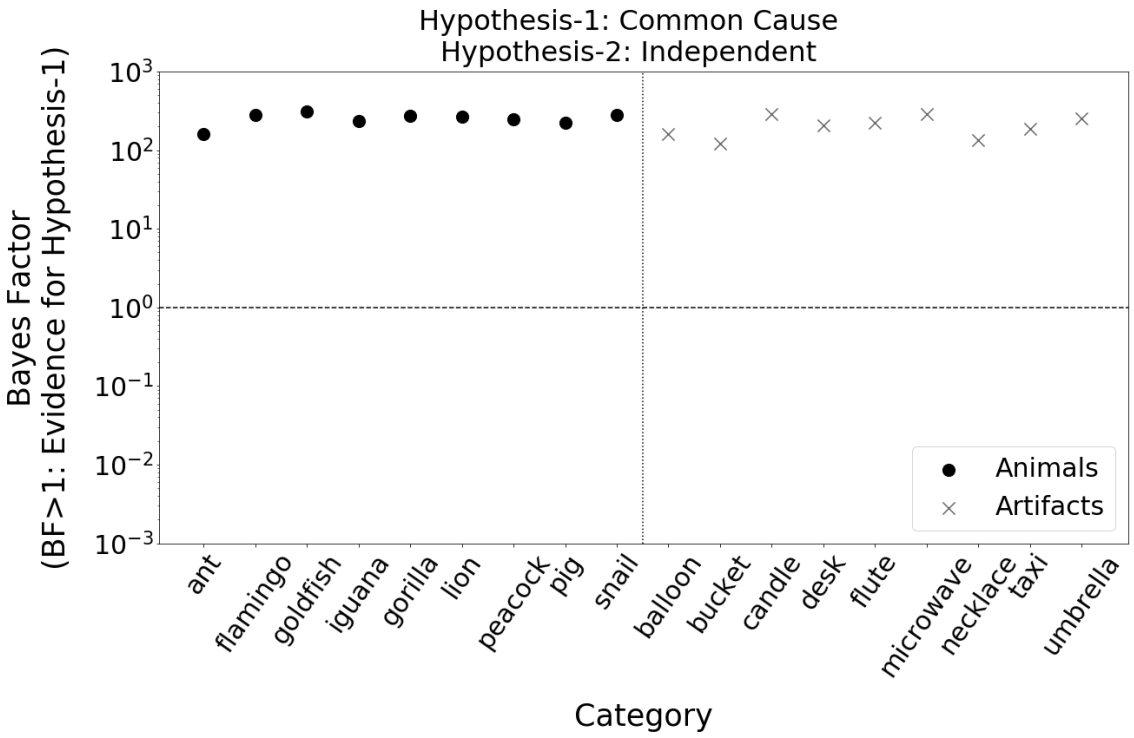


FIGURE 3.5: Bayes Factor(Common Cause / Independent)
The figure shows the performance of the common-cause graph as compared to the independent graph. A Bayes Factor >100 indicates very strong evidence for the common cause model as compared to the independent model. We find that for all categories, the common cause model is better supported by the data as indicated by the high Bayes factors. The top 4 features were used in this comparison.

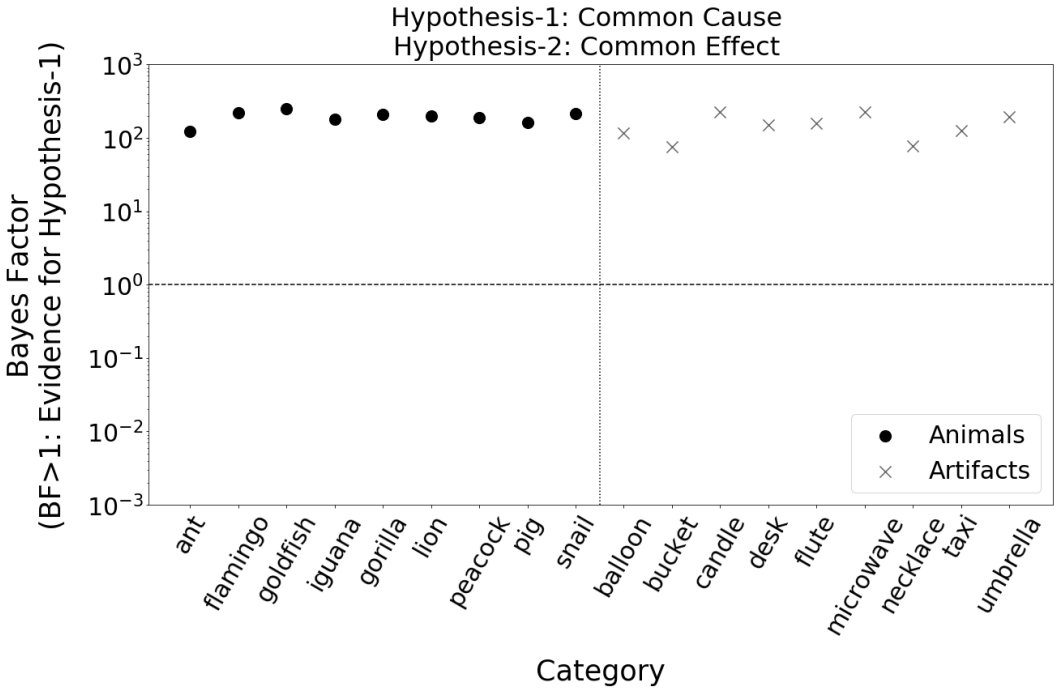


FIGURE 3.6: Bayes Factor(Common Cause / Common Effect)
The figure shows the performance of the common-cause graph as compared to the common effect graph. We find that for all categories, the common cause model is better supported by the data.

Common-Cause vs Common-Effect

As shown in Figure 3.6, the Bayes Factors reveal that the common cause structure is better supported by the data as compared to the common-effect structure. This is true for all categories and there is no variation between essential and non-essential categories.

Best-fit model vs Common-Cause

We also examined the difference between the common-cause structure and the best fit model. Low Bayes factors, as shown in Figure 3.7, indicate that the likelihood of the common-cause structure is not much lower than the likelihood of best-fit graph, for a number of essential and non-essential categories. In fact, for 3 animal categories (goldfish, gorilla and snail) and 1 artifact category (flute), the common cause model is the best-fit model. The lower Bayes factors suggest that for a number of categories, the common-cause model is a good approximation for the best-fit model.

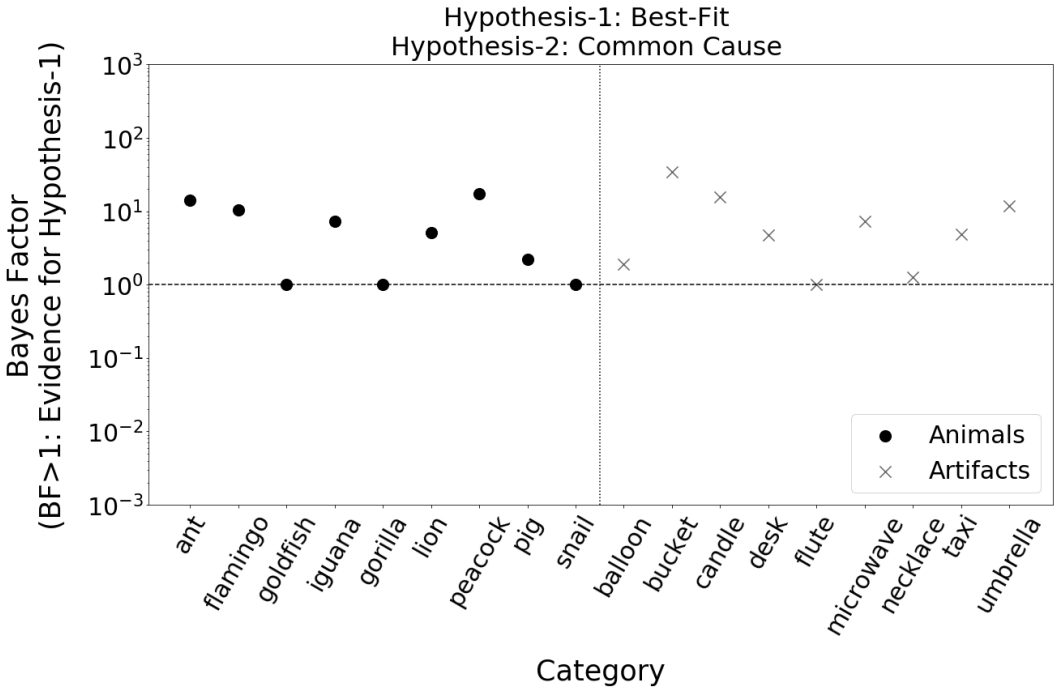


FIGURE 3.7: Bayes Factor(Best Fit / Common Cause)
The figure shows the ratio of the likelihood of the best-fit graph to the likelihood of the common-cause graph. For 3 essentialised and 1 non-essentialised category, the common cause graph is the graph which fits the data the best, as indicated by a Bayes Factor equal to one.

To get a better sense of how close the common cause structure is to the best-fit model, we calculated the out-degree of the essence dummy variable for each category. The out-degree of the essence variable is the number of connections of the form: $essence \longrightarrow feature$. A high out-degree indicates that a node is important in conceptual representations (Ahn et al., 2000). We find that the average out-degree for both essentialised and non-essentialised categories are close to each other. However, in a larger number of essentialized categories the dummy variable is connected to all four nodes. The out-degree of the dummy variable for various categories is shown in Figure 3.8.

The above analysis was performed using the four highest-weighted features. We examined if the results discussed above held valid using all combinations of 8 features taken 4 at a time, i.e. for each category, we constructed 70 different graphs. The average Bayes Factors for the common cause structure when compared to the independent model are shown in Figure 3.9. Figure 3.10

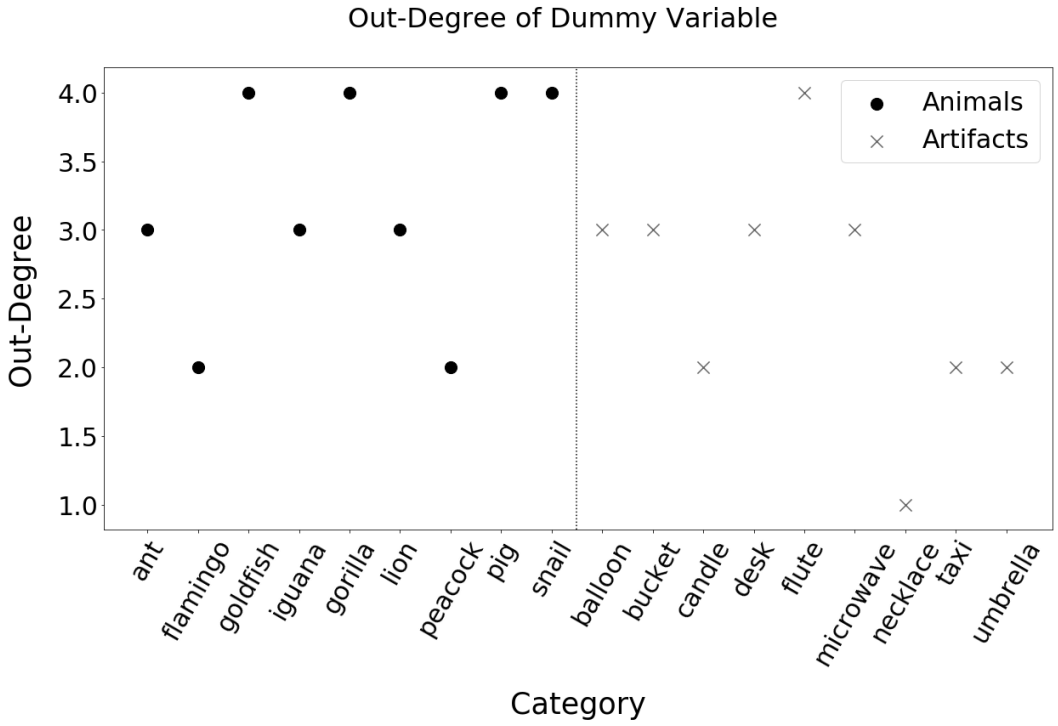


FIGURE 3.8: Out-degree of dummy variable
The out-degree of the essence (the number of variables which the essence is a cause for in the best fit structure) has been shown in the figure above. A node with a high out-degree is considered to be important in conceptual representation.

shows the average Bayes factor of the common cause model when compared to the common effect structure. The graphs indicate that for all categories. the common-cause model is better supported by the data than the common-effect model and the independent model.

The analysis was repeated using the top 3 features instead of the top 4 features. The results follow a similar pattern and can be found in Appendix B.

3.2.2 Using All 8 features

Due to a limited data sample, we did not compare the common-cause model with the common effect model[5] using all eight features. However, we could

[5]The conditional probability table for the dummy variable in the common-effect model has 2^7 rows, and would require at least $10 * 2^7$ data points for a reasonable probability estimate.

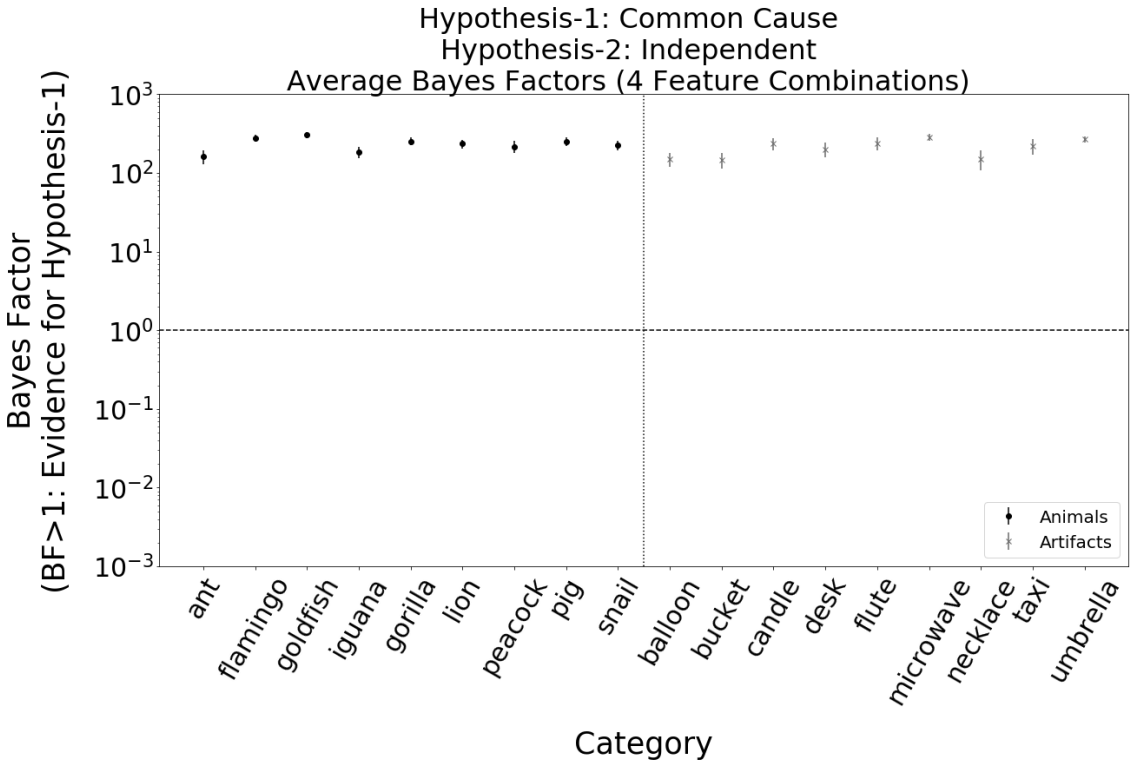


FIGURE 3.9: Average Bayes Factor(Common Cause / Independent)

In the figure above, for each category, the average Bayes factor of the common cause model when compared to the independent model has been plotted. The average has been calculated by averaging individual Bayes factors for 70 different graph structures by considering 4 variables (from a set of 8 variables) at a time. The graph shows the mean Bayes factor along with the standard deviation for each category. Average Bayes factors indicate that the common cause model is better supported by the data as compared to the independent model.

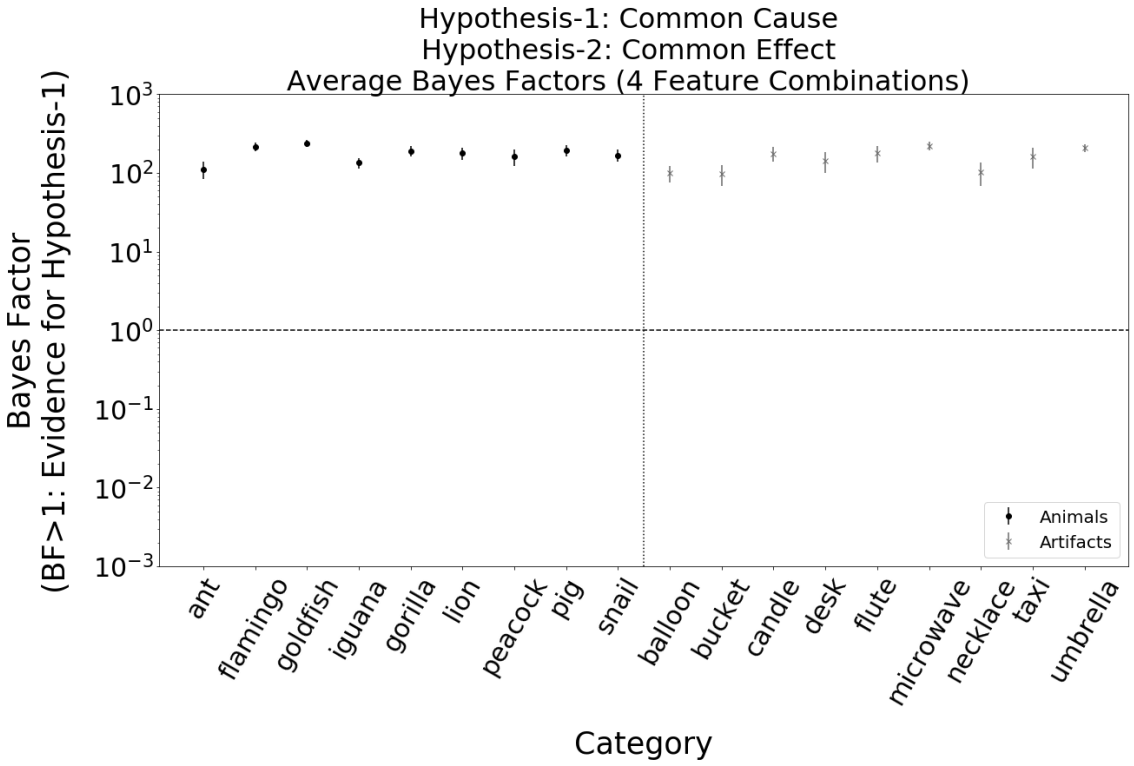


FIGURE 3.10: Average Bayes Factor(Common Cause / Common Effect)

In the figure above, for each category, the average Bayes factor of the common cause model when compared to the common effect model has been plotted. The average has been calculated by averaging individual Bayes factors for 70 different graph structures by considering 4 variables at a time. The graph shows the mean Bayes factor along with the standard deviation for each category. Average Bayes factors indicate that the common cause model is better supported by the data as compared to the independent model.

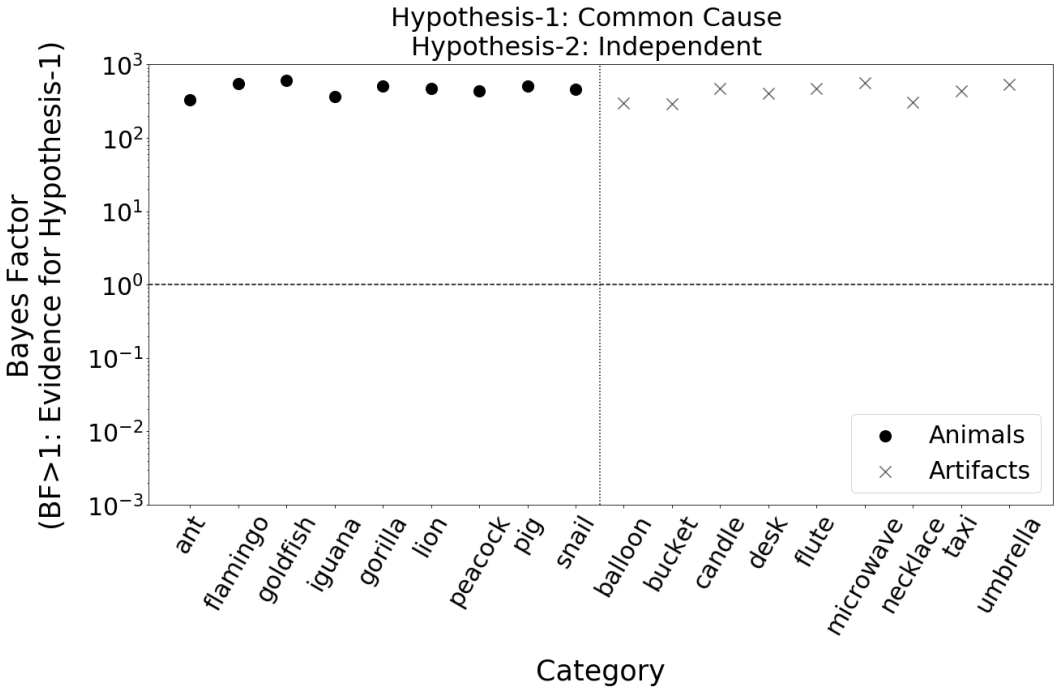


FIGURE 3.11: Bayes Factor(Common Cause / Independent) - All eight features

Ratio of the likelihood of the common cause model to the likelihood of the independent model has been plotted in the graph above. Data for all eight features was used to estimate the parameters of the graph. Bayes Factors indicate that the common cause model is better supported by the data as compared to the independent model.

compare the common-cause model with the independent model and the common-cause model with the best-fit and best-fit-independent model (with certain restrictions on number of parents).

Common-Cause vs Independent

We examined if the common-cause structure is better supported by the data when compared to the independent model even in the case if all 8 features are used. Figure 3.11 shows that for all categories, the data supports the common-cause structure as compared to the independent model, with no differences between essential and non-essential categories.

Common-Cause vs Best-Fit model

Further, we compared the best-fit model to the common-cause structure. Due to a small sample size, we put the restriction that every node can have a maximum of four parent nodes in the best fit model. This constraint ensured that

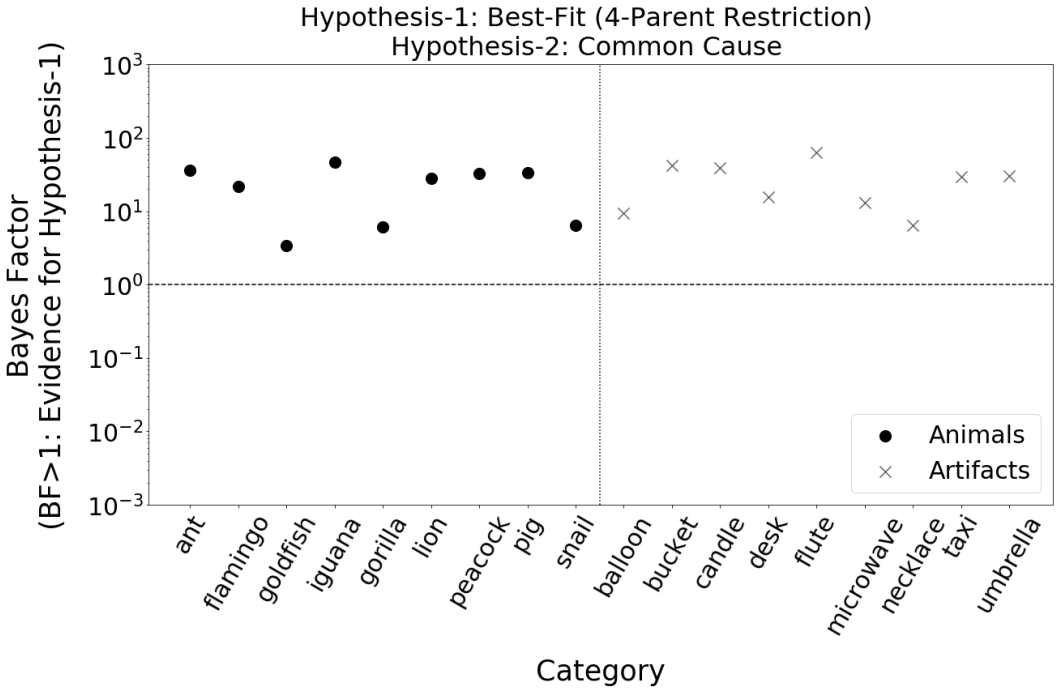


FIGURE 3.12: Bayes Factor(Best Fit / Common Cause) - All eight features

Ratio of the likelihood of the best fit model to the likelihood of the common cause model has been plotted in the graph above. Data for all eight features was used to estimate the parameters of the graph. The best-fit model was found using the Max-Min Hill Climbing algorithm with a constraint on the number of parents of a node - maximum of four.

we had at least 10 data points (approximately) to make an estimate for each of the conditional probabilities associated with the graph structure. Figure 3.12 shows that for both essential and non-essential categories, the common-cause model does not perform very poorly when compared to the best-fit model.

Figure 3.13 shows the out-degree of the essence dummy variable in the best-fit model using all 8 features (with the 4-parent constraint).

Table 3.2 shows the nodes which had the highest out-degree in the best-fit network for various essential and nonessential categories. We find that for 4 animal categories and 3 artifact categories, the dummy variable is the node with the highest out-degree.

The common-cause model was also better supported by the data than the

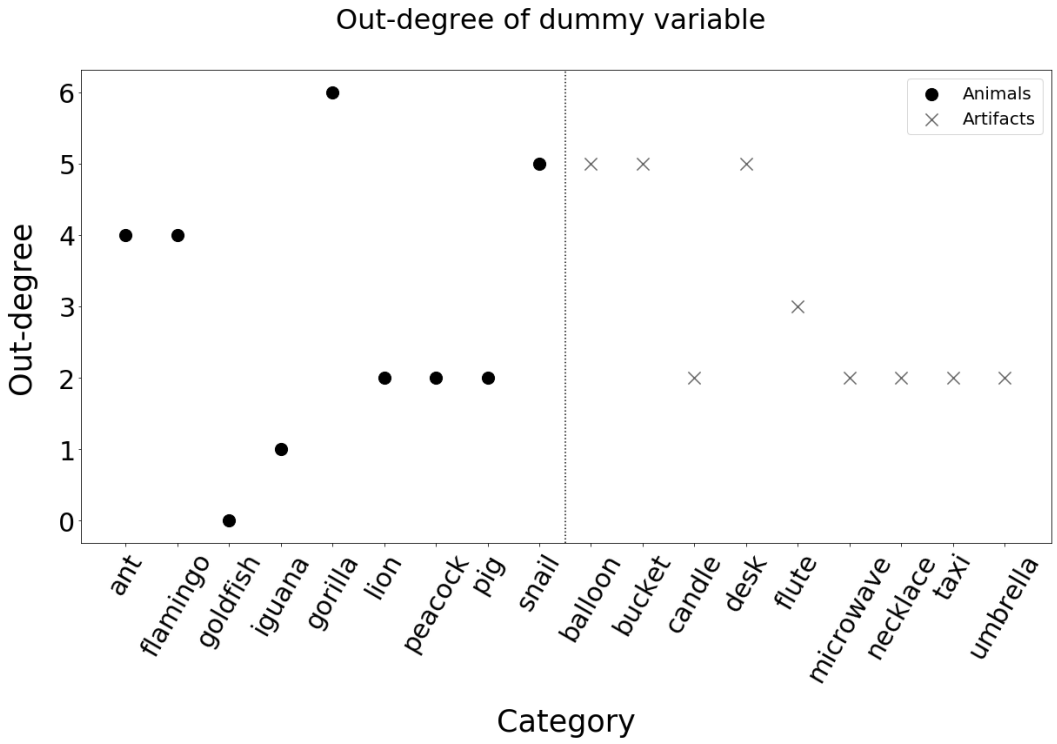


FIGURE 3.13: Out-degree of dummy variable in best-fit model
In the figure above, the out-degree of the essence variable in the best fit structure has been plotted for each category. A high out-degree indicates that a feature is important in conceptual representations.

category	features	degree
ant	dummy	4
flamingo	dummy	4
goldfish	has.fins	6
iguana	can.eat.insects,has.a.tongue,has.legs	2
gorilla	dummy	6
lion	can.roar	3
peacock	has.a.beak	6
pig	oinks	4
snail	dummy	5

TABLE 3.2: Nodes with the highest out-degree: Essential Categories

category	features	degree
balloon	dummy	5
bucket	dummy	5
candle	has.a.wick	3
desk	dummy	5
flute	produces.music,dummy	3
microwave	is.electric	3
necklace	is.worn.around.the.neck	4
taxi	used.for.passengers	3
umbrella	has.spokes	3

TABLE 3.3: Nodes with the highest out-degree: Non-Essential Categories

best-fit-independent model, as shown in Figure 3.14 (the best fit model with the constraint that the essence dummy variable is not connected to any of the features). The ratio of marginal likelihoods[6] of the two graphs were also compared (Graph in Appendix B). The results are similar, indicating support for the common cause graph as compared to the best-fit-independent graph.

A possible argument to our methodology of comparing causal structures using Dataset 3 is that we are using negative examples from all categories. For example, the data for the category GORILLA consists of 120 positive examples of gorillas and 50 negative examples of non-gorillas which include instances from all artefact categories and other animal categories. However, it is possible that people actually construct causal structures using negative examples only from related categories. For example, the negative examples for the category GORILLA should be limited to instances of other animals and should not include artefact categories. To verify if our results were affected by this, we repeated the analysis on a reduced dataset, where negative examples were from the same group, i.e. animal negatives for animal categories and artefact negatives for artefact categories. The results are consistent with the analysis presented above and can be found in Appendix B.

[6]Ratio of the likelihood of the distribution across the observed variables in the two graph structures, i.e. by marginalizing the dummy variable

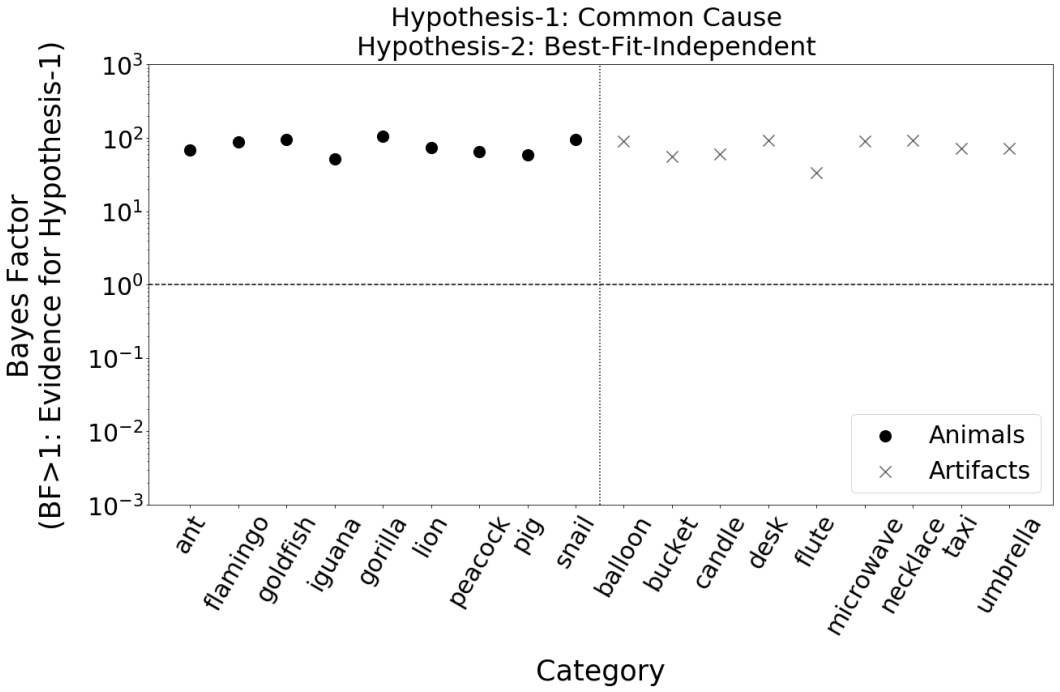


FIGURE 3.14: Bayes Factor (Common Cause/Best-Fit-Independent)

Bayes Factors of the Common Cause Graph when compared with the Best-Fit-Independent model. Best-fit-independent represents the case when the essence variable is not causally responsible for any perceptual features.

3.3 Identifying Latent Variables using the RFCI algorithm

The analysis above relies on the use of a dummy variable which plays the role of a causal essence which is shared by all members of a category. The second approach that we followed was to use the Really Fast Causal Inference Algorithm on Dataset 3 after binning the variables into binary variables. We identified pairs of variables which share a latent cause ($\longleftrightarrow$). The tuning parameter alpha[7] was increased in steps of 0.0025 from 0.01 to 0.06 and the causal structure was learnt at each step, i.e. a total of 20 causal structures were learnt.

Tables 3.4 and 3.5 shows the number of pairs of variables which had a latent

[7]The RFCI algorithm has a tuning parameter which is the significance level of the conditional independence test

Category	Pairs Sharing a Latent Cause
Ant	8
Flamingo	0
Goldfish	1
gorilla	1
iguana	1
lion	1
peacock	0
pig	1
snail	3

TABLE 3.4: Number of variable pairs sharing a common cause (essential categories)

Category	Pairs Sharing a Latent Cause
balloon	2
bucket	6
candle	2
desk	2
flute	1
microwave	2
necklace	1
taxi	0
umbrella	2

TABLE 3.5: Number of variable pairs sharing a common cause (non-essential categories)

cause between them, for various essential and non essential categories. We include a pair of variables if at least 3 causal structures out of 20 suggest that there is a latent cause between the pair of variables. The variable pairs which shared a common cause can be found in Appendix B.

We find that the for both essential and non-essential categories there are feature pairs which share a common cause, indicating that the data supports the inference of a common cause for both groups.

3.4 Discussion

We first summarize our results and discuss what the implications might be, following which we discuss if there were any limitations in our study which might have affected the results.

3.4.1 What do our results mean?

The analysis in this chapter suggests the following:

- The common cause structure is better supported by the data as compared to the common-effect, independent structures and best-fit-independent structures.
- The common cause structure is not much worse than the best fit model of the graph. This is evident from the lower Bayes factors and from the high out-degree of the essence dummy variable in most categories.
- Hidden causal variables exist in both essential and non-essential categories.

Essentialism suggests that people's representations of natural kinds consists of a hidden, causal property which is responsible for various surface features. The goal of the first part of the thesis was to explore whether perceptual features support the inference of hidden causes. Further, we expected to find differences in causal structures between essential and non-essential categories. As shown by our rational model of structure learning, the key take-away point is that the perceptual input supports the inference of hidden causes.

However, evidence for the common-cause model and presence of latent features is insufficient to justify why people are essentialists. We find that the

common-cause model is better supported by the data for both essential and non-essential categories, implying that although we can possibly explain why people have essentialist beliefs for animal categories based on support for the common-cause model, we cannot explain why they do not have essentialist beliefs for artefact categories.

In our analysis, we also find that for various categories from both groups (essential and non-essential), the essence variable has a high out-degree, i.e. it is a causally important feature. This aligns with the Causal Status Effect (Ahn et al., 2000) which states that the importance of a feature in a causal structure determines it's conceptual importance. A high out-degree of the essence variable explains why the essence is important in people's representations of animal categories. However, we are unable to explain why the same dummy variable is not important or does not exist in representations of artefact categories.

In the next section we speculate about potential reasons why people are not essentialists about artefacts.

3.4.2 Why are people not essentialists about artefacts?

Based on our analysis, we speculate that there are two reasons which can possibly explain why people's representations of artefacts do not consist of a hidden essence. First, it is possible that people use the perceptual information to learn about causes but do not attribute causes of the artefact categories to a hidden essence. Possibly, people have an inherent bias which influences their ability to think about causes for artefacts as a function rather than a essence (Bloom, 1996; Ahn, 1998; Rips, 1989). Second, it is possible that although the data and models suggest that people should infer a cause based on perceptual features, people do not use this perceptual information.

In the next two sections we highlight any limitations in our approach and identify areas for future research.

3.4.3 Does the latent cause have to be the 'essence'?

We use two strategies to estimate causal structures. First, we explicitly define a variable which is playing the role of an essence. This dummy variable allows us to compare causal structures in which the essence is causally responsible for perceptual features (common-cause) with structures where the essence is not important (independent, best-fit-independent etc.). Second, we use an approach which learns the causal structure and estimates hidden causes between pairs of variables. This helps us understand if the data suggests that there are pairs of variables which share a common cause and people might be using this information to represent an essence in their conceptual representations.

A possible argument or objection to our methods might be that that these latent variables (or the explicit variable) are not the same as the *essence* which essentialist theories describe. Strevens (2000) argues that a latent cause could also be attributed to a causal law governed by category membership (minimal hypothesis - Strevens, 2000). According to Strevens, essentialism implies a causal structure of the form shown in Figure 3.15 and a minimal causal law of category membership implies a structure shown in Figure 3.16. The causal structures that we are considering do not differentiate between the two.

We look at the existence of latent causal variables and evidence for the common-cause structure as a necessary condition for essentialism but not a sufficient condition. We are not making the claim that existence of latent causal variables is a formal proof for the theory of essentialism. However, an absence of latent causal variables would question the concept of essentialism.

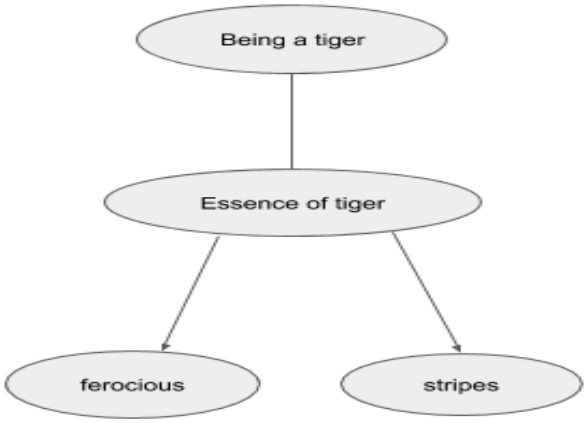


FIGURE 3.15: Essentialism

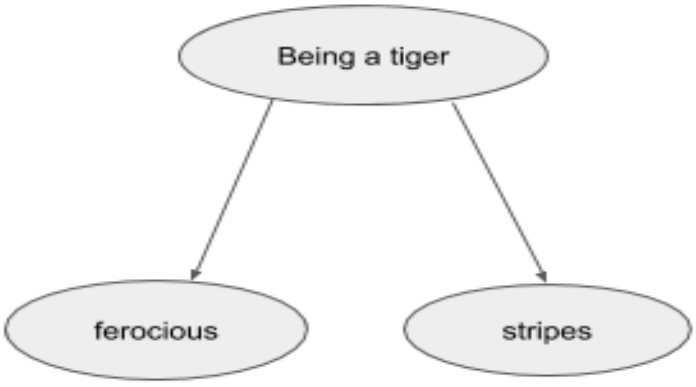


FIGURE 3.16: Minimal Hypothesis

3.4.4 Limitations in our approach

In this thesis, we compare causal structures at the basic-level (for example, dogs, peacocks, etc.). It is possible that essentialism does not exist at this level but exists higher up the category hierarchy at the superordinate level. For example, people might have a single essence in their representation of all objects that fly or all mammals etc. This possibility has not been explored in our study, however, our method can be easily extended to a hierarchical learning approach. However, most developmental evidence for essentialism is based on experiments which have studied differences at the basic-level hence we are not sure if a hierarchical method would actually capture essentialism.

It is possible that using semantic features from property norm studies does not actually capture the causal structure present in people's representations of categories. Possibly abstract features extracted from a Convolutional Neural Network might be closer to conceptual representations. Peterson et al. (2016) demonstrate that the hidden layer of a convolutional neural network can be re-weighted to represent category representations. However, a bayesian structure learning approach might not be the ideal choice for estimating causal structures due to the large number of variables in CNN representations (order of 1000000 variables). Even the best heuristic methods currently can deal with a maximum of 100000 variables assuming at least 1000 data points are available for each category.

A possible limitation in our approach might be that we are limited to 9 categories from each group. However, we have tried to capture a wide range of categories. For the essential categories, we use examples from mammals, amphibians, birds, etc. and for the non-essential categories, we use furniture, music instruments, utilities, etc. However, since our study is limited to only 9 categories from each group, it is possible that these categories have some

underlying systematic patterns and do not span the range of all essential and non-essential categories. It would be interesting to study the causal structures of categories such as plants and fruits. These are categories which do not have a function nor do they have an essence. An absence of latent causal variables would be interesting.

Another possible limitation might be that people are learning causal structures based on a different set of positive and negative examples. We estimate the likelihood of a structure based on a dataset comprising of positive and negative examples. We try two methods of learning the causal structure: first, using negative examples from all categories and second, using a reduced set of relevant negative examples. It is unlikely that people are learning the structure based on an entirely different set of examples, but it might be worth exploring.

Currently, in our study there is no differentiation between negative examples in the data set.[8] We have used a binary variable as a proxy for the ideal essence due to a small sample. For example, in the data for candle, for all the negative examples the dummy variable is set to 0. With a larger data sample, it would be interesting to examine whether these results replicate if the dummy binary variable is replaced with a categorical variable.

Our study relies on the assumption that adults and children are perceiving similar features. One possible argument might be that we have not been able to observe differences between artifacts and animals in our dataset because children perceive a different set of features. Future studies can be designed to capture data from child participants instead of adults.

Part 1 of the thesis allowed us to understand how the perceptual features might be playing a role in the development of essentialist beliefs. Using

[8] Essence variable set to 0

causal models helped us understand that the perceptual environment supports the inference of hidden causes, allowing us to speculate why people might be essentialists. In the second part of the thesis, we study how generic noun phrases vary across categories and discuss their role in the development of essentialist beliefs.

Part 2

Chapter 4

Part 2: Method

The second part of this thesis is concerned with studying the role that Generic Noun Phrases play in the development of essentialist beliefs in children. We examine how the use of generic speech varies across categories and test the hypothesis that adults and children produce a larger number of generic statements for essential categories as compared to non-essential categories. In order to do so, we developed a classifier to automatically identify generic noun-phrases in the CHILDES dataset. In this chapter, we discuss existing methods for the automated identification of generic noun phrases and the system which we developed.

4.1 Existing Work: Automated Identification of Generic Noun Phrases

Generic noun phrases are phrases that do not refer to a specific member of a class but refer to the class in general. For example, the statement, *Elephants take a bath with their long trunk* is a generic statement about the category elephants but the statement, *Charlie, the elephant, uses his trunk to have a bath* is not a generic statement since it refers to a particular individual of a class.

To the best of our knowledge, there have been three attempts to automate the identification of generic noun phrases. Suh (2006) proposed a rule-based approach which uses patterns of part of speech tags to automatically identify generic noun phrases. Reiter and Frank (2010) argue that identification of generic noun phrases should be solved as a classification task rather than a rule-based approach. Reiter and Frank use a Bayesian Network with a feature set comprising of semantic and syntactic features. They test their work on the ACE-2 corpus and show a considerable improvement to the method proposed by Suh (2006). Friedrich and Pinkal (2015) propose a new corpus called WikiGenerics with a new annotation scheme as compared to the ACE corpora. Further, they propose a sequence labelling model (using a conditional random field) for the automated identification of generic noun phrases.

The methods proposed by Reiter and Frank (2010) and Freidrich and Pinkal (2015) use a combination of syntactic and semantic features (including the wordnet lexical category of the noun). It is possible that the datasets which have been used have a larger number of generic statements from certain categories. This would imply that the classifiers might be biased towards certain categories. Since our goal was study how generic noun phrases vary across categories, it was important for us to ensure that the classifier we used did not depend on semantic features and was not biased towards certain categories.

Therefore, we developed a sequence based classifier, using Long Short Term Memory Networks (LSTMs) which achieves a similar performance as compared to the method proposed by Friedrich and Pinkal (2015) using a minimal feature set comprising of only syntactic features. The next section introduces recurrent neural networks and LSTMs, following which we discuss the architecture of the model that we used.

4.2 RNNs, LSTMs and GRUs

Recurrent Neural Networks (RNNs) are a class of neural networks for processing sequential data (X_1,X_2,....X_t). Unlike a feed-forward neural network where the output depends only on the current input, the output in an RNN depends on the current input as well as the previous output. An RNN can be thought of as neural network with a loop, which allows for information to remain in the sequence. Figure 4.1 shows a Recurrent Neural Network which receives an input X_t at time-step t and outputs a value h_t. The loop shows that the output h_t will be available to the network at the next time step, i.e. $t + 1$. This architecture serves the purpose of memory in the network, allowing sequential information to be preserved in the RNN's hidden states. The input data to an RNN does not necessarily have to be in time-steps like time series data, and could be any sequential data, for example, a sentence or a paragraph.

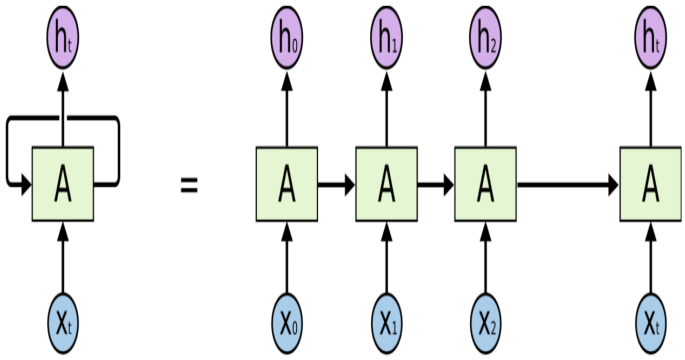


FIGURE 4.1: Recurrent Neural Network
Unfolding a recurrent neural network

In theory, Recurrent neural networks are supposed to be able to pick out correlations and dependencies across time steps in sequences. However, Bengio

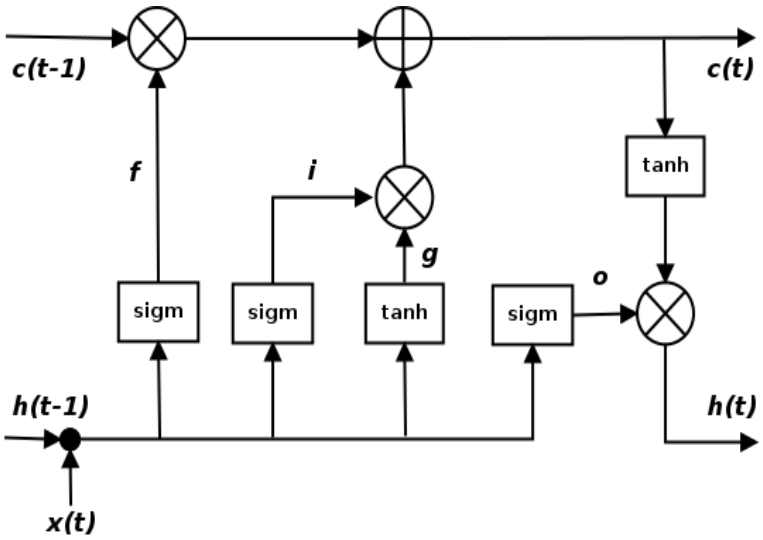


FIGURE 4.2: LSTM
Architecture of a Long Short Term Memory Network

et al. (1994) and Hochreiter (1991) show that in practice recurrent neural networks suffer from the problem of the vanishing gradient[1] and are unable to pick out long term dependencies, i.e. dependencies which are spread out over a large time interval. A Long Short-Term Memory Network (LSTM) is a recurrent network architecture which is capable of capturing these long-term dependencies.

LSTMs maintain an additional state, called the cell state C_t, which is available to the network at a time-step t, apart from the input X_t and the output from the previous hidden unit, i.e h_{t-1}. LSTMs have the ability to add and remove information from this cell state using structures called gates. Figure 4.2 shows the architecture of an LSTM at time-step t. f represents information being forgotten from the cell state using a gate. i and g represent new information being added to the cell state.

The Gated Recurrent Unit (GRU) is a variant of the LSTM in which the forget and input gates are merged into an update gate. The GRU does not maintain a separate cell state, but rather merges the hidden state and the cell state. Figure 4.3 shows the architecture of a GRU cell.

[1]The gradient becomes very small which effectively prevents learning in the network

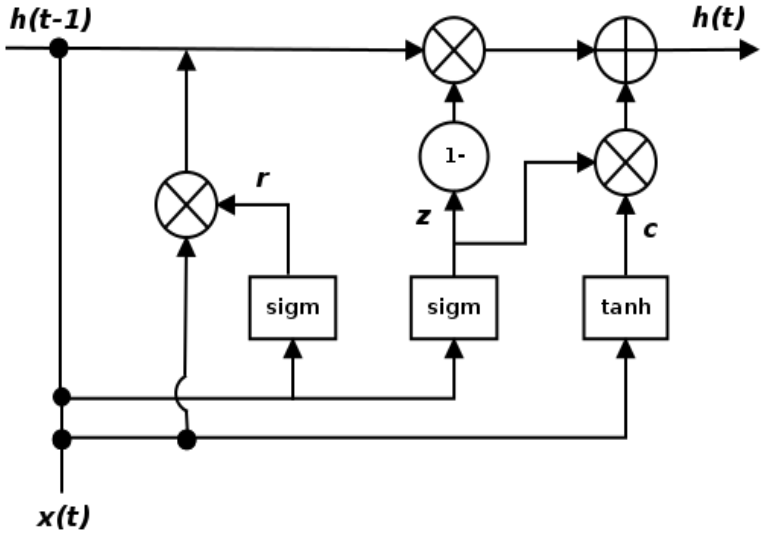


FIGURE 4.3: GRU
Architecture of a Gated Recurrent Unit

To develop a classifier for the identification of a generic noun phrase in a sentence, we use a combination of LSTMs and GRUs. The architecture of the model is discussed in the next section.

4.3 Automated Identification of Generic Noun Phrases

We use a sequence based approach to classify if a sentence consists of a generic noun phrase or not. For each step of the sequence (the sequence here is the sequence of words - the sentence), the input vector X_t is a concatenation of two one-hot encoded vectors.[2] These are:

- The Part of Speech (POS) Tag of the word [3]
- The Dependency label associated with the word in the Dependency Parse Tree. (Johansson et al., 2008)

[2] A One-Hot encoded vector is way of representing a categorical variable as a binary vector. For example, a variable X which can take 3 values, can be represented as [0,0,1]; [0,1,0] or [1,0,0] depending on the value that X takes.

[3] We use the universal part of speech tag-set which consists of 16 possible POS-tags. (Petrov et al., 2011)

Word	Pos-Tag	Dependency
Elephants	noun	nsubj
do	verb	aux
not	adv	neg
eat	verb	ROOT
birds	noun	dobj
.	punct	punct

TABLE 4.1: Features

For example, for the sentence, *Elephants do not eat birds.*, the POS Tags for each of the words and the corresponding dependency label would be as shown in Table 4.1:

The input vector X_t corresponding to each word is a concatenation of the two labels after one-hot encoding each label. This concatenated vector was a vector of length 64. The LSTM cell is combined with a fully connected neural network layer (Dense) for a classification task. Further, multiple LSTM and GRU cells can be stacked together. This is the equivalent of stacking multiple hidden layers in a feed forward network.

Our model consists of seven different architectures of networks which were trained independently. The decision on generic/non-generic was made by taking the majority vote across classifiers. The architectures of the classifiers were:

- $input(64) \rightarrow GRU(90) \rightarrow GRU(60) \rightarrow GRU(30) \rightarrow Dense(2) \rightarrow output$
- $input(64) \rightarrow GRU(45) \rightarrow GRU(45) \rightarrow GRU(45) \rightarrow Dense(2) \rightarrow output$
- $input(64) \rightarrow LSTM(30) \rightarrow GRU(45) \rightarrow GRU(45) \rightarrow GRU(45) \rightarrow GRU(30) \rightarrow Dense(2) \rightarrow output$
- $input(64) \rightarrow GRU(90) \rightarrow LSTM(90) \rightarrow Dense(2) \rightarrow output$

- $input(64) \rightarrow GRU(30) \rightarrow GRU(40) \rightarrow GRU(50) \rightarrow GRU(60) \rightarrow GRU(70) \rightarrow Dense(2) \rightarrow output$
- $input(64) \rightarrow LSTM(196) \rightarrow GRU(196) \rightarrow GRU(196) \rightarrow GRU(196) \rightarrow LSTM(196) \rightarrow Dense(2) \rightarrow output$
- $input(64) \rightarrow GRU(396) \rightarrow GRU(192) \rightarrow GRU(98) \rightarrow Dense(2) \rightarrow output$

The number in brackets denotes the dimensionality of the output of the layer. In each of the models, the activation function of the dense layer was the softmax function. Each model was trained with a weighted categorical cross entropy loss function[4] where the error associated with classifying a non-generic statement as generic (false positive) was weighed 1.5 times more than the error associated with classifying a generic statement as non-generic (false negative).

We tested our model on the WikiGenerics dataset (Friedrich and Pinkal, 2015). The WikiGenerics dataset consists of 102 classes of documents (each document consists of generic and non-generic statements). Friedrich and Pinkal (2015) test their model by using a leave-one-document-set-out cross validation strategy. In each cross validation step, examples from 101 document sets are used for training and the model is tested on the left out document set. We followed the same strategy. Table 4.2 shows how our majority voting classifier performed in comparison to the existing methods, i.e. Reiter and Frank (2010) and Friedrich and Pinkal (2015).[5]

We also implemented a variant of our model with an extended feature set comprising of syntactic and semantic features. We added the following semantic features for every word:

[4]The 'Adam' optimizer was used for optimization (Kingma and Ba, 2014).

[5]These numbers are as reported by Friedrich and Pinkal's paper (2015). We did not reimplement their models.

Model-Name	Accuracy	F-Measure
Bayes-Net (Reiter and Frank, 2010)	71.7	72.3
Conditional Random Field (Friedrich and Pinkal, 2015)	79.1	78.8
Majority Voting Classifier (syntactic features)	76.4	79.3

TABLE 4.2: Model Performance

- Word net lexical category of the word
- word2vec representation of the word

We find that on adding the semantic features, the accuracy of the model increases to 78.2 and the F-Measure increases to 80.5. However, since this increase may be a result of a bias towards certain categories in the dataset which might impact the classifiers performance on child-directed speech, we use our syntactic model which achieves a comparable performance to existing state-of-the-art methods.

We used our model to analyse corpora from the CHILDES data set to study how generic speech varies across age groups and categories. The analysis is discussed at length in the next chapter of the thesis.

Chapter 5

Part 2: Analysis

In the previous chapter, we implemented a classifier for the detection of generic noun-phrases. We use this model to understand how the use of generic noun phrases varies across categories and age groups in child speech and child-directed speech. In this chapter, the words generic and generic noun-phrases have been used interchangeably.

5.1 Data

We collected data from 28 studies from the English (UK) and English (North America) corpora available on CHILDES. To ensure that our dataset comprises of natural conversations that occur between children and adults, we restricted ourselves to 28 studies which were naturalistic and did not include studies in which children were being asked a specific set of questions or being told to play with a restricted set of toys. The names of the studies have been listed in Appendix C.

The number of utterances by children for various age categories are shown in Table 5.1. Due to a small number of utterances for the age groups 5-6, 6-7 and over 7, while performing any age-wise analysis, we merge the three into

Age-Category (in years)	Number of Utterances
Less Than 2	117551
2-3	567749
3-4	264232
4-5	212423
5-6	11279
6-7	1127
Over 7	997

TABLE 5.1: Number of utterances: Child Speech

Age-Category of Child (in years)	Number of Utterances
Less Than 2	220156
2-3	787690
3-4	352240
4-5	224508
5-6	7877
6-7	1258
Over 7	1576

TABLE 5.2: Number of Utterances: Adult Speech (as a function of the age of the child)

a single age-group (Over-5). We do not consider child speech lesser than the age of 2 in any of our analysis.[1]

Similarly, we extracted all the speech directed towards children. Table 5.2 shows the division of the number of utterances by adults for various age categories, i.e. the age of the child the speech is directed towards.

In the next section, we study how the use of generic phrases varies across categories and age groups.

[1]Most of the speech is limited to one or two words at a time, making it very difficult to predict genericity.

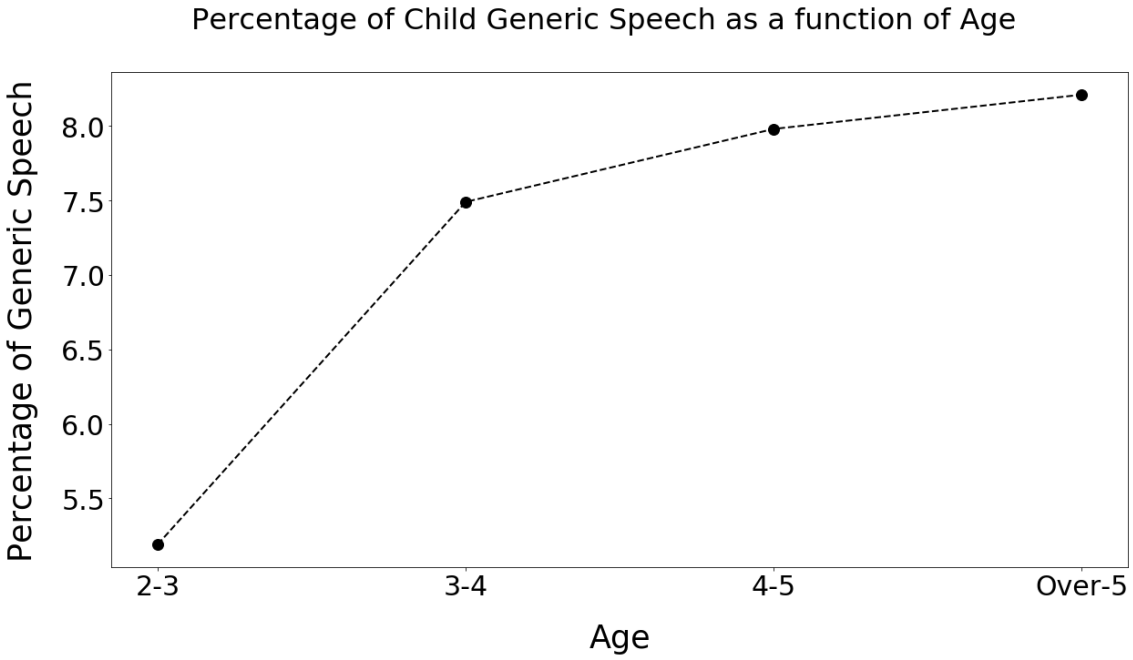


FIGURE 5.1: Percentage of Generics in Child Speech
The figure above shows the percentage of child speech which is generic for each age category.

5.2 Analysis

Genericity across age groups

We first examined the percentage of child speech which is generic as a function of the child's age. The results shown in Figure 5.1 are consistent with the findings of Gelman et al. (2008), which showed that the percentage of utterances which are generic increases with age. However, we find a higher percentage of generics as compared to the study conducted by Gelman et al.(2008). Figure 5.2 shows the percentage of adult speech which is generic as a function of the age of the child that is being addressed. Again, the results are consistent with the study conducted by Gelman et al. (2008), which showed that percentage of generic utterances increases till 3 years of age and then decreases.

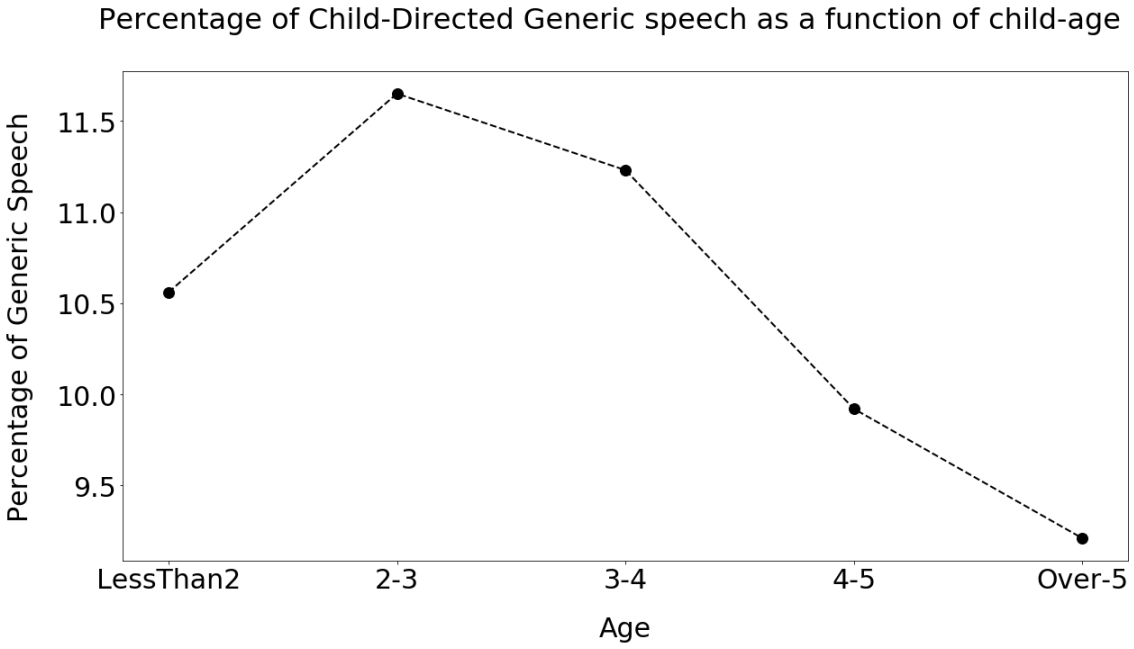


FIGURE 5.2: Percentage of Generics in Adult Speech
The figure above shows the percentage of child-directed (adult) speech which is generic as a function of the age of the child.

Category	Generic
Animal	snakes don't really have heads;I dunno I think raccoons might eat birds eggs
Artifact	I mean helicopters are often used
Person	Cowboys have big faces
Plant	it's called peel the skin the skin on fruit is called peel
Object	where's the shiny stars
Food	collard greens is good for you

TABLE 5.3: Examples of Generic Speech

Genericity across categories

Generic noun phrases were found from the following lexical categories[2]: animals, artifacts, plants, food, natural objects and person. Examples of Generic and Non-Generic sentences from each category are shown in Tables 5.3 and 5.4.

Figure 5.3 shows the division of generic and non-generic speech in absolute

[2]wordnet lexical categories: https://wordnet.princeton.edu/documentation/lexnames5wn

Category	Non-Generic
Animal	how about I trace this lizard
Artifact	tell her what happened to your toy,where's the top of this basket
Person	poor Aunt Dot she has nobody to love;I know a little girl who lost her finger
Plant	that's a very nice tree that's a big tree
Object	this star is already touching this star
Food	how about Skippy peanut butter

TABLE 5.4: Examples of Non-Generic Speech

numbers across categories in adult speech. We find that the majority of generics in adult speech are from the artifact category. However, non-generic statements follow a similar distribution across categories. This indicates that the higher proportion of artefact generics in our dataset is probably because the number of utterances about artefacts was higher than the number of utterances for other categories, rather than a tendency of adults to produce generic statements about artefacts. We also find that child speech follows a similar pattern for both generic and non-generic speech (Figure 5.4).

To get a better sense of generic usage across categories in child and child-directed speech, we compared the percentage of generic statements in each category, i.e. the number of generic statements from a particular category/Total number of statements which belong to the category (Figure 5.5).

We find that in both adult and child speech, natural objects is the category with the highest percentage of generic statements. While children tend to produce a higher percentage of generic statements for the animal category as compared to the artifact category, the differences are small (just over 1%). Adults on the other hand produce a slightly higher percentage of generic statements for the artifact category as compared to the animal category.

Figures C.1 to C.8 (Appendix C) highlight the variation in generic speech across categories (by percentage of genericity within a category) across age groups. The results across age groups are similar: by percentage of genericity, natural objects are spoken about most often in a generic sense, and the

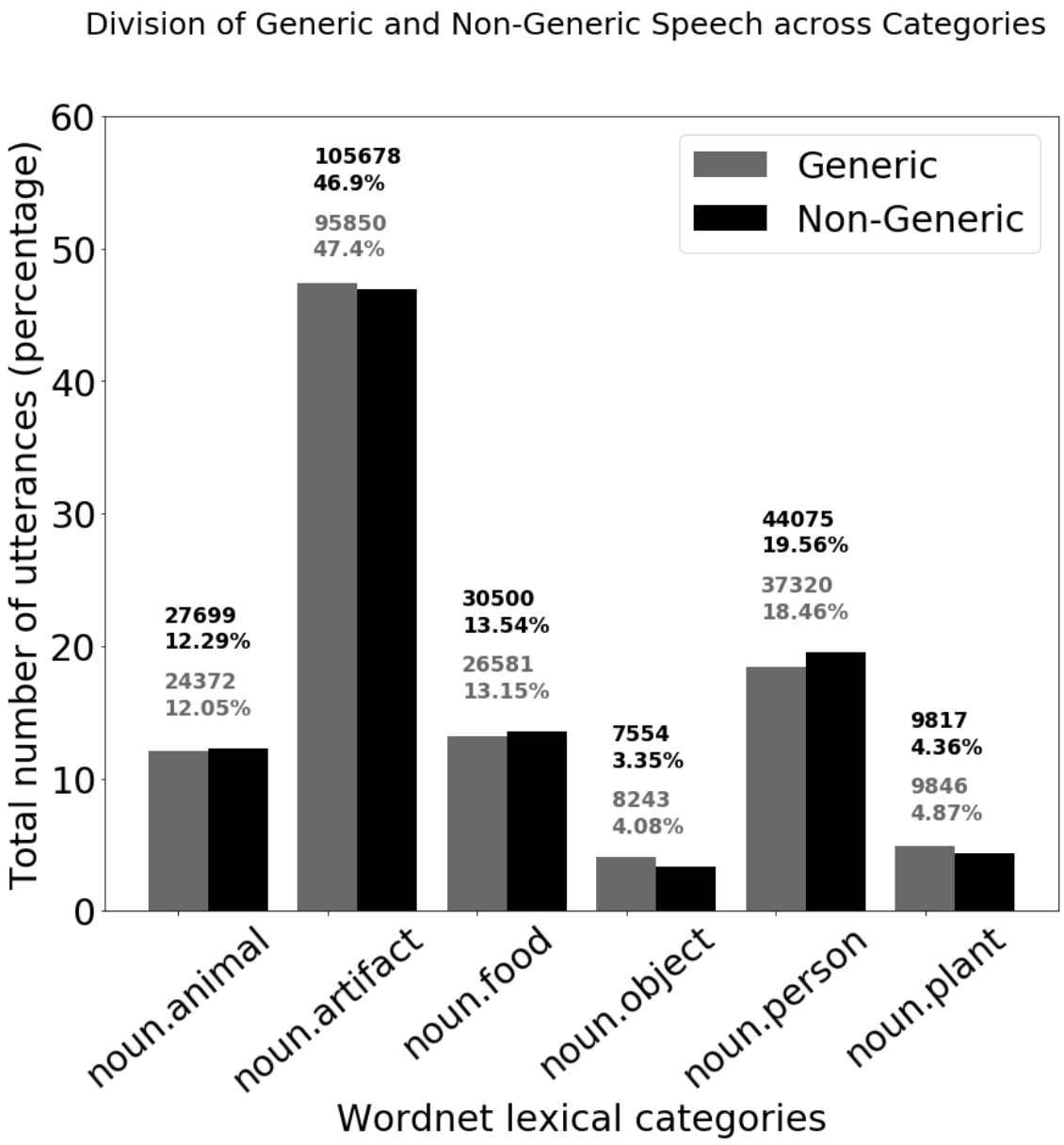


FIGURE 5.3: Division of Generics and Non-Generics in Adult Speech

The figure above shows the division of the raw counts of generic speech and non-generic speech across categories in child directed speech. Note that non-generic speech which does not consist of nouns has not been included in our analysis. We find that by division of raw counts, noun.artifact is the lexical category with the maximum number of generic and non-generic statements. Total number of statements have been reported above each bar (grey for generic speech, black for non-generic speech)

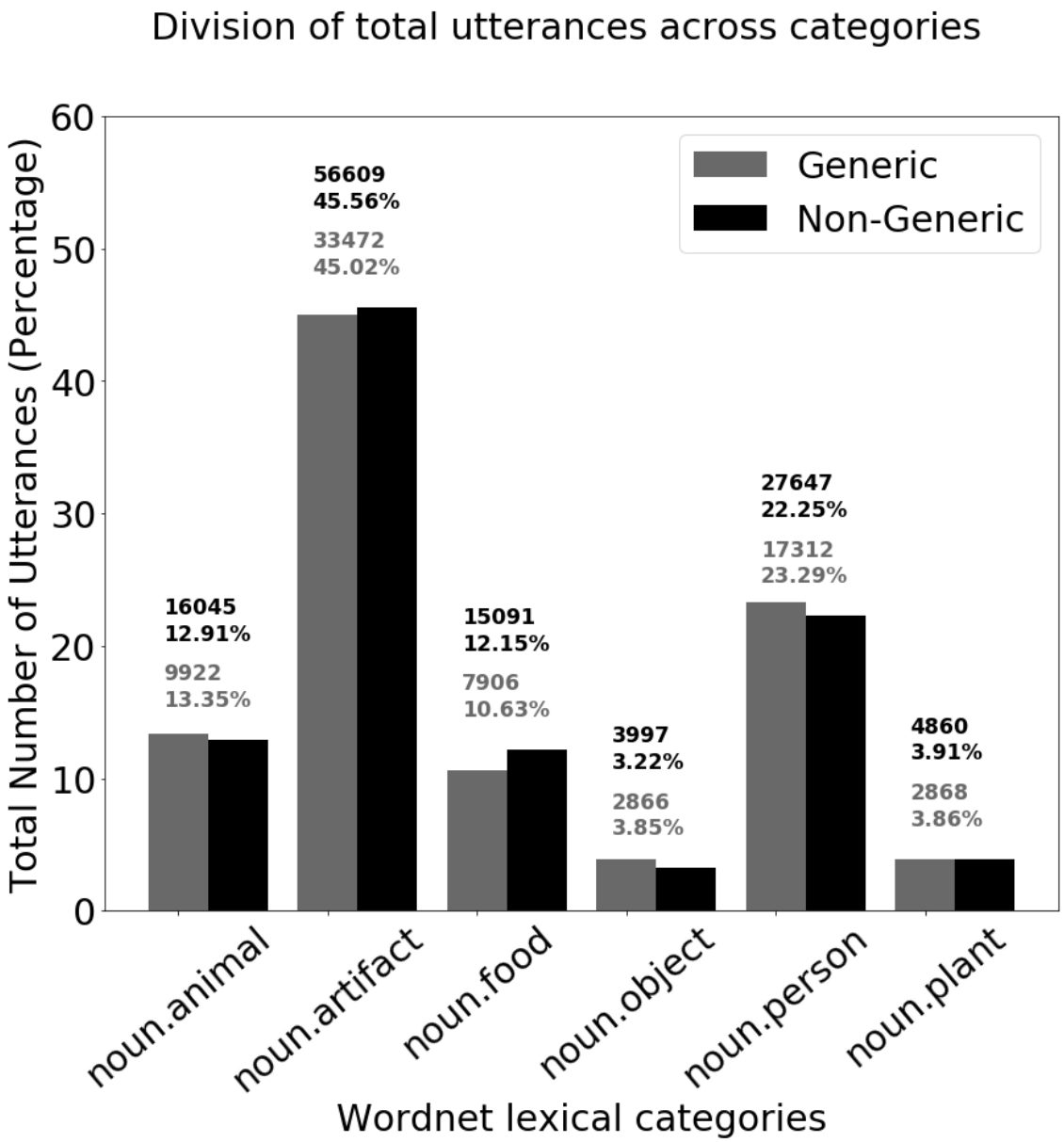


FIGURE 5.4: Division of Non-Generics in Child Speech
The figure above shows the division of the raw counts of generic speech and non-generic speech across categories in child speech. By division of raw counts, noun.artifact is the lexical category with the maximum number of generic and non-generic statements. Total number of statements have been reported above each bar (grey for generic speech and black for non-generic speech)

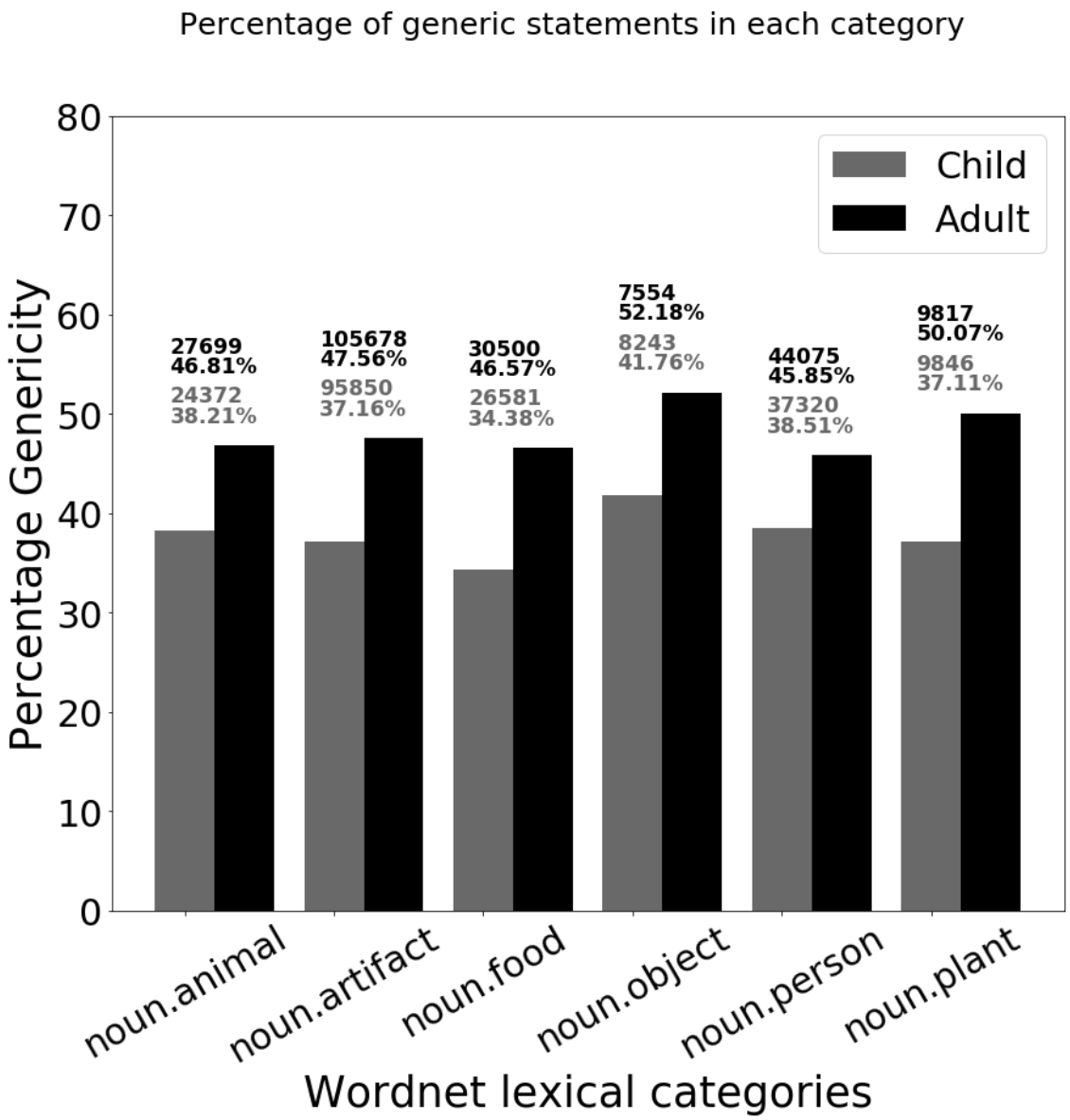


FIGURE 5.5: Genericity in each category: Child and Child-Directed Speech

Each bar represents the percentage of generic statements within a category. Total number of generic statements have been reported above each bar (grey for child speech and black for adult speech).

differences between animal and artifact genericity (by percentage) are small.

We find that our results are not consistent with results presented by Gelman et al. (2008). Their study suggested that both children and adults produce a higher number of generic statements for animal categories as compared to artifact categories. The study conducted by Gelman et al. (2008) indicates that over 40% of generics (by division of raw counts across categories) are from the animal category and approximately 25% are from the artifact category, in both child and adult speech. Our results, in Figure 5.3 and Figure 5.4, indicate that approximately 13% of the generics are from the animal category and approximately 45% of the generics are from the artefact category. Gelman et al. (2008) do not report the percentage genericity of each category. However, they argue that a higher percentage of generic speech by raw counts for the animal category is not because people speak more about animals. Our results on the other hand suggest that both generic and non-generic statements follow a similar distribution, indicating that percentage genericity is a better metric. Our results suggest that the difference in percentage of genericity of the animal category as compared to the artefact category is small. In fact, by percentage of genericity, adults produce more generic statements from artefact categories as compared to animal categories, indicating that generic noun phrases might not be playing a role in the development of essentialist beliefs in children.

To get a better sense of the variation of generic statements across categories, we extracted the top 20 nouns which occur in generic statements. Table 5.3 shows the top 20 nouns in adult speech and Table 5.4 shows the top 20 nouns in child speech.

We find that by both, percentage of genericty as well as by raw counts, there are words from artefact as well as animal categories in the top 20 nouns. It is interesting to note that children produce generic noun-phrases about nouns

Category	Generic (Raw Counts)	Non-Generic (Raw Counts)	Generic (by percentage)
Artifact	car, house, train, floor, truck, toy	house, car, room, train, truck, bag, ball	floor, road, train, sock, chair,truck, bed, toy
Person	baby, man, boy, girl	baby, man, boy, friend, tea	Kid, man, child
Animal	cat, bird, dog, egg	cat, dog, bird	bear, bird, elephant
Plant	flower, tree	flower	grass, tree
Object			sun, sky
Food	apple, biscuit, milk, breakfast	apple, milk, breakfast, tea	apple, biscuit

TABLE 5.5: Top 20 nouns in generic and non-generic speech (adult speech)

Columns 1 and 2 are based on raw frequency of the words. However, since words with a high generic frequency are probably in the list because of a high total frequency, we calculate the top 20 words by percentage of generic use, i.e. frequency of the noun when used in a generic sense/total frequency of the word. We include words which have a frequency of at least 100 in the entire corpus.

Category	Generic (Raw Counts)	Non-Generic (Raw Counts)	Generic (by percentage)
Artifact	car, house, room, floor, ball, engine, pant	car, house, room, train, truck, boat, bus, bag	floor, road, bed, toy
Person	baby, boy, girl, friend	Man, baby, girl	Kid, *monsters*, mother, *doctor*
Animal	dog, cat, cow, lion	cat, dog, horse, bear, bird	*dinosaur*, snake, bird, cat
Plant	flower, tree	flower	tree
Object			sun, *star*
Food	apple, egg, chocolate	apple, milk, cookie,	apple, biscuit

TABLE 5.6: Top 20 nouns in generic and non-generic speech (child speech)

Top 20 nouns by raw frequency as well as by percentage of generic use.

which do not necessarily have a high percentage of genericity in adult speech (example: doctor, monsters, dinosaurs etc.). These results follow a similar pattern when we consider the top 50 and top 100 nouns as well.

5.3 Discussion

5.3.1 Summary

The analysis conducted above, suggests the following:

- Natural Objects is the category with the highest percentage of generic statements.
- Differences in use of generic noun phrases between artifact and animal categories are small.
- Use of generic statements in child speech is not limited to those categories which have a high percentage of generic use in child-directed speech.

We discuss where our results differ from the original study conducted by Gelman et al. (2008) and how our results fit into the discussions about the development of essentialism.

5.3.2 Differences with Gelman et al. (2008)

Earlier studies (Gelman et al., 2008; Gelman and Tardif, 1998) have indicated that generic speech in child speech and child-directed speech is content specific. A higher number of generics were found to be from the animal categories and it has been argued that these results are not because people are speaking more about animals in general. The study conducted by Gelman et

al. (2008) does not report a metric of percentage genericity, and claims about animal bias in child-directed speech are based on division of raw counts across categories. We find that by division of raw counts, generic speech and non-generic speech follow a similar pattern, and by percentage genericity the differences across the categories are very small.

Why do our results differ? We speculate that there are two possible reasons:

One possible argument might be that the classifier is not performing well, which in turn affects our analysis. However, performance of the classifier is unlikely to account for the differences in the two studies. Currently the classifier has an F-Measure of 79.3, which is better than (or similar in performance to) existing state of the art methods. Gelman et al. (2008) claim that there is a difference in the order of 15-20% when generics about animals are compared with artefact generics.[3] Our results suggest a difference in the order of close to 1%. Even if the classifier is not capturing all the generic statements or making a few errors, it should still capture the patterns of difference, if the differences actually exist. Further, since we are using a classifier which does not depend on semantic features, these errors would occur in both animal and artefact categories, and hence although the actual count of number of generics might be over or under estimated, the patterns of genericity should not be affected.

Another possible reason for the difference in results is that the corpora used by Gelman et al. (2008) is biased towards certain categories. This is possible considering that their study uses data from only eight children. By substantially increasing the size of the dataset, we reduce the chances of biases in the dataset towards certain categories. In the future, we plan to extend our

[3]Note that Gelman et al. (2008) do not report percentage genericity and only report division based on raw counts. Hence we assume that as per their claims, even percentage genericity would follow a similar pattern of distribution of raw counts.

analysis to capture differences in each corpora by studying how generic output produced by a child varies with input. This analysis would allow us to examine if particular corpora have a content specific bias.

5.3.3 What do our results mean for Essentialism?

The fact that natural objects are spoken about in a generic sense, can explain why people develop essentialist beliefs about such categories. However, the lack of differences in generic use between animals and artefacts suggests that generic noun phrases are not the only reason why people develop essentialist beliefs. Developmental studies have shown that children develop essentialist beliefs even in the absence of generic noun phrases (Gelman, 2003). Similarly, our results indicate that the differences in patterns of generic speech between essentialised and non-essentialised categories are small. This potentially means that generic noun phrases are not a source of essentialist beliefs. Possibly people have an inherent bias for animal categories which is strengthened on hearing generic noun phrases for natural kinds.

In the next chapter, we summarize our findings from both the parts and discuss future directions of research.

Chapter 6

Discussion

The goal of this thesis was to study why people have essentialist beliefs. In this chapter we summarize our analysis, discuss the implications of our results, limitations of our study and directions for future research.

6.1 Summary

In Chapter 1 we discuss that the representation of an essentialised category is structured, coherent, has an inductive potential and consists of an underlying structure of latent causal features which are stable over transformations. In this thesis, we tried to explain why representations of categories consist of two of these properties. We study why conceptual representations consist of an essence which is a latent causal feature and why people consider certain categories to have an inductive potential.

Hidden, Causal Variable

Essentialism theories suggest that people's representations of categories consist of a hidden variable, i.e. the essence, which is causally responsible for various surface features (Gelman, 2003). In the first part of this thesis, we

tried to explain why representations of categories might consist of a hidden causal variable, by examining whether people might be inferring hidden causes because of the perceptual features of a category. We find that for both essential and non-essential categories, perceptual features serve as cues for the inference of hidden causal variables. This potentially explains why people develop essentialist beliefs for animal categories based on the hidden causes inferred but it does not explain why people do not develop essentialist beliefs for artefact categories.

Inductive potential

In the second part of the thesis, we studied the role that generic noun phrases might play in the development of essentialist beliefs. It has been suggested that the use of generic noun phrases for a category implies that the category is coherent and has an inductive potential. Previous studies (Gelman and Tardif (1998) and Gelman et al. (2008)) have shown that generics in child-directed speech are content specific and parents produce a higher number of generics when talking about animals as compared to artefacts. We extend the study conducted by Gelman et al. (2008) to a larger data set and examined whether generics used in child and child-directed speech are content specific. We find that both adults and children produce a higher percentage of generics when they are talking about natural objects. Therefore, we might be able to explain why people have essentialist beliefs for natural kinds such as stars and beaches. However, we find that the differences in generic speech between animal and artefact categories are small. Hence, we are unable to explain if generic speech is actually responsible for the development of essentialist beliefs.

What are the implications of our results? Our results suggest that the environment offers a variety of perceptual and linguistic input for people to develop essentialist beliefs. However, our results also open up the question

of why do people essentialize differently for animals and artefacts in spite of similar input - generic noun phrases and latent variables exist for animal and artefact categories.

6.2 Limitations and Future Directions

Our study indicates that the environment (particularly, the perceptual features) might be the source of essentialist beliefs in people. We suggest that future work should be aimed at analysing the reason for the difference in essentialist beliefs for natural and artefact kinds. In our approach of trying to explain why are people essentialists, we have followed a strategy of explaining two different aspects of essentialism - causality of surface features and inductive potential of a category. Neither of our studies can fully explain why people's representations of categories consist of these properties for animal categories but not for artefact categories. We have tried to separate linguistic and perceptual input and tried to explain how each of these might be contributing to essentialism. We expected to find differences in both the studies for essential and non-essential categories, which in turn would explain the acquisition of different aspects of essentialism. However, based on the results we suggest that future studies should try to combine linguistic and perceptual input while trying to answer why people are essentialists. It is possible that differences between animal and artefact kinds are visible when a combination of linguistic and perceptual input is available.

6.3 Conclusion

We have tried to answer the question of why are people essentialists. By developing a framework of using causal graphical models to explain the causal

structures in representations of features, we try to explain if the perceptual features are responsible for the inference of a hidden causal essence. We also studied if generic noun phrases might be responsible for people believing that certain categories have an inductive potential. We are able to explain certain aspects of representation of essential categories but are unable to explain why the same properties do not develop for artefacts and non-essential categories. Our results potentially indicate that people have an inherent bias which differentiates between natural and artefact kinds. Further, the environment offers perceptual input which might strengthen this difference. Our study however, is often limited due to a small data sample and we have outlined future directions of research which can be undertaken with a larger data set. By developing a framework of combining property norm studies with causal graphical models, we open up new directions of research for answering questions which existing models of conceptual representation cannot answer.

Appendix A

Additional details of the experiment

A.1 List of Categories and Features

- LION: can roar, has a mane, has a tail, has fur, has teeth, is ferocious, is large, lives in jungles, is yellow, lives in Africa
- GORILLA : likes to eats bananas, can swing from trees, is black, is dangerous, has fur/hair, is large, is strong, beats its chest
- FLAMINGO : can fly, can stand on one leg, has a beak, has a long neck, has feathers, has long legs, is pink, lives in water
- PEACOCK : has feathers, is colourful, has long tail feathers, is blue, is noisy/loud, is beautiful, has a beak, is proud
- GOLDFISH : can swim, has fins, has scales, is orange, is small, lives in aquariums, lives in water, lives in a bowl
- ANT : has legs, can bite, can crawl, is black, is red, is small, is strong, lives in colonies

- SNAIL : has a shell, is slow, is slimy, is found in gardens, has eyes, leaves a trail, has an antenna, is edible
- IGUANA : is green, is scaly, has a tail, has a tongue, has legs, can eat insects, is big/large, lives in hot climates
- PIG : has a snout, is pink, has a curly tail, likes mud, is found on farms, is fat, oinks, can be eaten as bacon
- BALLOON : can float, made of rubber, is colourful, requires helium, can burst, is for parties, is round, requires air
- FLUTE : made of metal, is long, has holes, used by blowing air through, is silver, produces music, used in orchestras, is thin
- NECKLACE : is worn around the neck, made of gold, made of pearls, made of silver, is expensive, has a clasp, has a pendant, is for females
- MICROWAVE : is found in kitchens, can cook food, can heat, made of metal, has a door/doors, is electric, is rectangular/square, is fast
- TAXI : is yellow, is black, made of metal, has a meter, used for transportation, is expensive, used for passengers, has a sign
- CANDLE : has a wick, made of wax, provides light, produces heat, melts, is scented, different shapes, is romantic
- BUCKET : has a handle/handles, made of metal, made of plastic, can contain liquid, used for holding things,is circular/round, is found on beaches, is used for gardening
- UMBRELLA : protects from the rain, has a handle/handles, is watertight/waterproof, is collapsible, keeps things dry, made of fabric/cloth material, is carried, has spokes

- DESK : made of wood, has legs, has drawers, is flat , found in offices , made of metal , is strong ,used for working on

A.2 Category Typicality

Category	Overall	High	Medium	Low
bucket	0.83	0.93	0.70	0.49
flute	0.83	0.92	0.69	0.53
microwave	0.84	0.94	0.70	0.50
candle	0.85	0.97	0.69	0.49
desk	0.81	0.90	0.70	0.53
peacock	0.87	0.98	0.71	0.48
iguana	0.83	0.94	0.69	0.47
gorilla	0.85	0.96	0.70	0.48
pig	0.84	0.94	0.72	0.53
balloon	0.87	0.98	0.72	0.49
flamingo	0.87	0.97	0.72	0.51
ant	0.83	0.94	0.67	0.49
taxi	0.85	0.95	0.70	0.50
necklace	0.85	0.95	0.73	0.51
snail	0.84	0.95	0.69	0.51
goldfish	0.87	0.98	0.71	0.51
lion	0.84	0.95	0.69	0.52
umbrella	0.85	0.95	0.70	0.51

TABLE A.1: Average Category typicality ratings
The output of the softmax layer of a pre-trained neural network has been used as a proxy for category typicality/representativeness of an image. High denotes images from the range 0.8-1 of category typicality. Medium and Low represent images from the range 0.6-0.8 and 0.4-0.6 respectively.

Appendix B

Additional Details: Part 1

Additional analysis of the data using different models. We present similar results to our original analysis when we use the top 3 features, a reduced data set and 3 Bins instead of Binary variables.

B.1 Top 3 Features

We repeated the analysis using the top 3 features, instead of the top 4 features. The results follow a similar pattern as when the top 4 features are used. Figures B.1 and B.2 show that the common cause model is better supported by the data as compared to the independent and common-effect structure.

Figure B.3 shows the performance of the best-fit graph as compared to the common-cause graph and Figure B.4 shows the out-degree of the dummy variable in the best-fit graph.

B.2 Reduced Data Set

We constructed a reduced data-set where negative examples were from the appropriate group, i.e. data for animal categories consists of positive examples of the particular animal and negative examples of only animals. To

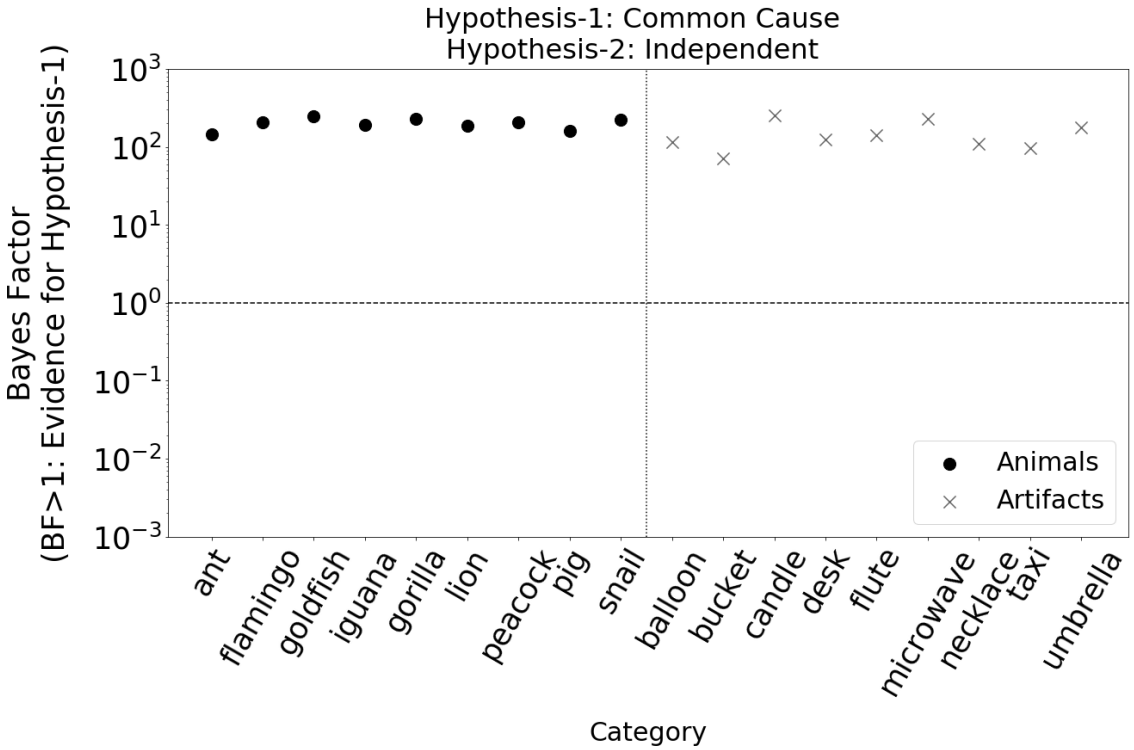


FIGURE B.1: Bayes Factor (Common Cause/Independent) - Top 3 features

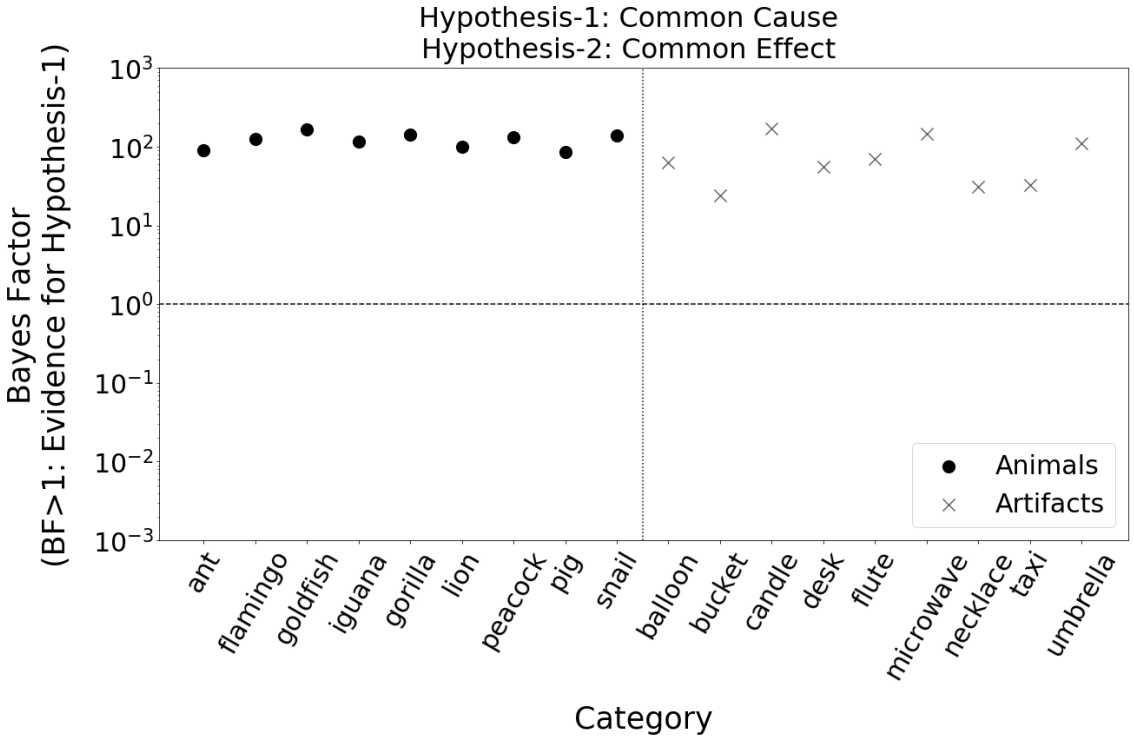


FIGURE B.2: Bayes Factor (Common Cause/Common Effect) - Top 3 features

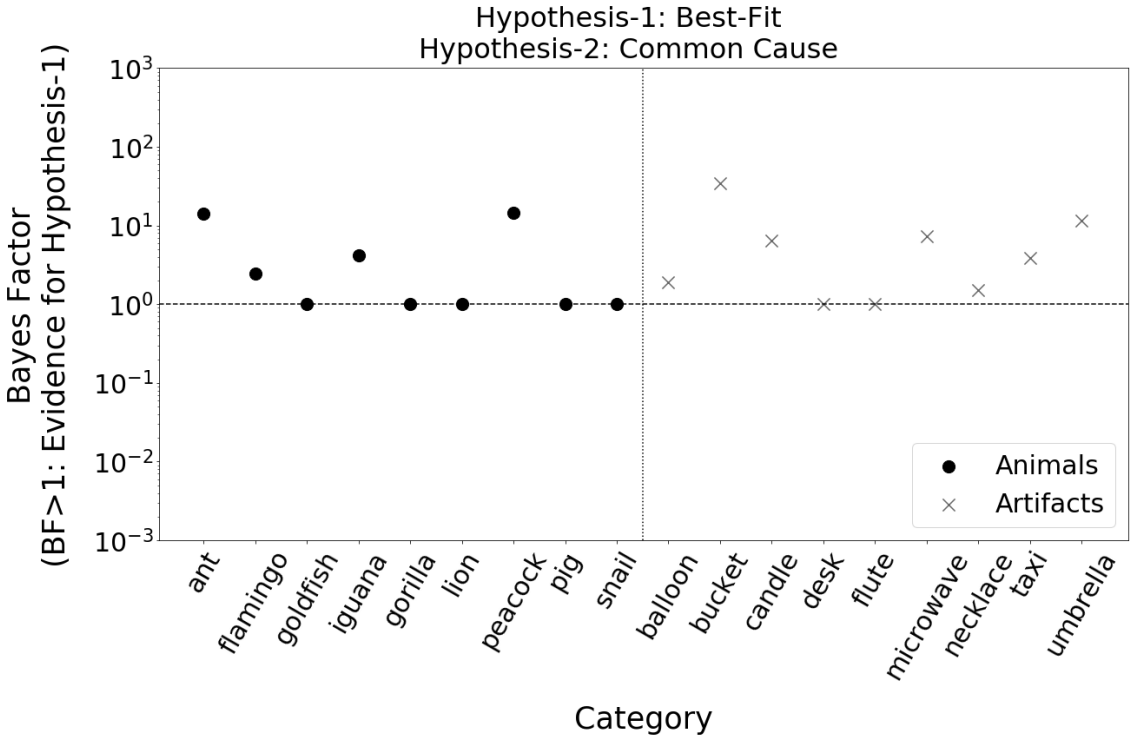


FIGURE B.3: Bayes Factor (Best-Fit/Common Cause) - Top 3 features

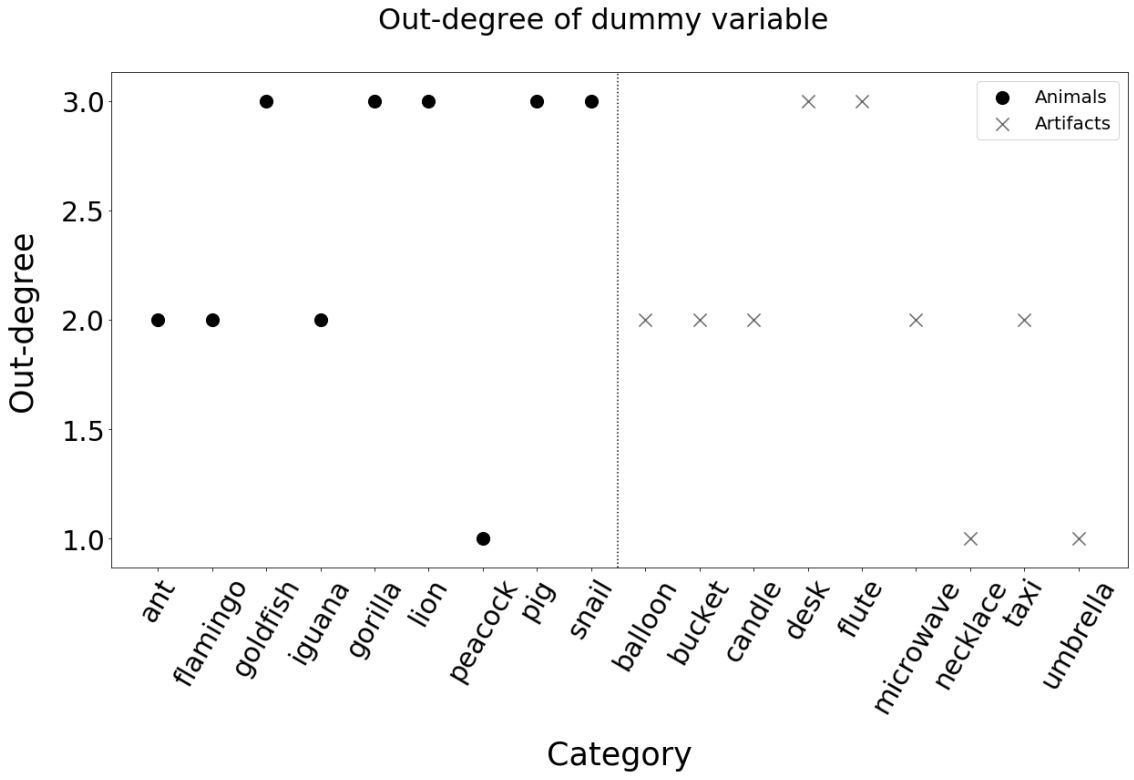


FIGURE B.4: Out-degree of Dummy Variable - Top 3 features

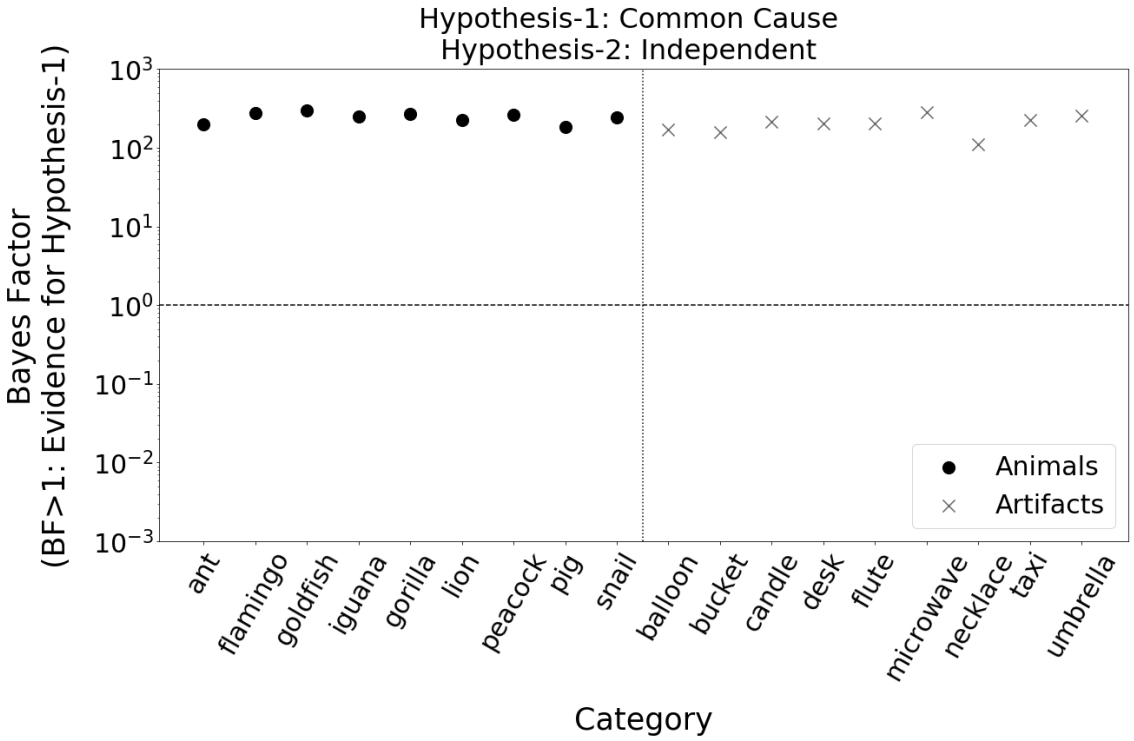


FIGURE B.5: Bayes Factor (Common Cause/Independent) - Reduced Data set

ensure that our data was not skewed towards positive examples (since we are reducing the negative examples), we reduce the positive examples from 120 to 80 using random sampling.

We used all 8 features in our analysis. We find that the common-cause structure is strongly supported by the data as compared to the independent model, which is consistent with our initial analysis. (Figure B.5)

We also find that the common cause model is not much worse that the best-fit model (Figure B.6). The best-fit model was learnt using a 3 parent node restriction instead of the 4 parent node restriction because the size of our data set is smaller. The common-cause model also fits the data better than the best-fit-independent model (Figure B.7). Both these results are consistent with our original analysis.

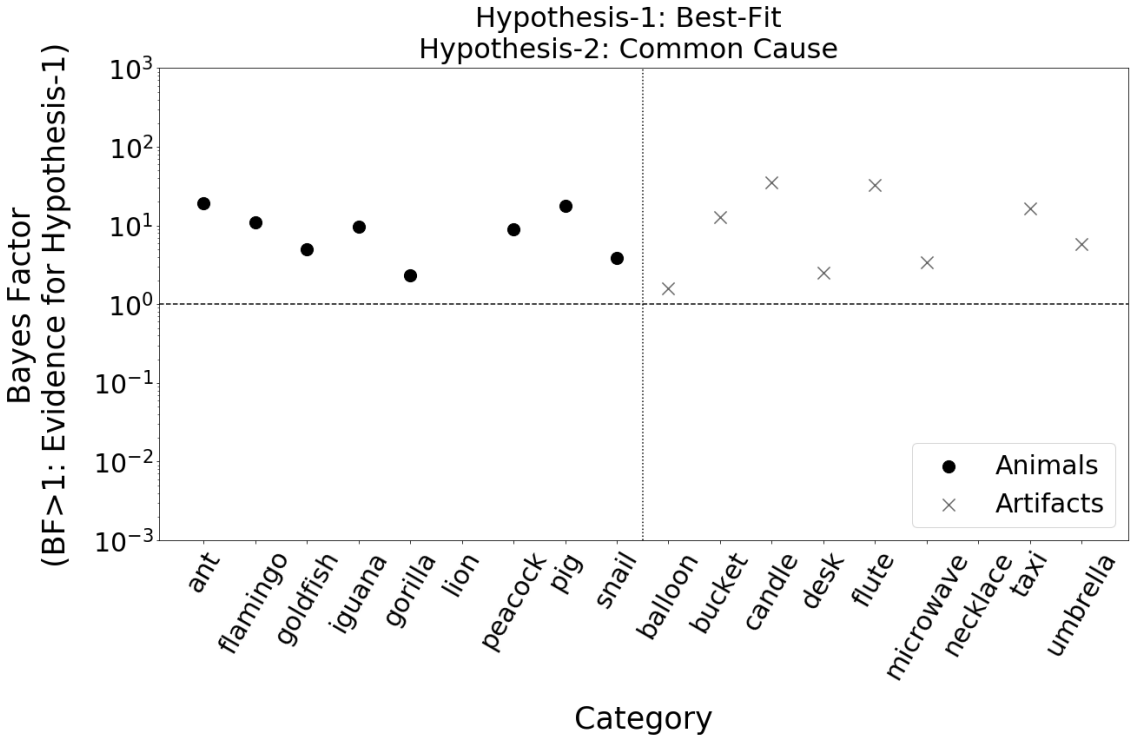


FIGURE B.6: Bayes Factor (Best Fit/Common Cause) - Reduced Data set

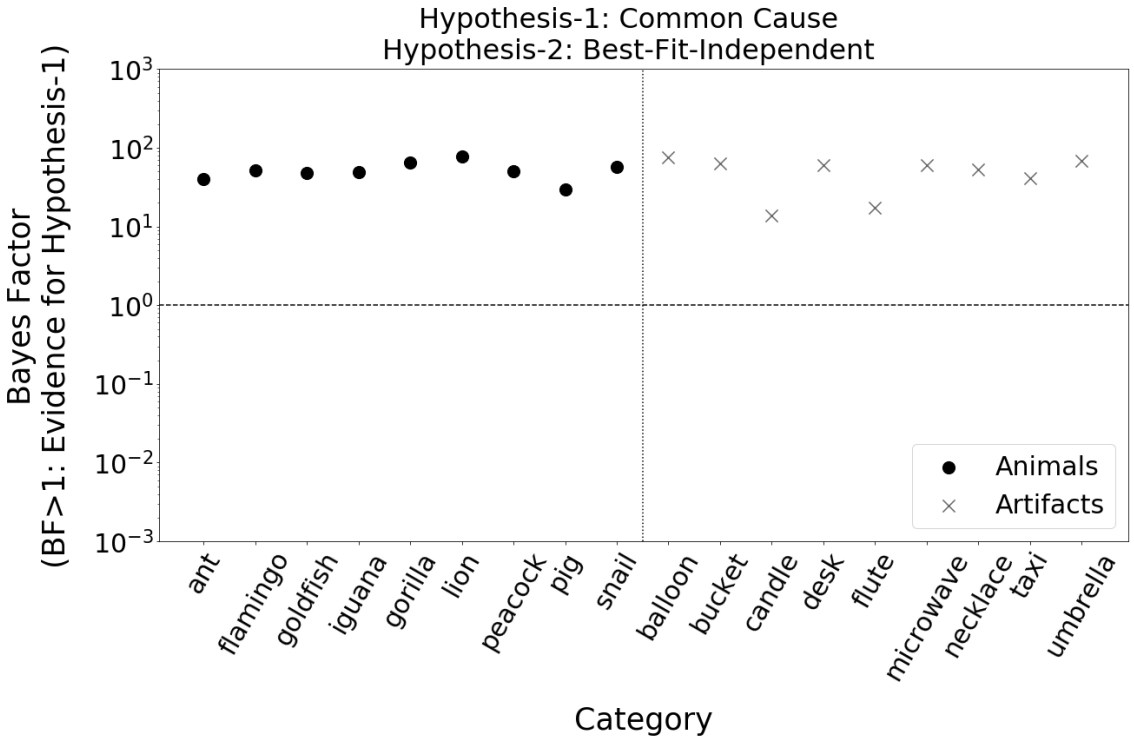


FIGURE B.7: Bayes Factor (Common Cause/Best-Fit-Independent) - Reduced Data set

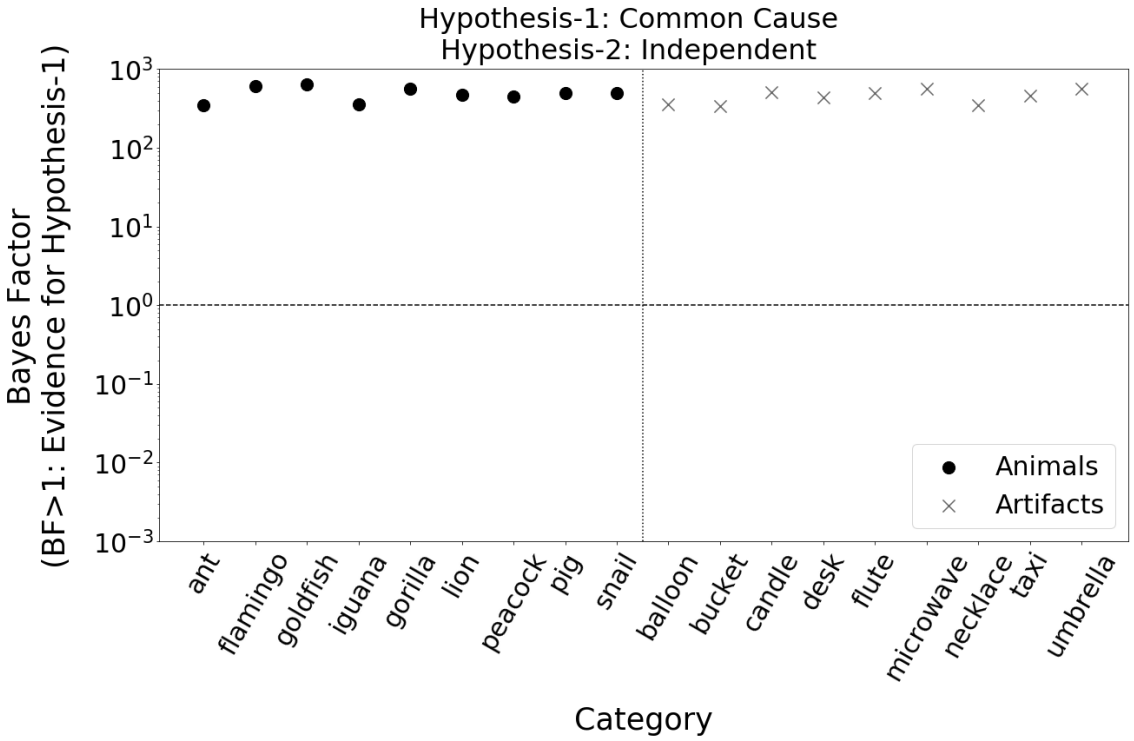


FIGURE B.8: Bayes Factor (Common-Cause/Independent) - 3 Bins

B.3 3 Bins

In Chapter 3 our analysis uses binary variables. However, to ensure that using 2 bins (binary variables) was not losing information, we repeated a part of the analysis using 3 bins. We find that the results are consistent with our original analysis. The common cause model is better supported than the independent model (Figure B.8) and the best-fit-independent model (Figure B.10). Further, the common cause model is not much worse than the best-fit model (Figure B.9). The best-fit-independent and best-fit models were learnt using a three parent node restriction. All 8 features were used.

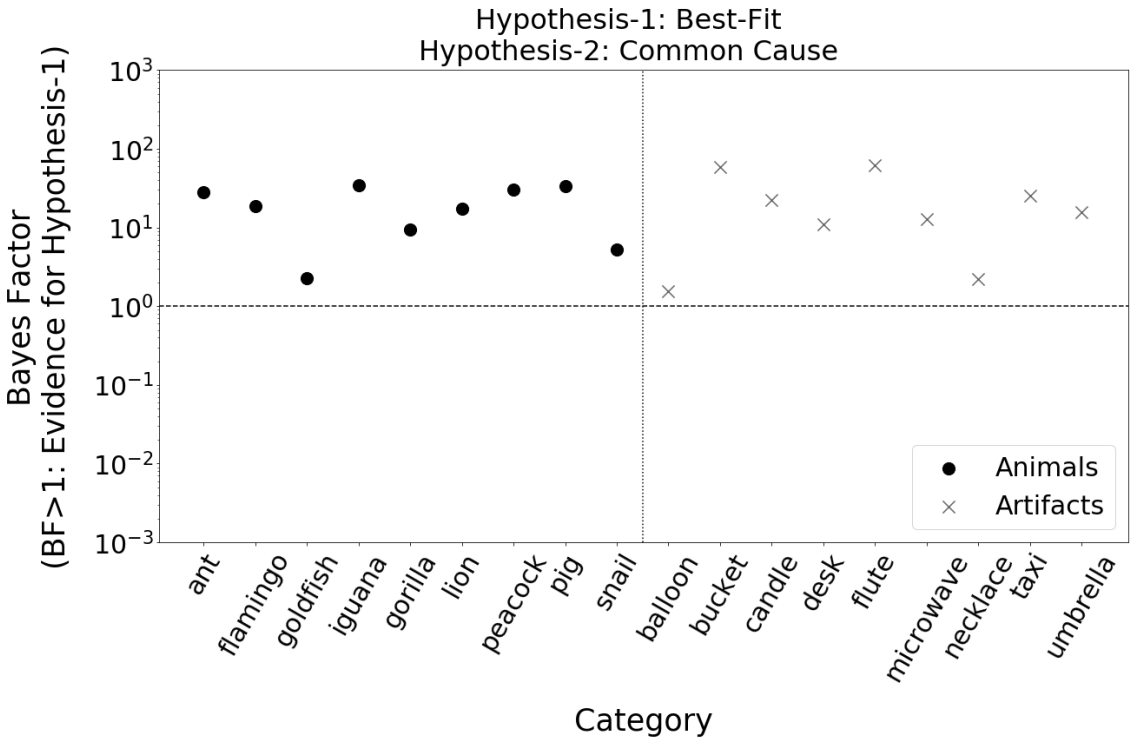


FIGURE B.9: Bayes Factor (Best Fit/Common Cause)- 3 Bins

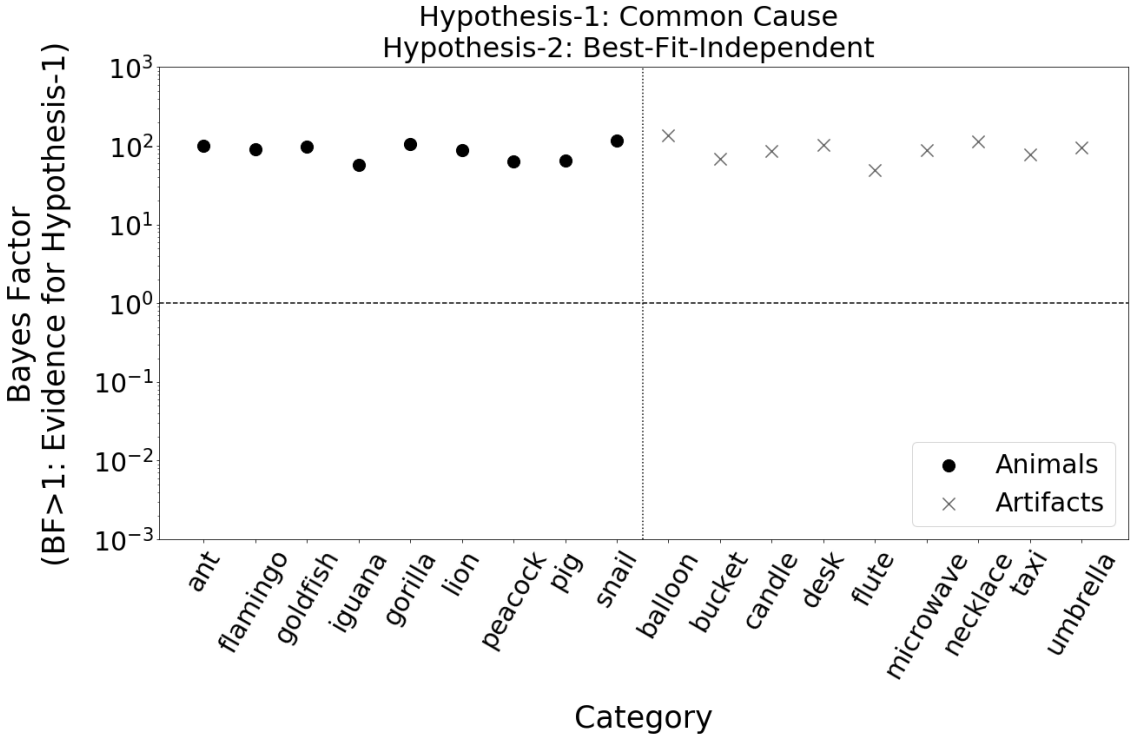


FIGURE B.10: Bayes Factor (Common Cause/Best-Fit-Independent)- 3 Bins

B.4 Marginal Likelihood

In Chapter 3, we compare the common-cause model with the best-fit-independent model by comparing their likelihoods. We also calculated the ratio of the marginal likelihood of the two graphs by marginalizing the dummy variable.

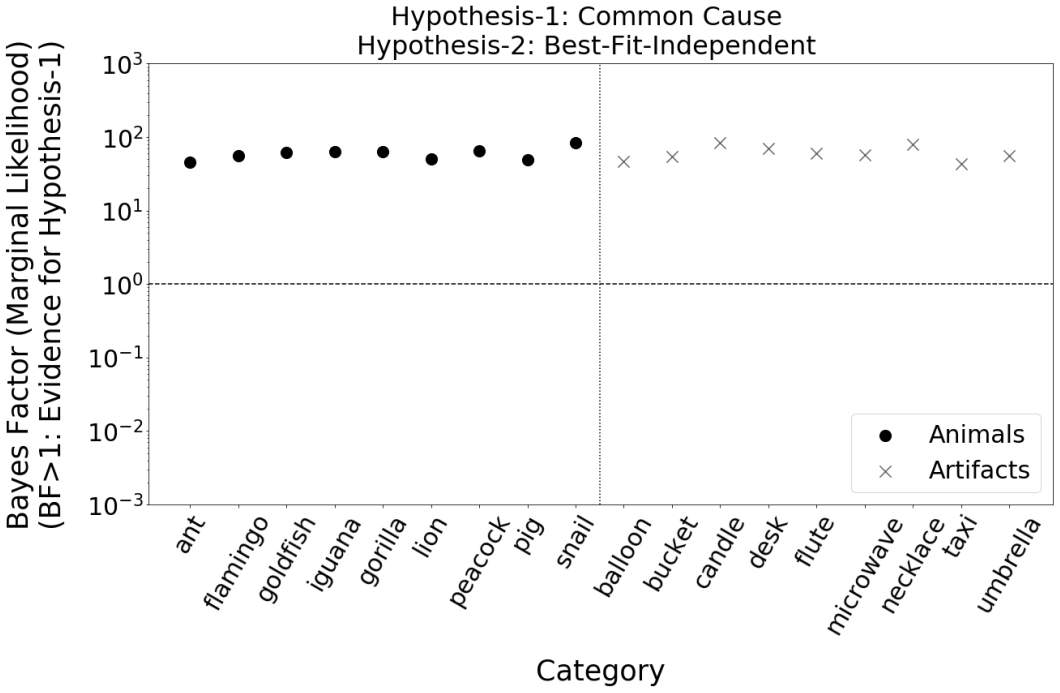


FIGURE B.11: Ratio of Marginal Likelihood (Common Cause/Best-Fit-Independent)

B.5 Latent Causes

We used the RFCI algorithm to calculate the directed acyclic graph in the presence of hidden variables. Pairs of features which share a latent cause have been shown in Tables B.1 and B.2. Results indicate that there are feature pairs in both essential and non-essential categories which share a latent cause.

category	features
ant	can.crawl,is.small,10
	can.crawl,is.strong,11
	has.legs,is.strong,6
	has.legs,lives.in.colonies,3
	is.black,is.red,8
	is.black,is.small,10
	is.black,lives.in.colonies,14
	is.small,lives.in.colonies,5
flamingo	
goldfish	can.swim,has.scales,6
iguana	is.scaly,lives.in.hot.climates,15
gorilla	is.black,is.dangerous,15
lion	has.teeth,lives.in.jungles,6
peacock	
pig	is.found.on.farms,likes.mud,3
snail	has.a.shell,leaves.a.trail,3
	is.edible,is.slimy,12
	is.slimy,leaves.a.trail,4

TABLE B.1: Feature pairs which share a latent cause - Animals
The number along with each feature pair is the number of causal structures (out of 20) in which the pair shared a common cause.

category	features
balloon	can.burst,requires.helium,11
	made.of.rubber,requires.helium,13
bucket	has.a.handle.handles,made.of.plastic,9
	is.circular.round,is.found.on.beaches,6
	is.circular.round,used.for.holding.things,12
	is.found.on.beaches,is.used.for.gardening,6
	made.of.metal,made.of.plastic,13
	made.of.plastic,used.for.holding.things,15
candle	is.romantic,is.scented,10
	is.romantic,provides.light,6
desk	has.legs,is.strong,5
	is.strong,used.for.working.on,5
flute	has.holes,is.long,20
microwave	has.a.door.doors,is.rectangular.square,14
	is.found.in.kitchens,is.rectangular.square,5
necklace	is.expensive,made.of.gold,10
taxi	
umbrella	has.a.handle.handles,keeps.things.dry,15
	is.carried,is.watertight.waterproof,8

TABLE B.2: Feature pairs which share a latent cause - Artefacts
The number along with each feature pair is the number of causal structures (out of 20) in which the pair shared a common cause.

B.6 Correlation between features

Figures B.11 to B.28 show the individual heat maps of correlation between features for each category.

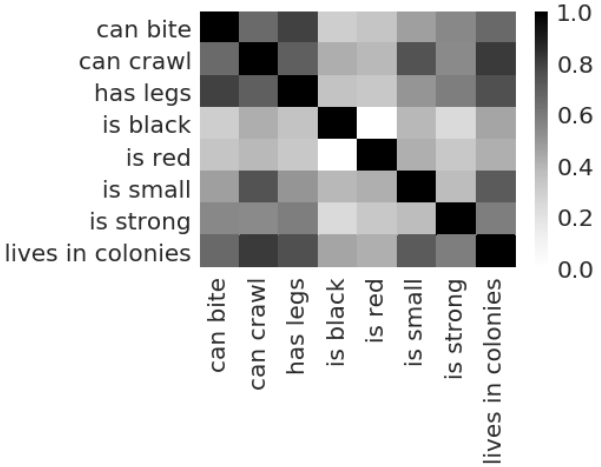


FIGURE B.12: Correlation - ANT

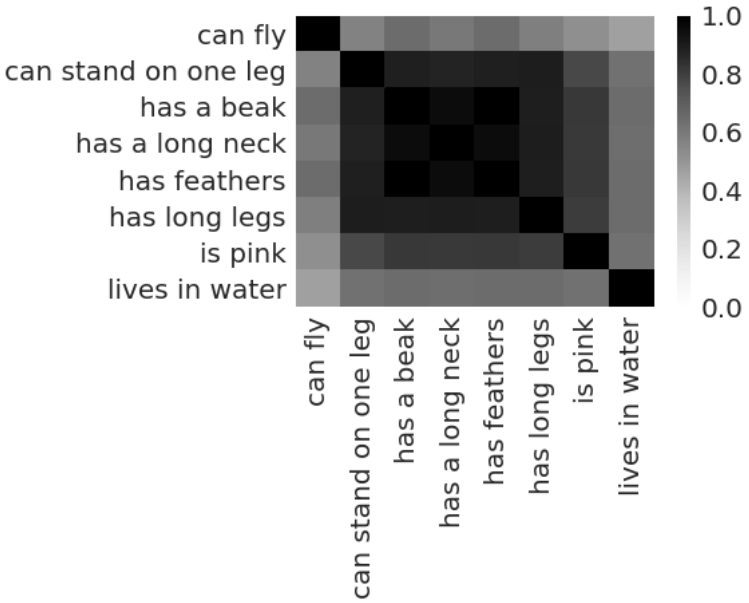


FIGURE B.13: Correlation - FLAMINGO

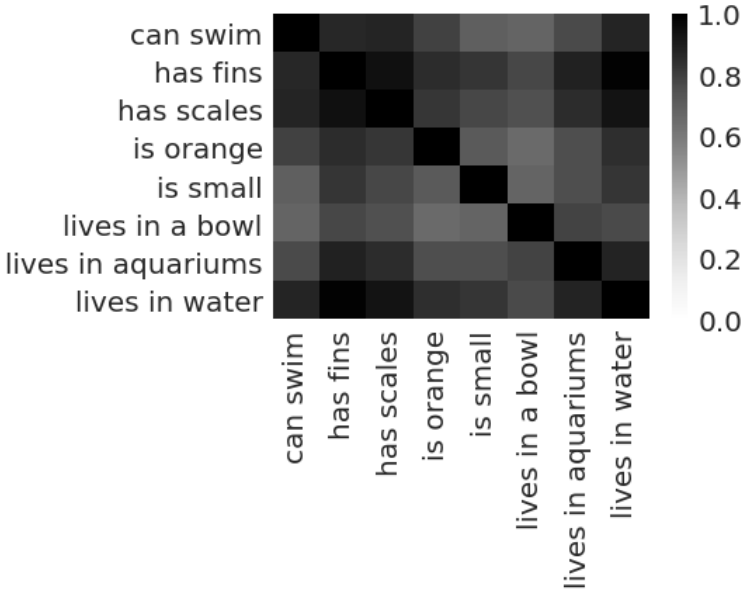


FIGURE B.14: Correlation - GOLDFISH

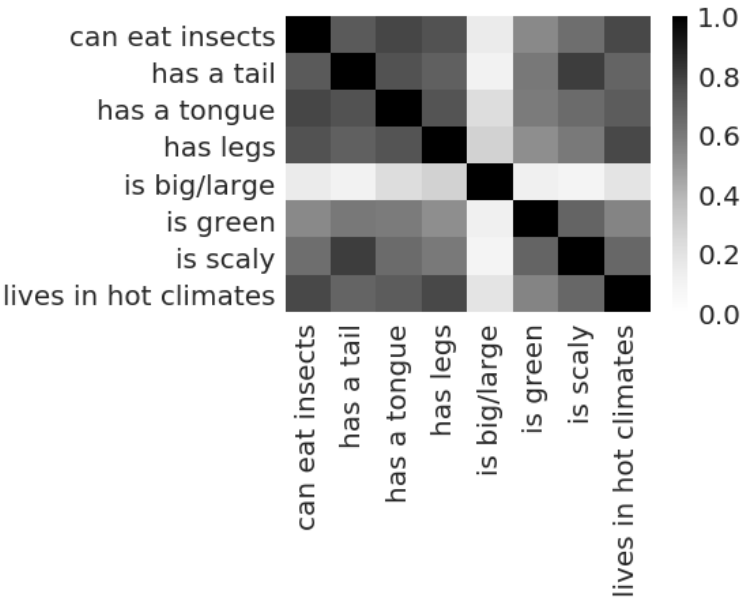


FIGURE B.15: Correlation - IGUANA

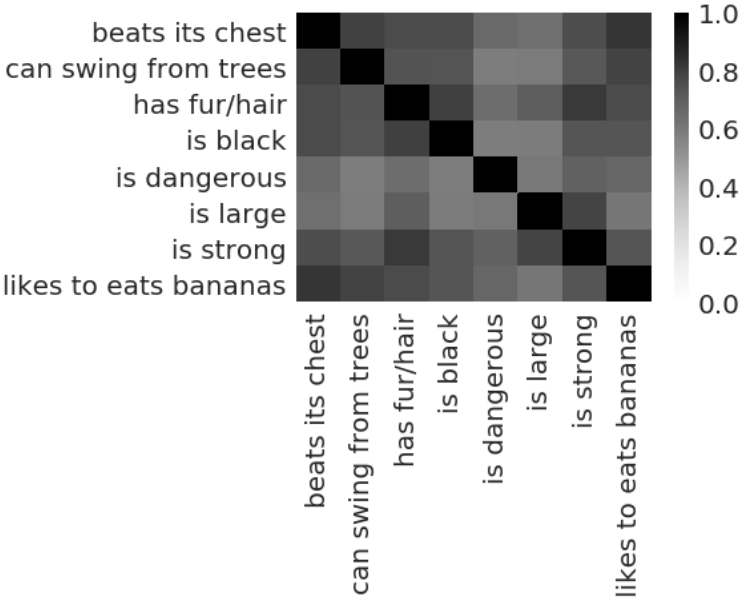


FIGURE B.16: Correlation - GORILLA

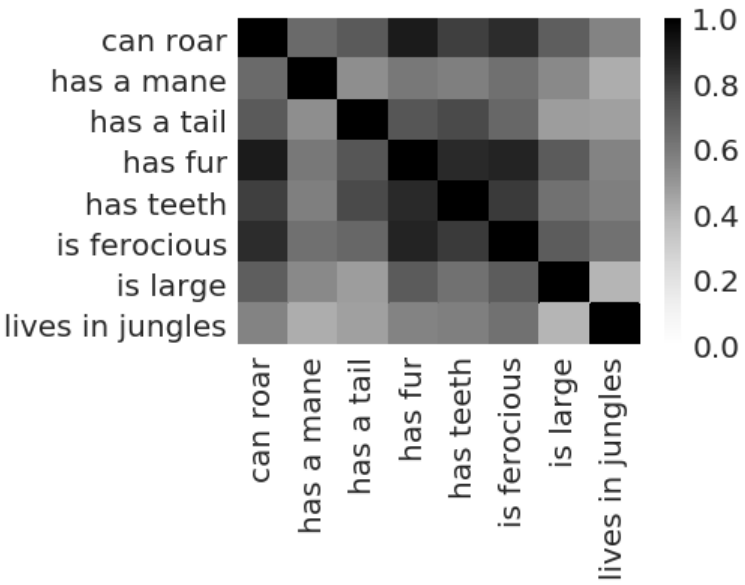


FIGURE B.17: Correlation - LION

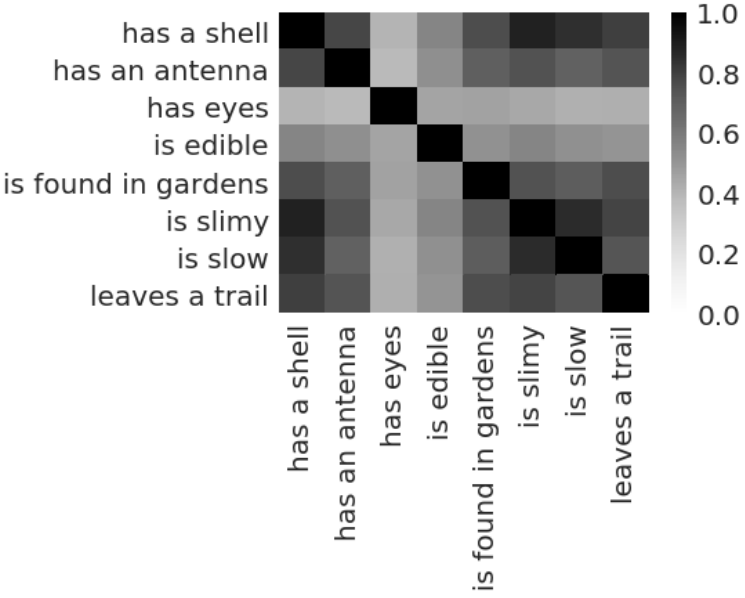


FIGURE B.18: Correlation - SNAIL

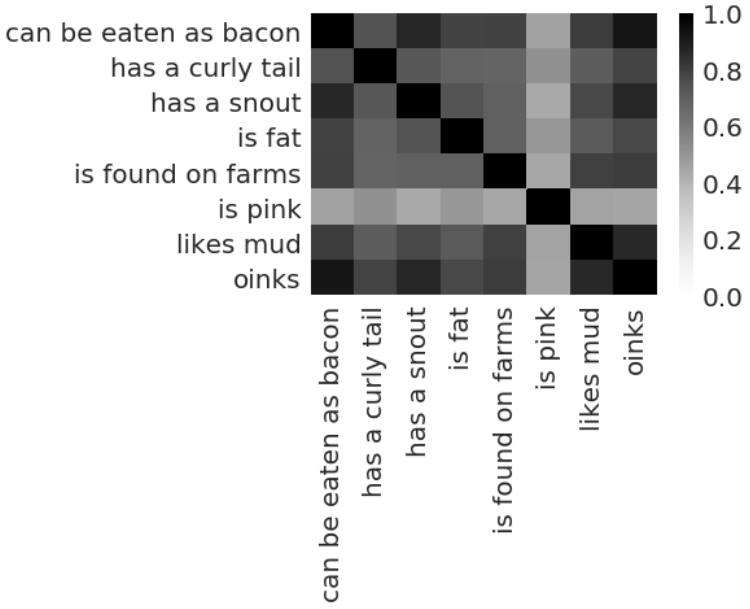


FIGURE B.19: Correlation - PIG

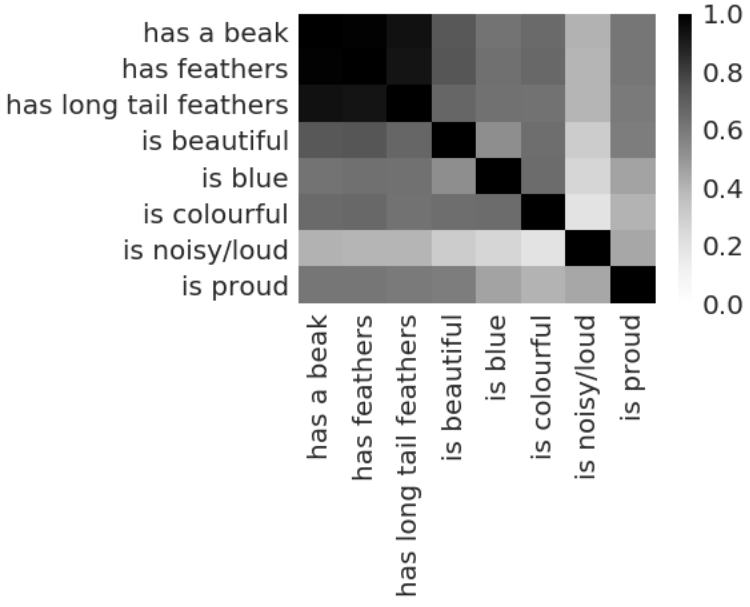


FIGURE B.20: Correlation - PEACOCK

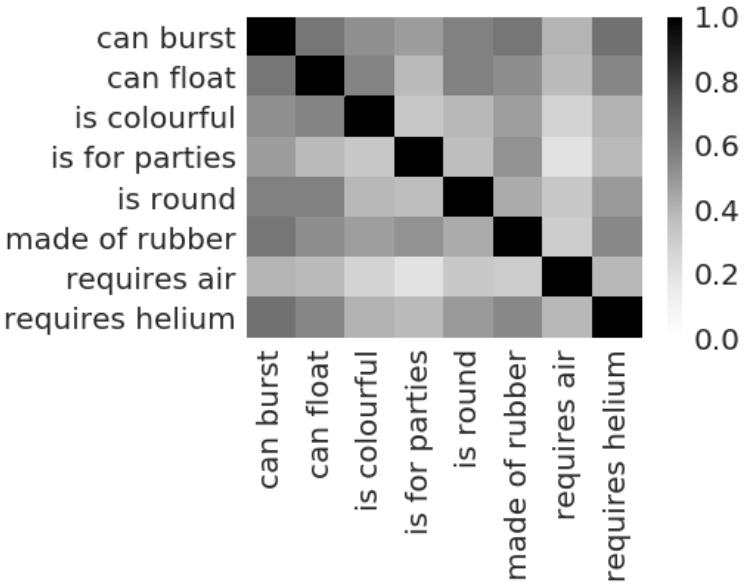


FIGURE B.21: Correlation - BALLOON

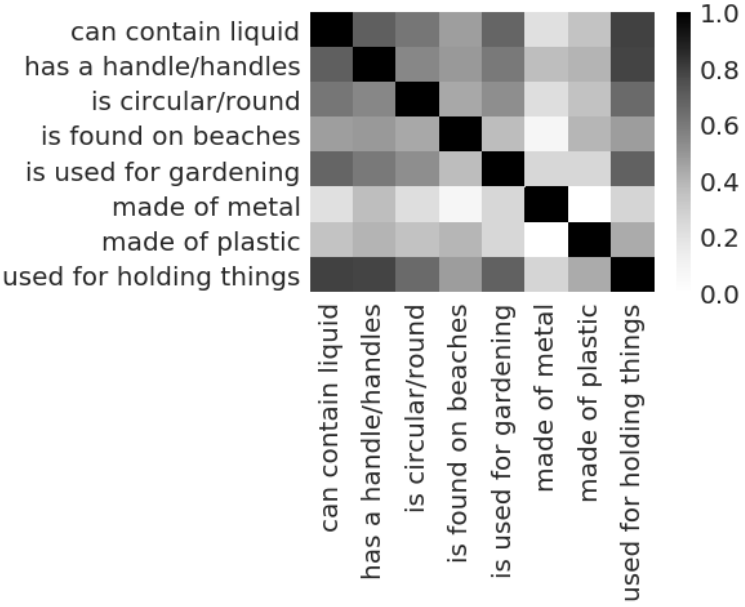


FIGURE B.22: Correlation - BUCKET

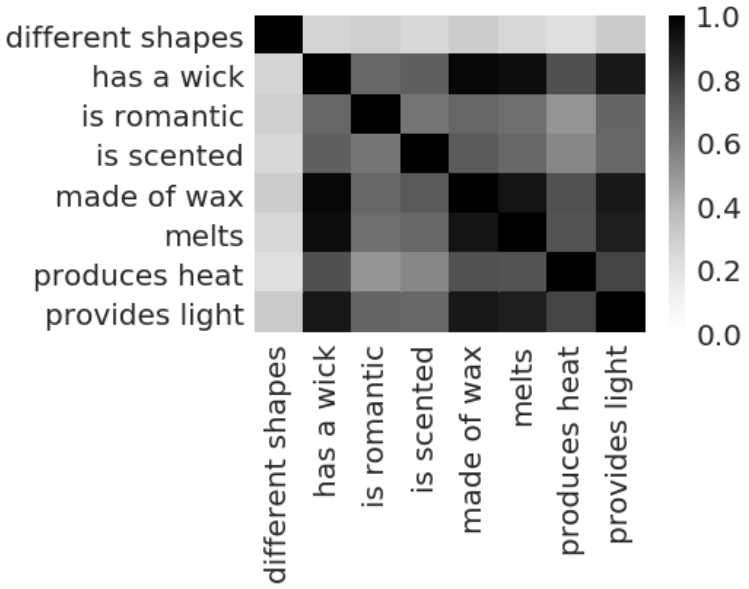


FIGURE B.23: Correlation - CANDLE

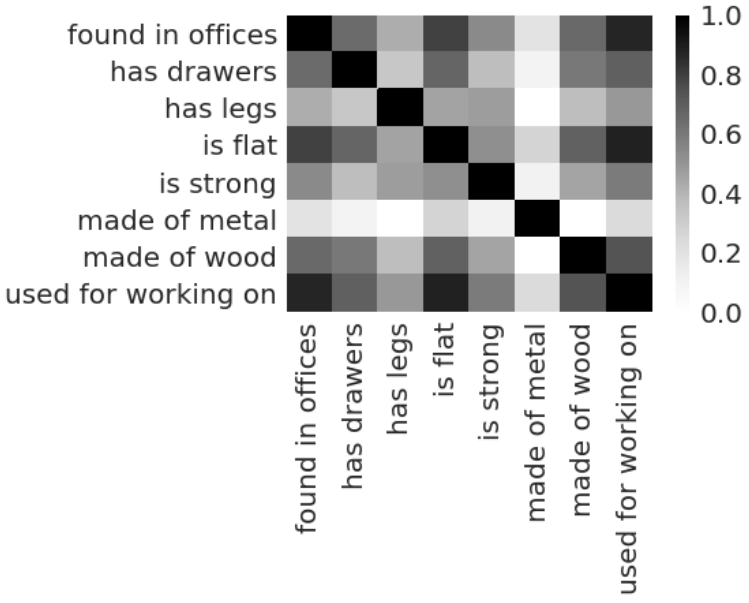


FIGURE B.24: Correlation - DESK

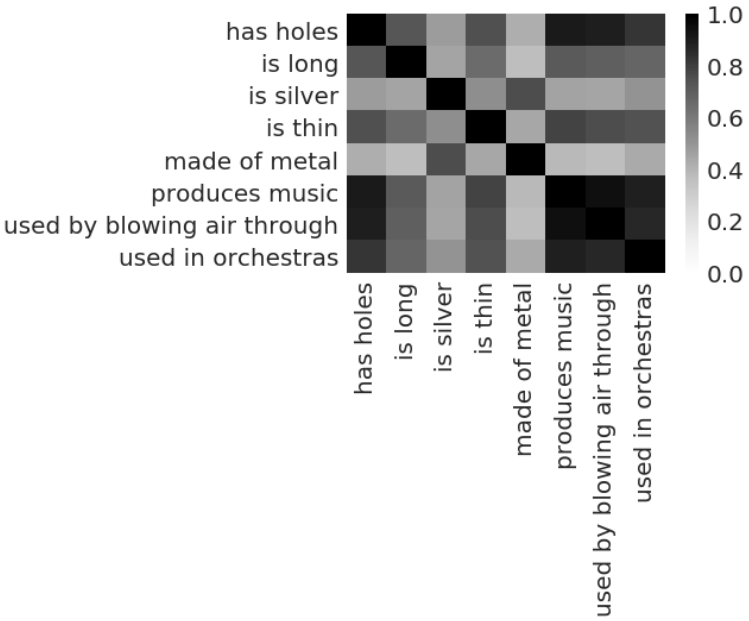


FIGURE B.25: Correlation - FLUTE

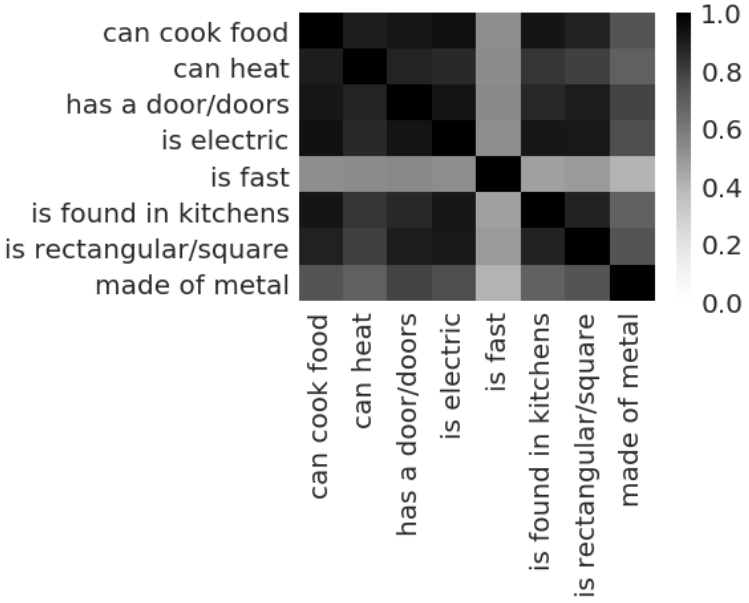


FIGURE B.26: Correlation - MICROWAVE

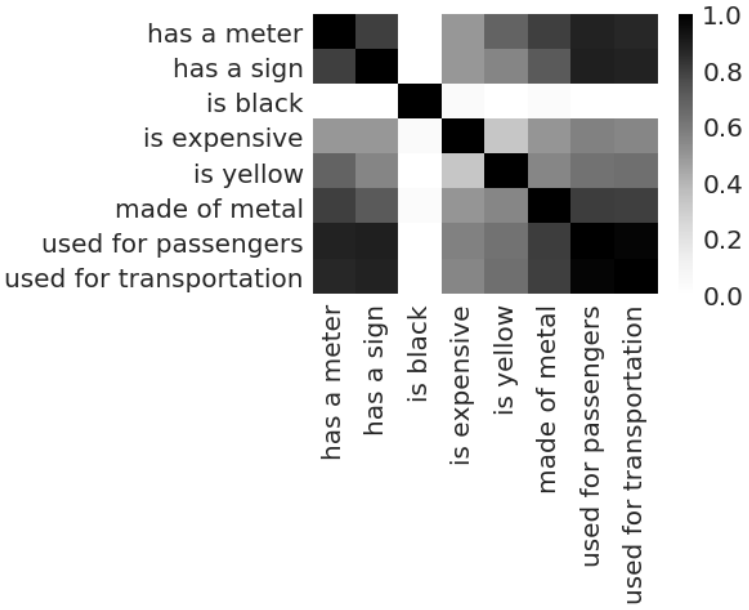


FIGURE B.27: Correlation - TAXI

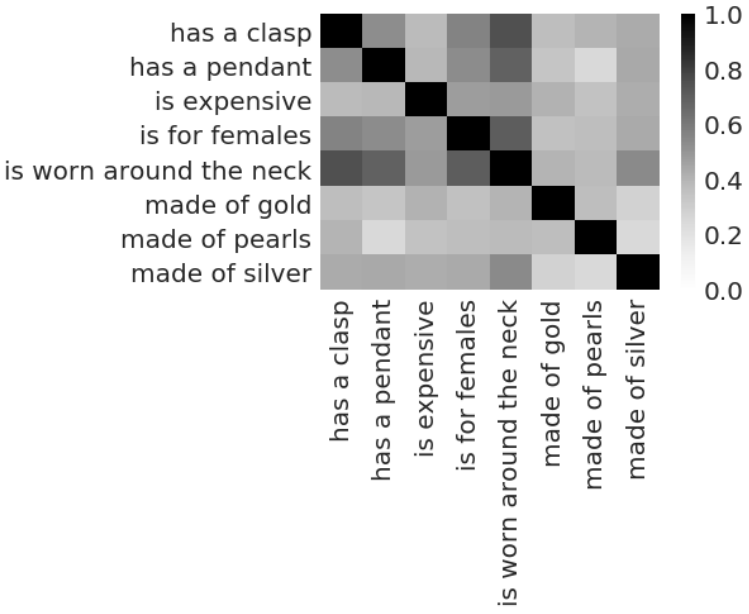


FIGURE B.28: Correlation - NECKLACE

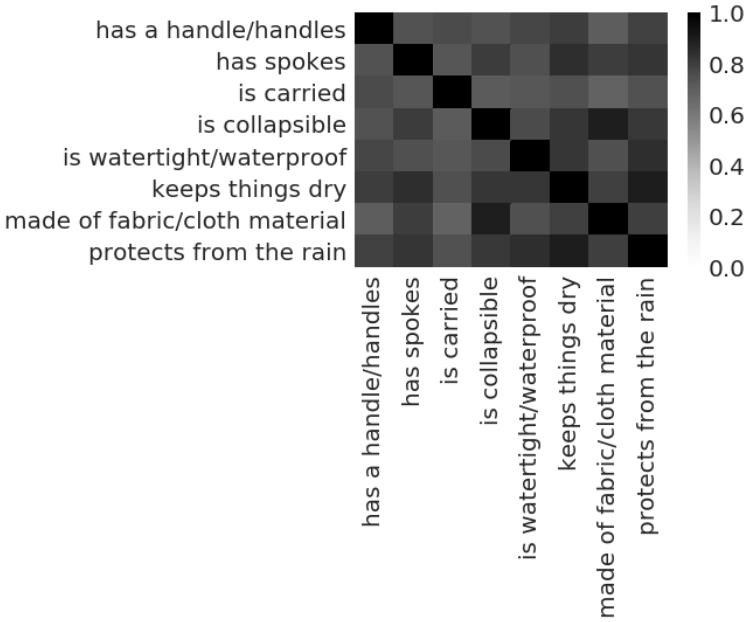


FIGURE B.29: Correlation - UMBRELLA

Appendix C

Additional Details: Part2

C.1 Variation of Generic Speech with Age

Adult Speech

Figures C.1 to C.5 highlight how the percentage of genericity within each category varies in adult speech as a function of the age of the child. Percentage of generic statements in the artifact category is higher than percentage of generic statements in the animal category for all age groups (although the differences are small).

Child Speech

The figures C.6 to C.9 highlight how the percentage of genericity within each category varies in child speech with age. Percentage of generic statements in the animal category is higher than the artifact category till the age of 4.

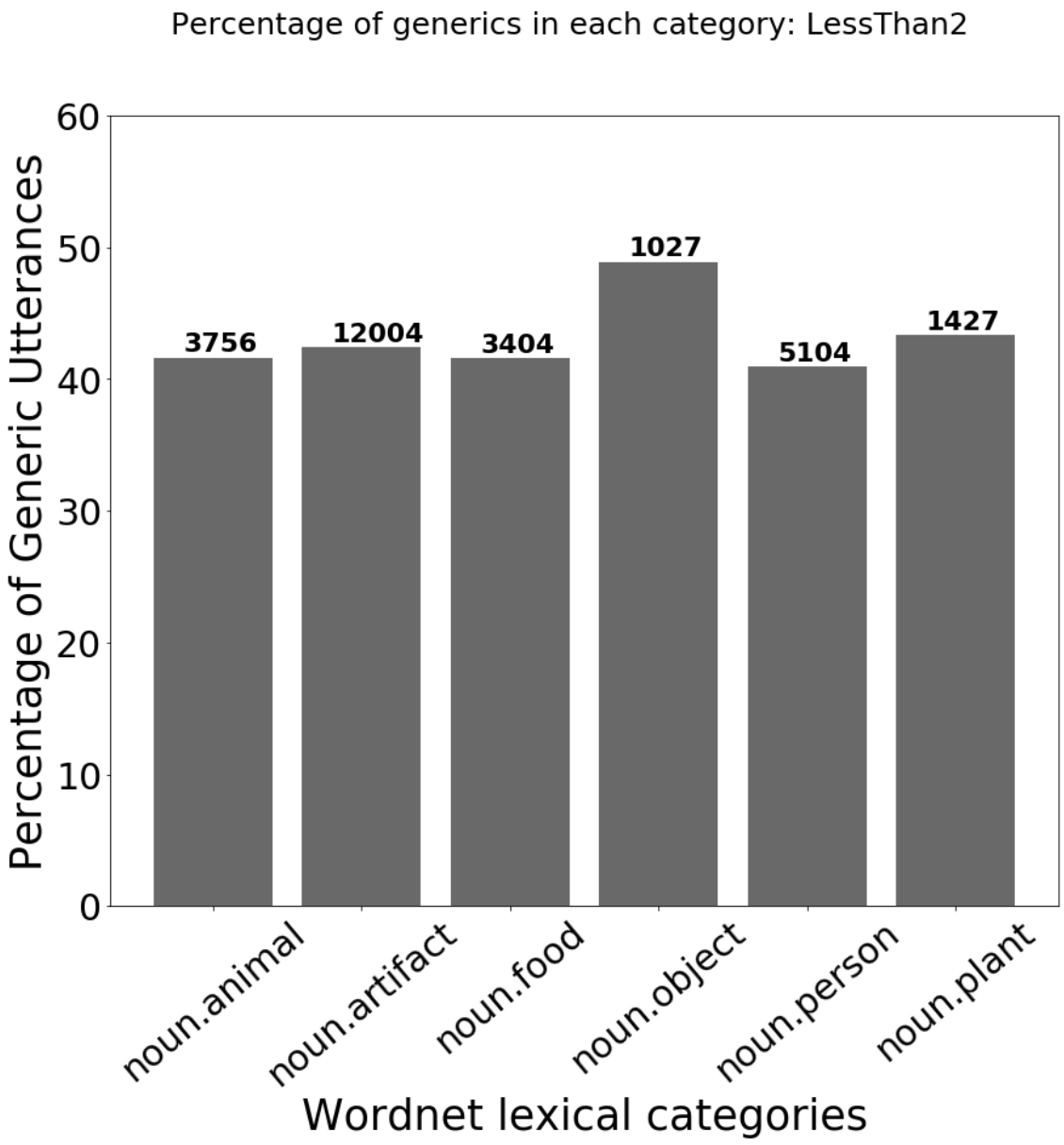


FIGURE C.1: Percentage Genericity in each category: Adult Speech

Age of the child: Less than 2 years. Each bar is calculated as the number of generic statements/total number of statements. The number above each bar is the number of generic statements from the particular category.

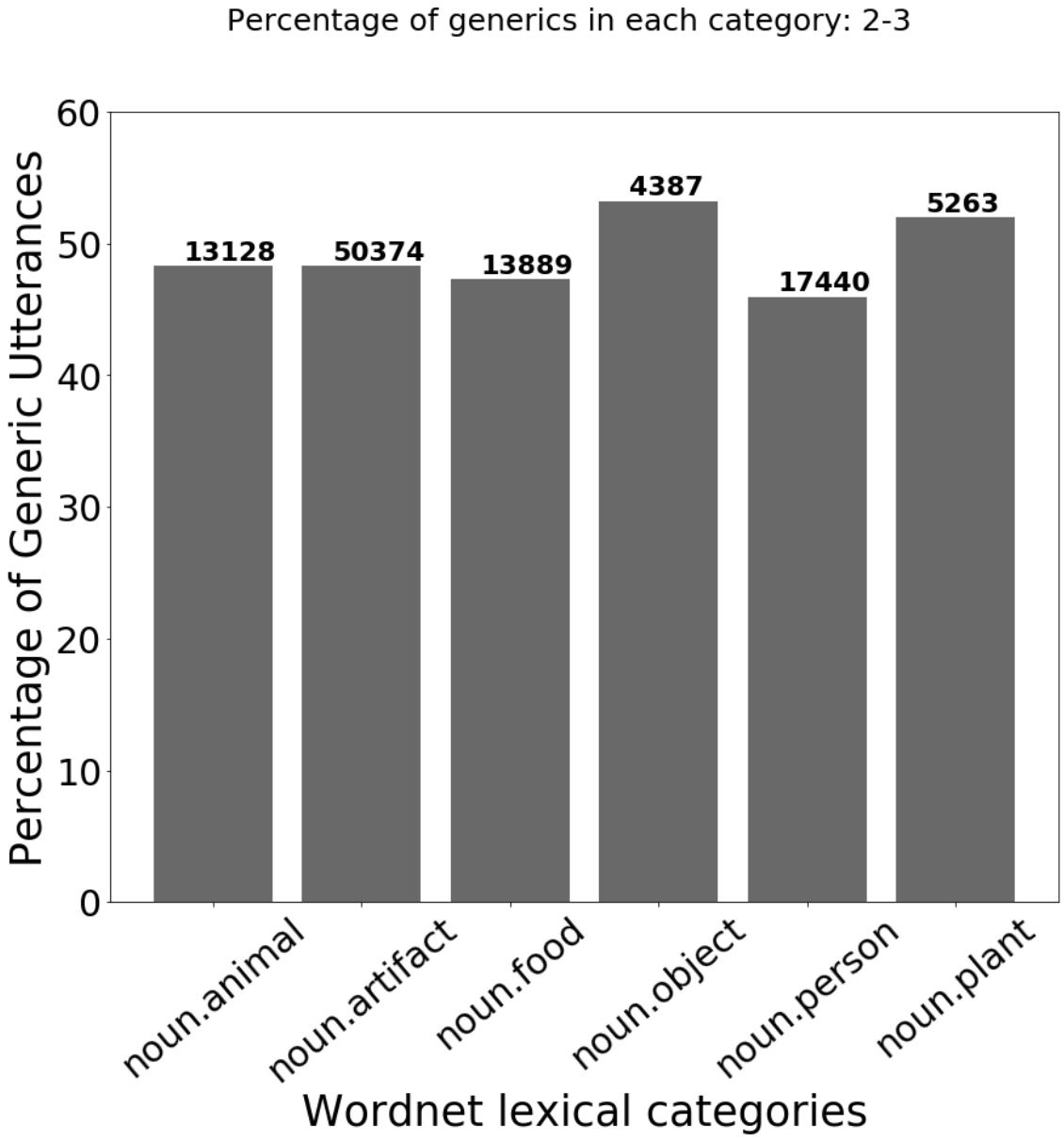


FIGURE C.2: Percentage Genericity in each category: Adult Speech

Age of the child: 2-3 years. Each bar is calculated as the number of generic statements/total number of statements.

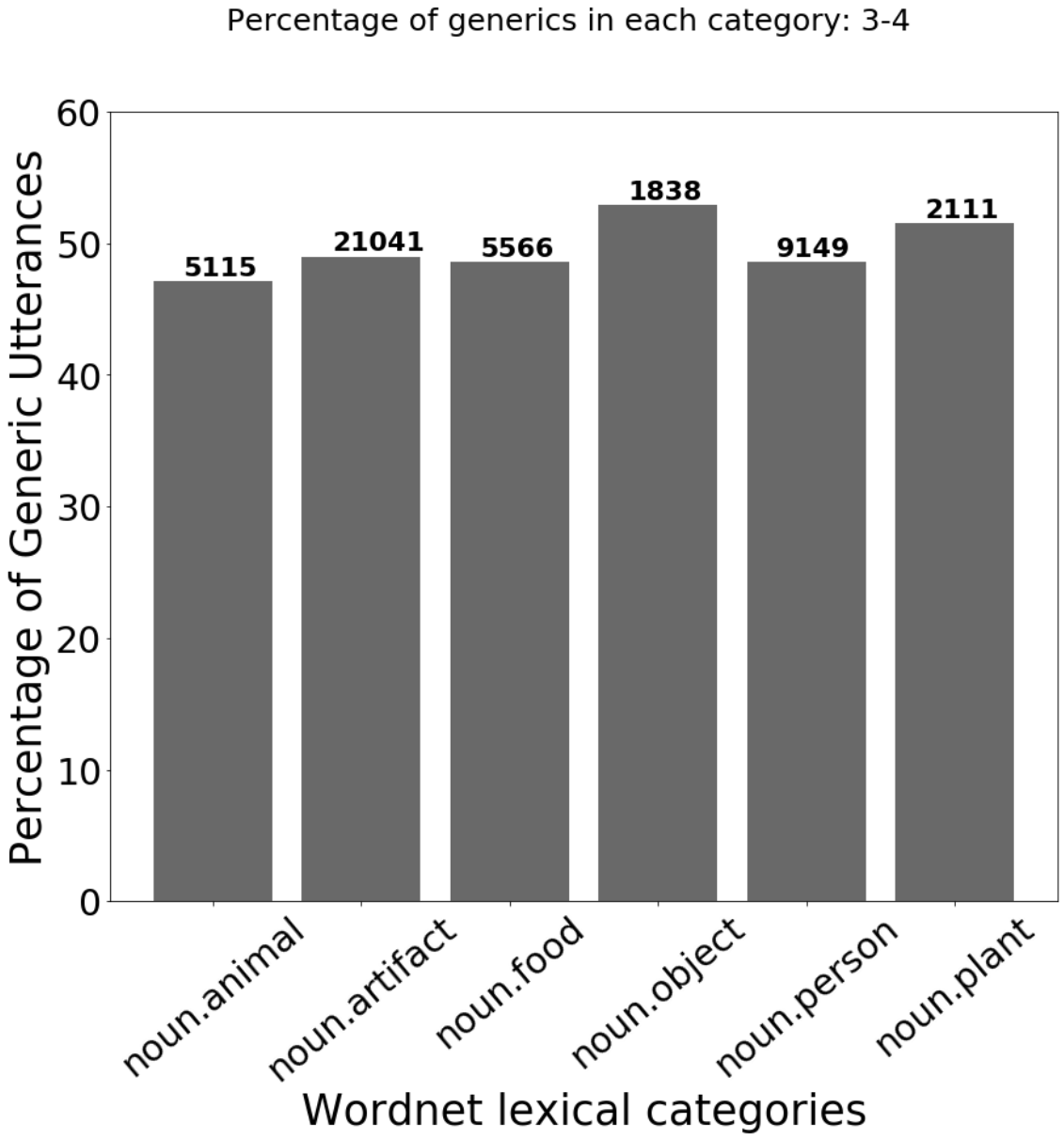


FIGURE C.3: Percentage Genericity in each category: Adult Speech

Age of the child: 3-4 years. Each bar is calculated as the number of generic statements/total number of statements.

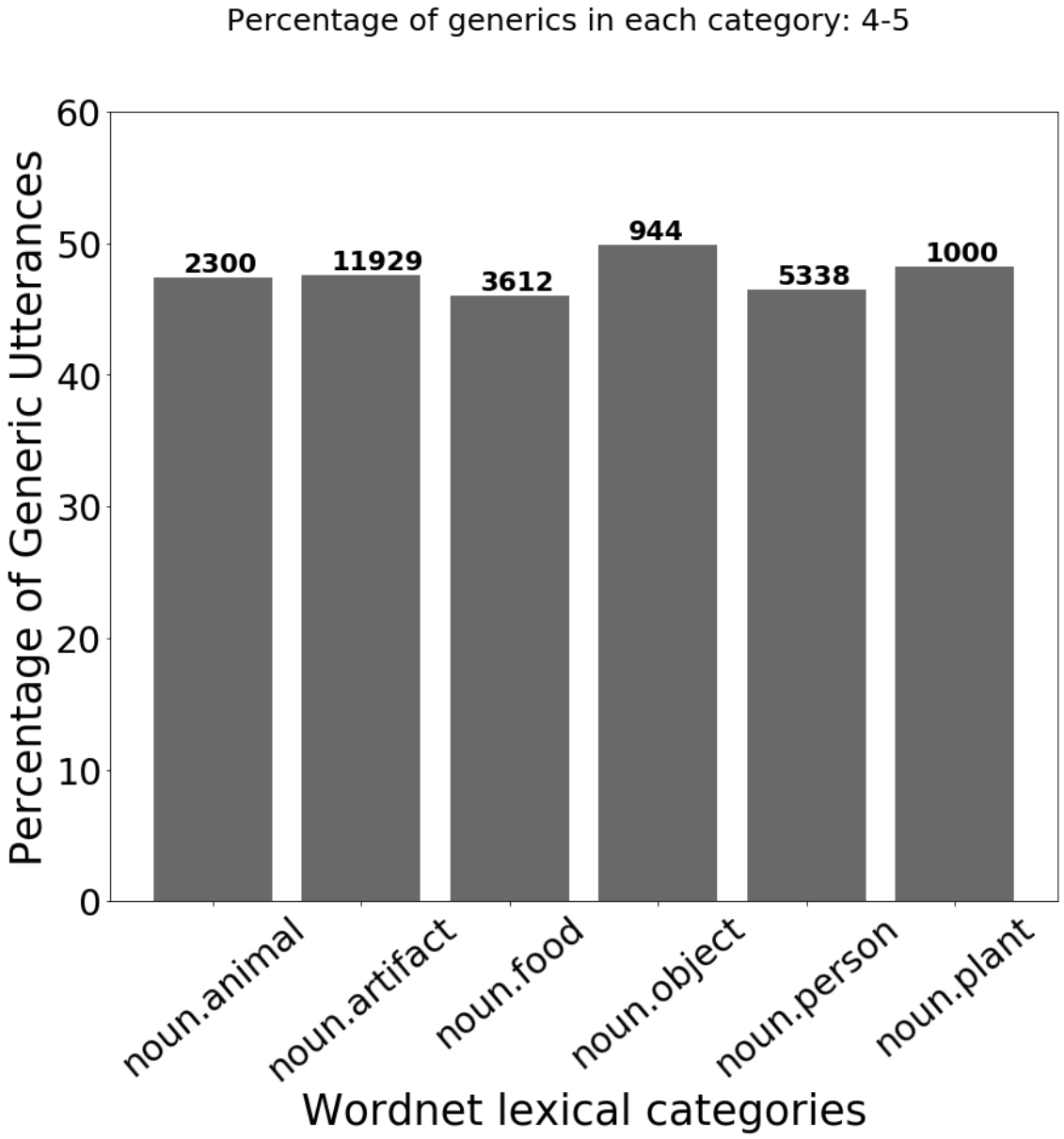


FIGURE C.4: Percentage Genericity in each category: Adult Speech

Age of the child: 4-5 years. Each bar is calculated as the number of generic statements/total number of statements.

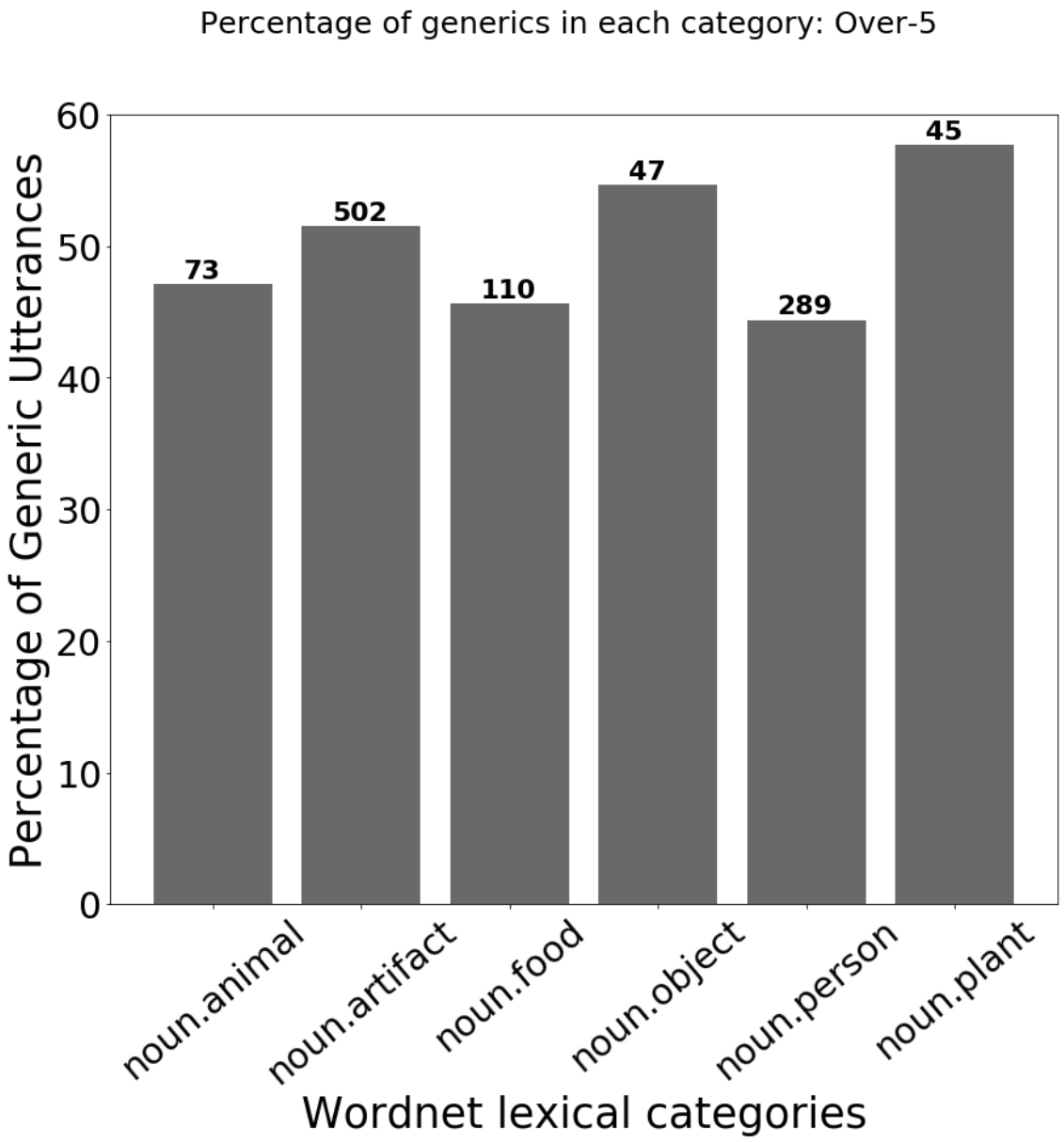


FIGURE C.5: Percentage Genericity in each category: Adult Speech

Age of the child: Over 5 years. Each bar is calculated as the number of generic statements/total number of statements.

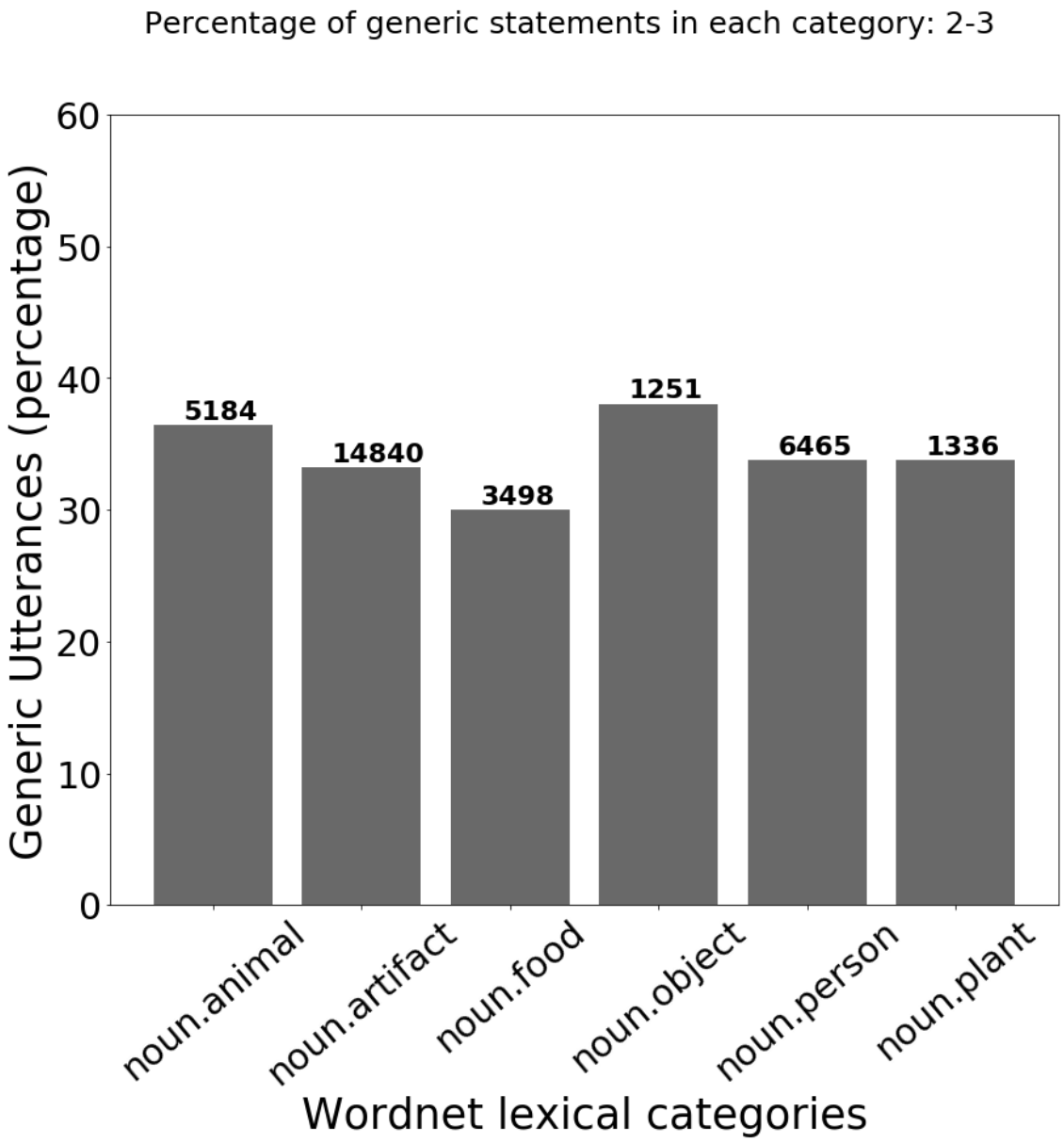


FIGURE C.6: Percentage Genericity in each category: Child Speech

Age of the child: 2-3 years. Each bar is calculated as the number of generic statements/total number of statements.

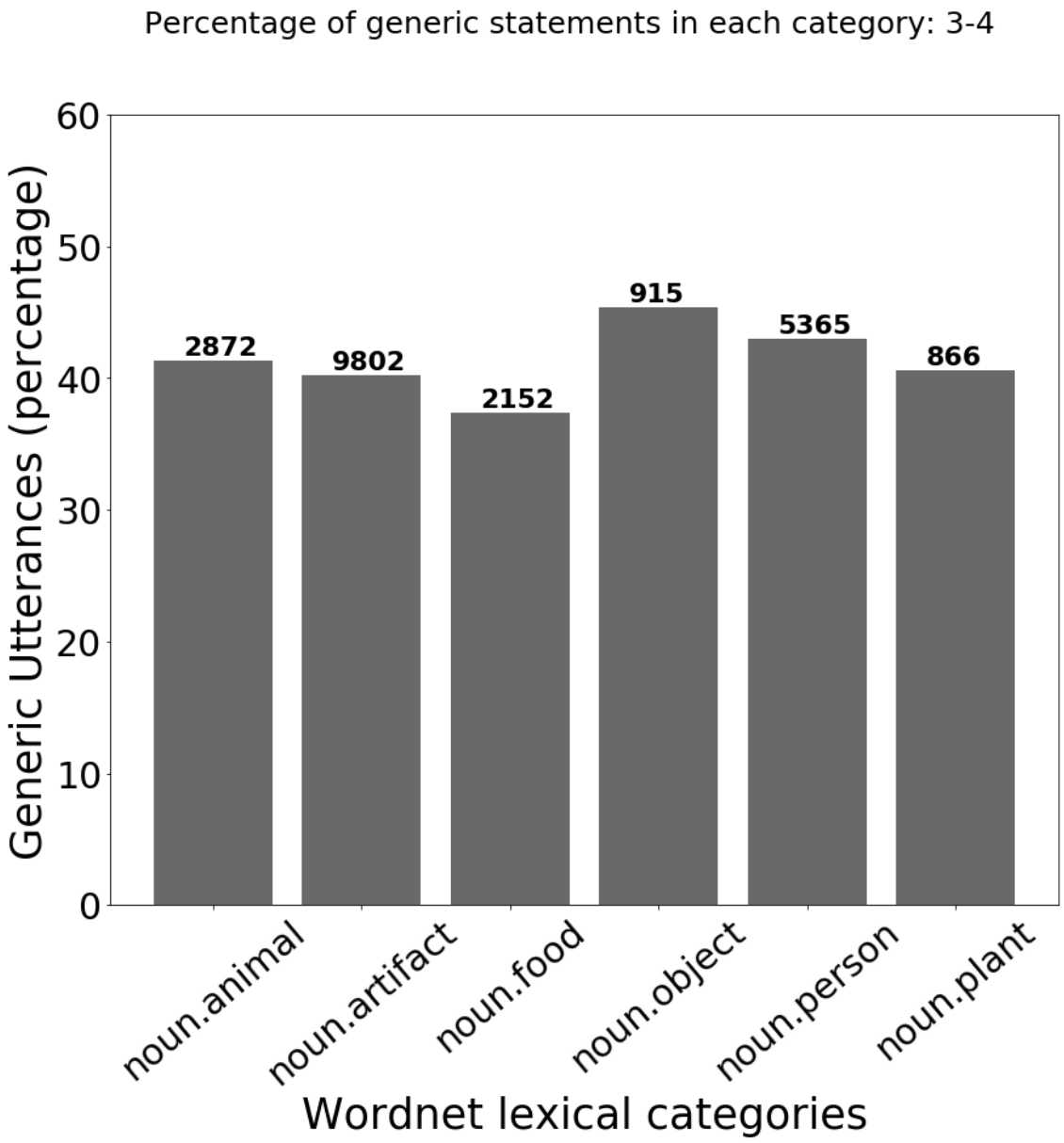


FIGURE C.7: Percentage Genericity in each category: Child Speech

Age of the child: 3-4 years. Each bar is calculated as the number of generic statements/total number of statements.

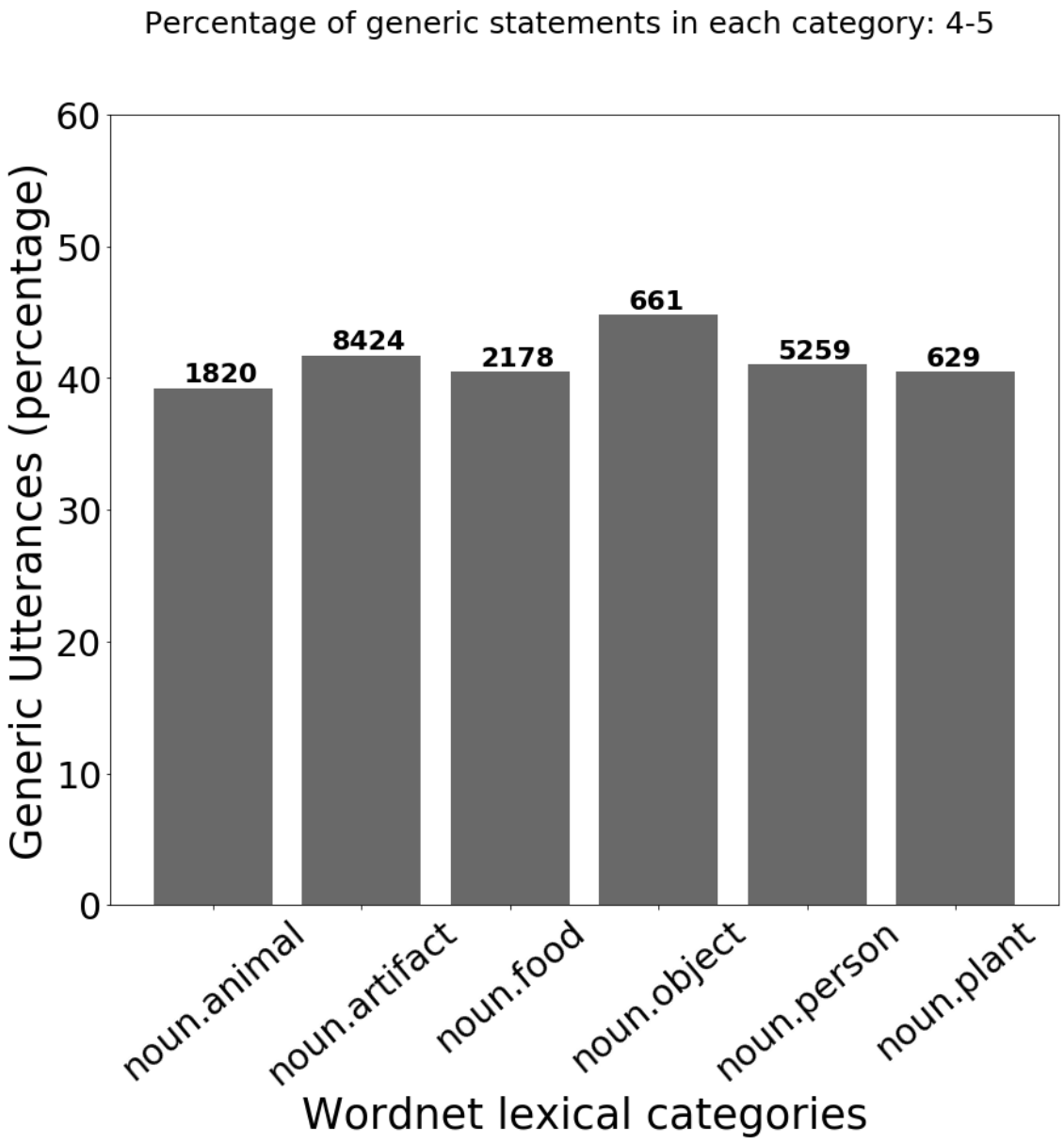


FIGURE C.8: Percentage Genericity in each category: Child Speech

Age of the child: 4-5 years. Each bar is calculated as the number of generic statements/total number of statements.

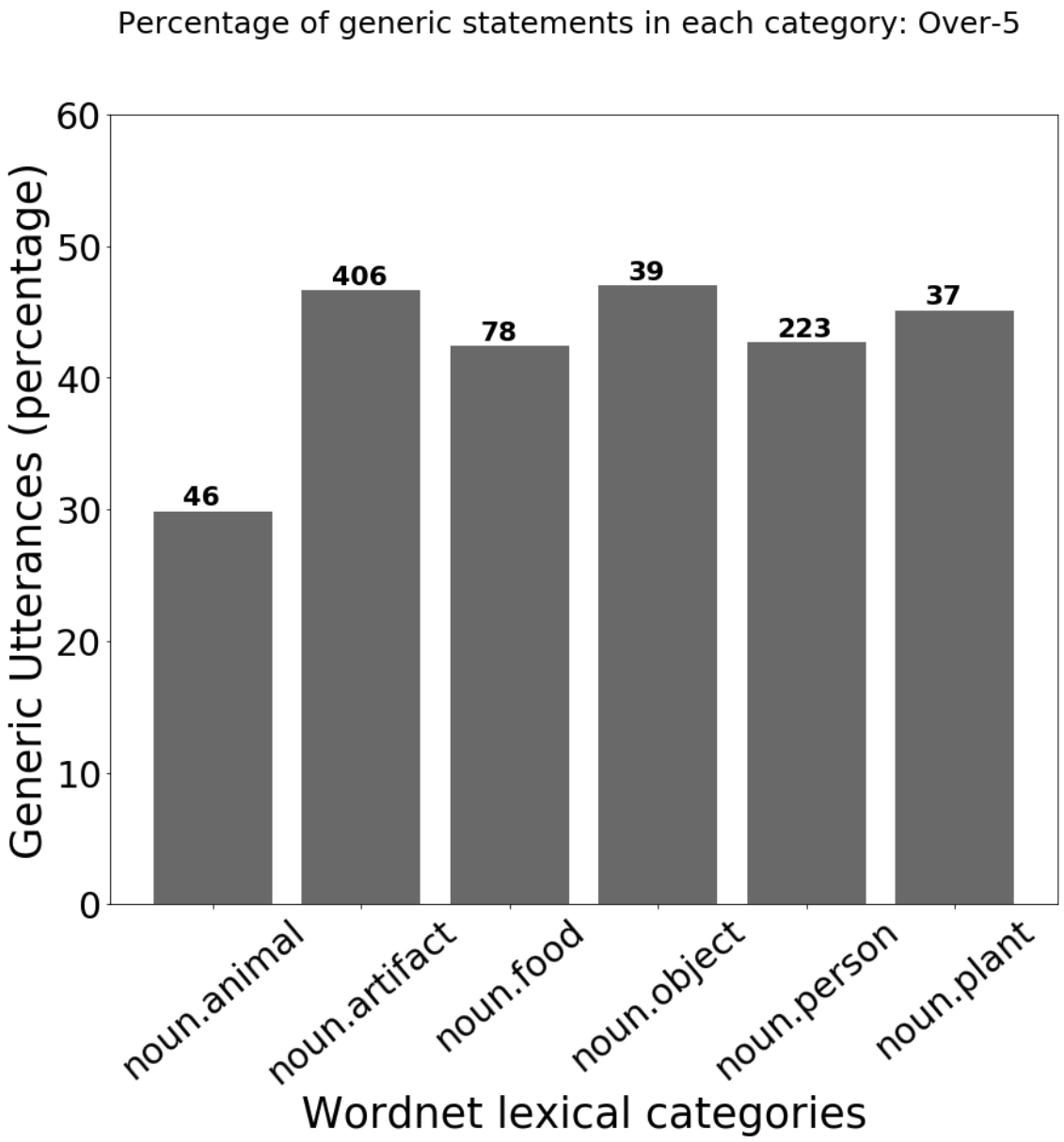


FIGURE C.9: Percentage Genericity in each category: Child Speech

Age of the child: Over 5 years. Each bar is calculated as the number of generic statements/total number of statements.

C.2 List of corpora (CHILDES)

- Brown
- Warren
- Forrester
- Demetras1
- MacWhinney
- Wells
- Demetras2
- Bloom70
- Peters
- Providence
- Sawyer
- Davis
- Braunwald
- Normal
- Soderstrom
- Belfast
- Snow
- Clark
- Higginson
- Feldman
- Morisset

- Cornell
- MPI-EVA-Manchester
- Hall
- Suppes
- Thomas
- Kuczaj
- Sachs

References

Ahn, W. et al. (2000). "Causal Status as a Determinant of Feature Centrality". In: *Cognitive Psychology* 41, pp. 361–416.

Ahn, W. et al. (2001). "Why essences are essential in the psychology of concepts". In: *Cognition* 82, pp. 59–69.

Bengio, Y., P. Simard, and P. Frasconi (1994). "Learning long-term dependencies with gradient descent is difficult." In: *IEEE Transactions on Neural Networks* 5, pp. 157–166.

Bloom, P (1996). "Intention, history, and artifact concepts". In: *Cognition* 60, pp. 1–29.

Cho, K. et al. (2014). "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *Proceedings of EMNLP*.

Colombo, D. et al. (2012). "Learning high-dimensional directed acyclic graphs with latent and selection variables". In: *Annals of Statistics* 40, pp. 294–321.

Cooper, G.F. and E. Herskovits (1992). "A Bayesian method for the induction of probabilistic networks from data". In: *Machine Learning* 9, pp. 308–347.

Deng, J. et al. (2009). "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR09*.

Devereux, B.J. et al. (2014). "The Centre for Speech, Language and the Brain (CSLB) Concept Property Norms". In: *Behaviour Research Methods* 46, pp. 1119–1127.

Friedrich, A. and M. Pinkal (2015). "Discourse-sensitive Automatic Identification of Generic Expressions". In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics*.

Gelman, S. (1988). "The development of induction within natural kind and artifact categories". In: *Cognitive Psychology* 20, pp. 65–95.

Gelman, S. (2003). *The Essential Child: Origins of Essentialism in Everyday Thought*. Oxford University Press.

Gelman, S. and E. Markman (1986). "Categories and induction in young children". In: *Cognition* 23, pp. 183–209.

Gelman, S. and H.M. Wellman (1991). "Insides and essences: Early understandings of the non-obvious". In: *Cognition* 38, pp. 213–244.

Gelman, S. A. and K. E. Kremer (1991). "Understanding natural causes: Children's explanations of how objects and their properties originate". In: *Child Development* 62, pp. 396–414.

Gelman, S. A. and T. Tardif (1998). "A cross-linguistic comparison of generic noun phrases in English and Mandarin". In: *Cognition* 66, pp. 215–248.

Gelman, S. A. et al. (2008). "A cross-linguistic comparison of generic noun phrases in English and Mandarin". In: *Language Learning and Development* 4, pp. 1–31.

Gopnik, A. and D.M. Sobel (2000). "Detecting Blickets: How Young Children Use Information about Novel Causal Powers in Categorization and Induction". In: *Child Development* 71, pp. 1205–1222.

Hochreiter, S. (1991). "Untersuchungen zu dynamischen neuronalen Netzen". PhD thesis. TU Munich.

Hochreiter, S. and J. Schmidhuber (1997). "Long Short-Term Memory". In: *Neural Computation* 9, pp. 1735–1780.

Johansson, R., D. Das, and P. Nugues (2008). "Dependency-based Semantic Role Labeling of PropBank". In: *Proceedings of the Twelfth Conference on Computational Natural Language Learning*.

Keil, F.C. (1989). *Concepts, Kinds, and Cognitive Development*. MIT Press.

Lake, B.M. et al. (2015). "Deep Neural Networks Predict Category Typicality Ratings for Images". In: *Proceedings of the 37th Annual Cognitive Science Society*.

MacWhinney, B. and C. Snow (1985). "The Child Language Data Exchange System". In: *Journal of Child Language* 12, pp. 271–295.

Margaritis, D. (2003). "Learning Bayesian Network Model Structure from Data". PhD thesis. Carnegie Mellon University.

Mason, W. and S. Suri (2012). "Conducting behavioral research on Amazon's Mechanical Turk". In: *Behaviour Research Methods* 44, pp. 1–23.

McRae, K. et al. (2005). "Semantic feature production norms for a large set of living and nonliving things". In: *Behaviour Research Methods* 37, pp. 547–559.

Medin, D.L. and A. Ortony (1989). "Psychological Essentialism". In: *Similarity and analogical reasoning*, pp. 179–195.

Peterson, J.C., J.T. Abbott, and T.L. Griffiths (2016). "Adapting Deep Network Features to Capture Psychological Representations". In: *Proceedings of the 38th Annual Conference of the Cognitive Science Society*.

Petrov, S., D. Das, and R. McDonald (2012). "A Universal Part-of-Speech Tagset". In: *Proceedings of LREC*.

Reiter, N. and A. Frank (2010). "Identifying Generic Noun Phrases". In: *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*.

Rips, L.J. (1989). "Similarity, typicality, and categorization". In: *Similarity and analogical reasoning*, pp. 21–59.

Schwarz, G. (1978). "Estimating the Dimension of a Model". In: *The Annals of Statistics* 6, pp. 461–464.

Spirtes, P. (2001). "An Anytime Algorithm for Causal Inference". In: *Presence of Latent Variables and Selection Bias in Computation, Causation and Discovery*. MIT Press, pp. 121–128.

Spirtes, P., C. Glymour, and R. Scheines (1993). *Causation, Prediction and Search*. Springer.

Strevens, M. (2000). "The Essentialist Aspect of Naive Theories". In: *Cognition* 74, pp. 149–175.

Suh, S. (2006). "Extracting Generic Statements for the Semantic Web". MA thesis. University of Edinburgh.

Szegedy, C. et al. (2015). "Going Deeper with Convolutions". In: *Computer Vision and Pattern Recognition (CVPR)*. URL: http://arxiv.org/abs/1409.4842.

Tsamardinos, I., L.E. Brown, and C.F. Aliferis (2012). "The max-min hill-climbing Bayesian network structure learning algorithm". In: *Machine Learning* 65, pp. 31–78.