2019

Relating Conceptual Structure With Flexible Concept Use

Sarah Solomon
University of Pennsylvania, sarahharrietsolomon@gmail.com

Follow this and additional works at: https://repository.upenn.edu/edissertations

Part of the Cognitive Psychology Commons, and the Neuroscience and Neurobiology Commons

Recommended Citation
https://repository.upenn.edu/edissertations/3414

This paper is posted at ScholarlyCommons. https://repository.upenn.edu/edissertations/3414
For more information, please contact repository@pobox.upenn.edu.
Relating Conceptual Structure With Flexible Concept Use

Abstract
Our mental words are populated with concepts — rich representations of knowledge about things in the world (e.g., diamonds, pumpkins). In language, words are used to refer to these concepts (e.g., “diamond”, “pumpkin”) and to communicate with others. This is quite impressive given that a word does not activate the same information each time it is used: conceptual information is flexibly activated based on the context. For example, the phrases “raw diamond”, “baseball diamond”, and “diamond eyes” evoke different kinds of diamond information. This flexible concept use is not only exemplified in creative language, but in creative thought and natural language more generally. The goal of this thesis was to leverage methods in cognitive neuroscience, network science, and computational modeling to explore the kinds of conceptual structure that can support this flexible concept use. In the first study (Chapter 2) I capture the global structure of concepts in novel feature-based networks, and show that aspects of this network structure relate to text-based and empirical measures of flexible concept use. I subsequently narrow in on the local representations of conceptual features that relate to flexible concept use by observing what happens when concepts combine. In one fMRI study (Chapter 3) I show that feature uncertainty predicts the extent to which features (e.g., green, salty) are flexibly modulated in the brain during comprehension of adjective-noun combinations (e.g., “green pumpkin”, “salty cookie”). In follow-up studies (Chapter 4) I further reveal the relationship between feature uncertainty and flexible feature activations in combined concepts. In combinations that modify conceptual brightness (e.g., “dark diamond”, “light night”), an explicit behavioral measure of conceptual feature modulation is predicted by feature uncertainty as well as by a related predictive combinatorial Bayesian model. An associated fMRI study reveals that flexible feature modulation and feature uncertainty relate to responses in left inferior frontal gyrus (LIFG) and left anterior temporal lobe (LATL), suggesting roles for these regions in flexible concept activation. Taken together, this work reveals relationships between conceptual structure and flexible concept use in behavior and in the brain.

Degree Type
Dissertation

Degree Name
Doctor of Philosophy (PhD)

Graduate Group
Psychology

First Advisor
Sharon Thompson-Schill

Keywords
cognitive combination, conceptual knowledge, fMRI, language, network science, semantic memory

Subject Categories
Cognitive Psychology | Neuroscience and Neurobiology | Psychology

This dissertation is available at ScholarlyCommons: https://repository.upenn.edu/edissertations/3414
RELATING CONCEPTUAL STRUCTURE WITH FLEXIBLE CONCEPT USE
Sarah Solomon
A DISSERTATION
in
Psychology
Presented to the Faculties of the University of Pennsylvania
in
Partial Fulfillment of the Requirements for the
Degree of Doctor of Philosophy
2019

Supervisor of Dissertation

Graduate Group Chairperson

Dr. Sharon Thompson-Schill
Christopher H. Browne Distinguished Professor of Psychology

Dr. Sara Jaffee
Professor of Psychology

Dissertation Committee:
Sharon Thompson-Schill, Christopher H. Browne Distinguished Professor of Psychology
Russell Epstein, Professor of Psychology
Lyle Ungar, Professor of Computer and Information Science
ACKNOWLEDGMENTS

Many people, animals, and places supported me while I completed the work in this dissertation and I would like to extend my gratitude to all of them. I first want to thank my advisor, Dr. Sharon Thompson-Schill, for the many years of mentorship. I benefited immensely from your analytical and creative insights, which gave me new ways to think about and solve problems. Thank you for letting me pursue my risky or offbeat ideas — some of which are now contained in this thesis. I am also appreciative of your culinary mentorship which introduced me to foods that I may not have tried otherwise.

My research output and research experience would not be what it is without the contribution of past and current members of the Thompson-Schill lab. The fresh ideas, constructive criticism, and professional and personal companionship you all provided was immensely valuable. Thank you to everyone who put up with me popping up at your desks unexpectedly to discuss the scientific or non-scientific crisis du jour.

I am also thankful for the larger communities at Penn that have supported me at various points and in various ways. I learned a lot from seminars and workshops hosted by the Center for Cognitive Neuroscience, Institute for Research in Cognitive Science, Computational Neuroscience Initiative, MindCORE, the NSF IGERT Traineeship in Complex Scene Perception, and the Psychology and Neuroscience departments. Thank you to individual faculty — especially committee members Dr. Russell Epstein and Dr. Lyle Ungar — for your valuable insights and feedback. And a big thank you to the fellow graduate students in the Psychology department who made this experience what it was (#gradschool).

I am also grateful for all of the Philadelphia cafés, restaurants, and bars in which I completed much of this work. Thank you for the space, the coffee, the delicious snacks, and the wine.

Thank you to Samantha (i.e., Sammies, Samsies, Sammers) the cat — you were with me for all of the highs and all of the lows. You are the cutest thing in the world.

And of course, thank you to my family for your ever-flowing support.
ABSTRACT

RELATING CONCEPTUAL STRUCTURE WITH FLEXIBLE CONCEPT USE

Sarah Solomon
Sharon Thompson-Schill

Our mental words are populated with concepts — rich representations of knowledge about things in the world (e.g., diamonds, pumpkins). In language, words are used to refer to these concepts (e.g., “diamond”, “pumpkin”) and to communicate with others. This is quite impressive given that a word does not activate the same information each time it is used: conceptual information is flexibly activated based on the context. For example, the phrases “raw diamond”, “baseball diamond”, and “diamond eyes” evoke different kinds of diamond information. This flexible concept use is not only exemplified in creative language, but in creative thought and natural language more generally. The goal of this thesis was to leverage methods in cognitive neuroscience, network science, and computational modeling to explore the kinds of conceptual structure that can support this flexible concept use. In the first study (Chapter 2) I capture the global structure of concepts in novel feature-based networks, and show that aspects of this network structure relate to text-based and empirical measures of flexible concept use. I subsequently narrow in on the local representations of conceptual features that relate to flexible concept use by observing what happens when concepts combine. In one fMRI study (Chapter 3) I show that feature uncertainty predicts the extent to which features (e.g., green, salty) are flexibly modulated in the brain during comprehension of adjective-noun combinations (e.g., “green pumpkin”, “salty cookie”). In follow-up studies (Chapter 4) I further reveal the relationship between feature uncertainty and flexible feature activations in combined concepts. In combinations that modify conceptual brightness (e.g., “dark diamond”, “light night”), an explicit behavioral measure of conceptual feature modulation is predicted by feature uncertainty as well as by a related predictive combinatorial Bayesian model. An associated fMRI study reveals that flexible feature modulation and feature uncertainty relate to responses in left inferior frontal gyrus (LIFG) and left anterior temporal lobe (LATL), suggesting roles for these regions in flexible concept activation. Taken together, this work reveals relationships between conceptual structure and flexible concept use in behavior and in the brain.
# TABLE OF CONTENTS

ACKNOWLEDGMENT .................................................................................................................. II

ABSTRACT .................................................................................................................................. III

LIST OF TABLES .......................................................................................................................... VII

LIST OF ILLUSTRATIONS ........................................................................................................... VIII

1: GENERAL INTRODUCTION ..................................................................................................... 1

1.1 Conceptual compositionality ............................................................................................... 2
   1.1.1 Neural models .............................................................................................................. 2
   1.1.2 Cognitive models ...................................................................................................... 4
   1.1.3 Computational models .............................................................................................. 6

1.2 Conceptual flexibility ............................................................................................................. 8
   1.2.1 Property flexibility .................................................................................................... 8
   1.2.2 State flexibility ......................................................................................................... 9
   1.2.3 Referent flexibility ................................................................................................. 10

1.3 Previous models of conceptual flexibility ........................................................................ 11

1.4 Conceptual combination ..................................................................................................... 13

1.5 Current approach .............................................................................................................. 16

2: IMPLEMENTING A CONCEPT NETWORK MODEL ........................................................ 17

2.1 Introduction ....................................................................................................................... 17

2.2 Methods ............................................................................................................................. 23
   2.2.1 Network Methods ................................................................................................... 23
   2.2.2 Figurative Language Task ..................................................................................... 30
   2.2.2 Alternative Uses Task .......................................................................................... 32

2.3 Results ............................................................................................................................... 34
   2.3.1 Classification Results ............................................................................................. 34
   2.3.2 Network Measures of Conceptual Structure ......................................................... 35
   2.3.3 Bootstrap Results ................................................................................................. 37
   2.3.4 Similes ................................................................................................................... 39
   2.3.5 Alternative Uses .................................................................................................... 39
2.4 Discussion .................................................................................................................. 40

3: FEATURE UNCERTAINTY PREDICTS THE FLEXIBLE NEURAL ACTIVATION OF CONCEPTUAL FEATURES IN COMBINED CONCEPTS .................................................................................................................. 53

3.1 Introduction .................................................................................................................. 55

3.2 Methods ...................................................................................................................... 60
  3.2.1 Participants ............................................................................................................. 60
  3.2.2 Adjective and noun stimuli ..................................................................................... 60
  3.2.3 Feature surprisal and feature uncertainty ............................................................... 61
  3.2.4 Analyzing concept network core-periphery structure ............................................. 62
  3.2.5 Task order and design .......................................................................................... 63
  3.2.6 Functional MRI acquisition and analysis ............................................................... 64
  3.2.7 Searchlight voxel selection .................................................................................... 64
  3.2.8 Property modulation in adjective-noun combinations ........................................... 65
  3.2.9 Validating VOIs and analysis methods ................................................................. 67

3.2 Results ....................................................................................................................... 68
  3.3.1 Feature uncertainty predicts degree of property modulation .................................. 68
  3.3.2 Concept network structure predicts degree of property modulation ...................... 70
  3.3.3 Neural representation of properties in unmodified noun concepts ......................... 71

3.2 Discussion .................................................................................................................. 71

4: PROBABLISTIC MODELS OF FEATURE UNCERTAINTY PREDICT BEHAVIORAL AND NEURAL RESPONSES TO COMBINED CONCEPTS .................................................................................................................. 79

4.1 Introduction .................................................................................................................. 79

4.2 Methods ...................................................................................................................... 80
  4.2.1 Participants ............................................................................................................. 80
  4.2.2 Adjective and noun stimuli ..................................................................................... 81
  4.2.3 Behavioral measures of conceptual brightness ....................................................... 81
  4.2.4 Predictive models of adjective-noun combination .................................................. 83
  4.2.5 fMRI task and design .......................................................................................... 86
  4.2.6 fMRI acquisition and analysis .............................................................................. 87
  4.2.7 Defining a priori ROIs ......................................................................................... 88
  4.2.8 Neural adjective effects ....................................................................................... 89
  4.2.9 Analyzing neural effects ..................................................................................... 89

4.3 Results ....................................................................................................................... 90
  4.3.1 Feature uncertainty predicts ground-truth adjective effects .................................... 90
  4.3.2 Bayesian model best predicts ground-truth adjective effects .................................. 91
  4.3.3 Neural responses to combined concepts ............................................................... 92

4.4 Discussion .................................................................................................................. 95
4.5 Appendix................................................................................................................. 99
4.6 Supplemental Materials......................................................................................... 100

GENERAL DISCUSSION .............................................................................................. 102

BIBLIOGRAPHY .......................................................................................................... 109
LIST OF TABLES

Table 1: Statistics and correlations between network measures.

Table S1: Included and excluded subordinates for all concepts.

Table A1: Noun concepts.

Table S2: Significant clusters in whole-brain searchlight analysis.
LIST OF FIGURES

Figure 1: A visualization of network topologies and measures.

Figure 2: Visualizing the chocolate network.

Figure 3: Example images used to generate test data in classification analysis.

Figure 4: Simile ratings.

Figure 5: Classification results.

Figure 6: Network predictors of Semantic Diversity.

Figure 7: Results of bootstrap analyses.

Figure 8: Network predictors of creative concept use.

Figure 9: Predicting conceptual combination.

Figure S1: Subordinate bootstrap analysis of network measures and simile goodness.

Figure S2: Property bootstrap analysis of network measures and simile goodness.

Figure S3: Subordinate bootstrap analysis of network measures and AUT creativity.

Figure S4: Property bootstrap analysis of network measures and AUT creativity.

Figure 10: Adjective and noun stimuli.

Figure 11: Calculating neural property modulation.

Figure 12: Feature surprisal and feature uncertainty.

Figure 13: Concept network analysis.

Figure 14: Explicit measures of conceptual brightness.

Figure 15: Predictive combinatorial models.

Figure 16: A priori neural regions of interest (ROIs).

Figure 17: Behavioral results.

Figure 18: fMRI results.

Figure S5: Overlap between ground-truth effects and feature uncertainty.
1: GENERAL INTRODUCTION

Concepts are the currency of meaning. We rely on concepts — such as green, sharp, diamond, and shadow — to understand the world around us and to communicate this information to others using language. Our ability to successfully use this conceptual currency to support both linguistic and nonlinguistic aspects of cognition is remarkable when one considers the flexibility of conceptual meaning across contexts. The information evoked by diamond varies across the phrases raw diamond, baseball diamond, and diamond eyes. Despite this flexibility, the conceptual system is not in a hopeless state of anarchy: conceptual structure provides constraints on the cognitive system such that we are able to settle on an appropriate meaning in each case.

The challenge is to characterize the kinds of conceptual structure that enable this flexibility to emerge: this structure can be analyzed from cognitive, neural, and computational perspectives. The cognitive structure of conceptual knowledge refers to the way in which semantic knowledge is organized and represented, and the kinds of cognitive processing that these structures enable. Around the time Tulving (1972) coined the term “semantic memory”, different theories of semantic structure emerged, including network-based (Collins & Quillian, 1969; Collins & Loftus, 1975) and feature-based theories (Smith et al., 1974; Rosch, 1975; Rosch & Mervis, 1975). Whereas classic network-based theories represent semantic knowledge as a web of interconnected concepts, feature-based theories represent individual concepts in terms of their corresponding features. These feature-based theories are “compositional”, a characteristic of conceptual structure that is important for theories of conceptual flexibility. The neural structure of conceptual knowledge refers to how conceptual information is embedded and organized within the brain; neural theories of conceptual structure rely on empirical data from neurological patients and neuroimaging studies to localize specific kinds of semantic information to different anatomical regions or functional networks. Finally, computational approaches can help validate these cognitive and neural theories and can even inform theories of conceptual structure (Love, 2015).

The general goal of this thesis is to relate theories of conceptual structure with flexible concept use. As illustrated in the diamond example, the information evoked by a concept or word varies across contexts. These contexts can be linguistic in nature (e.g., complex phrases, conceptual combination, figurative language) or non-linguistic (e.g., generating novel uses for a common object). In Chapter 2, I model individual concepts (e.g., cookie, knife) as distinct feature-based networks and explore whether aspects of this network structure relate to
linguistic and non-linguistic forms of flexible concept use (e.g., figurative language, novel uses). In Chapters 3 and 4, I use conceptual combination (e.g., *sharp diamond, dirty baseball*) as a tool to study flexible behavioral and neural responses to concepts in different verbal contexts.

In the empirical work reported here, I use a compositional, feature-based conceptual structure to explain aspects of conceptual flexibility. I will thus begin with a review of compositionality in cognitive, neural, and computational theories of conceptual knowledge (1.1), followed by a review of conceptual flexibility (1.2) and how previous models have attempted to capture this flexibility (1.3). I will then specifically review cognitive, neural, and computational models of conceptual combination (1.4) before summarizing the empirical work contained in this thesis (1.5).

1.1 Conceptual compositionality

Compositionality is a principle that states that the meaning of one structure depends on the meaning of its constituent elements and the ways in which those elements are combined. In the domain of conceptual knowledge, compositionality suggests that the meaning of a concept emerges as a function of its properties (or “features”) and their associations. This is a central component of all feature-based theories, in which concepts are considered to be distributed representations over featural information spanning multiple types of knowledge. Conceptual compositionality has been implemented in cognitive, neural, and computational architectures. Some theories incorporate an additional component which integrates feature-specific information into a transmodal, or amodal, representation. These components are commonly referred to as “convergence zones”, “hubs”, and “hidden layers” in cognitive, neural, and computational models, respectively. I will first review the compositional neural perspectives of conceptual knowledge. I will then discuss compositional cognitive theories of conceptual organization and the computational models used to implement them.

1.1.1 Neural models

From a cognitive neuroscience perspective, feature-based (i.e., compositional) models are neurally plausible. This is suggested by many studies revealing how the brain processes distinct features of the environment. In visual neuroscience, distinct neural regions or processes have been associated with specific visual features such as orientation (Hubel & Wiesel, 1962; Haynes & Rees, 2005), shape (Kanwisher et al., 1997; Kourtzi & Kanwisher, 2001), and color (Martin et al., 1995; Beauchamp et al., 1999). Outside of the visual modality, researchers have explored the neural organization or locations of other properties such as real-world size (Konkle & Oliva, 2012), texture (Lederman et al., 2001; Cavina-Pratesi et al., 2010), and manipulation (Boronat et al., 2005; Buxbaum et al.,
Having a grasp on how and where these features are processed in the brain allows us to probe the activation of these features during conceptual processing. For example, Coutanche & Thompson-Schill (2014) found that the appropriate activation of color features (i.e. GREEN, ORANGE) and shape features (i.e., ROUND, ELONGATED) predicted the success of object concept classification (e.g., LIME, CARROT).

The compositional assumption has met success in predicting concept-evoked brain activity. In vector-space models, each concept is defined as a vector in multidimensional space based on its rate of co-occurrence with other words in language. Distributed neural activation relating to concrete objects (e.g., CELERY, AIRPLANE) can be predicted by the extent to which those concepts co-occur with a set of feature concepts (e.g., TASTE, FLY) in a large text corpus (Mitchell et al., 2008). This concept-evoked neural activity can also be predicted by object-property associations found in feature-norms data (Chang et al., 2011). Neuroimaging evidence thus suggests that concepts can be decomposed into features, the activation of which appears to contribute to the informational content of the concept as a whole (Mitchell et al., 2008; Chang et al., 2011; Coutanche & Thompson-Schill, 2014).

The convergence zone framework of Damasio (1989) provides another feature-based neural model of conceptual processing. In this view, conceptual content depends on the activation of feature fragments in posterior sensorimotor cortices (e.g., visual cortex, auditory cortex), coordinated by convergence-divergence zones (CDZs) in more anterior regions (Meyer & Damasio, 2009). These CDZs act as repositories for combinatorial codes that coordinate appropriate activity in posterior modality-specific regions of the brain: the meaning of a concept resides in the synchronized co-activation of feature fragments and CDZs, with the feature fragments corresponding to modality-specific features of objects (e.g., color, texture, sound). This model is compositional and distributed, and links conceptual content to a neural architecture. Though in its original form it is computationally under-specified, CDZs are sometimes incorporated into connectionist-style models in the form of a hidden layer (see McNorgan et al., 2011).

Patterns of semantic deficits in neurological patients have provided extremely informative data regarding the ways in which conceptual information is organized in the brain. Some patients with brain damage exhibit category-specific deficits in which knowledge in only one semantic domain is affected. The most common pattern of deficits is that knowledge of living things is selectively degraded, whereas knowledge of artifacts is spared, though the reverse is also possible (for a review, see Tyler & Moss, 2001). An early explanation for these findings was that the conceptual system is organized into distinct stores of sensory and functional properties (Warrington & Shallice, 1984), and that damage to these
stores results in deficits that appear category specific. Caramazza & Shelton (1998) claimed that this organization does not explain the more complex patterns of semantic deficits, and hypothesized that domains of knowledge themselves (e.g., animals, artifacts, food) are the first-order organizing principle. Further claiming that this segregated organization does not adequately account for the patterns of patient deficits and observed neuroimaging results, Tyler and colleagues (Tyler et al., 2000; Tyler & Moss, 2001) proposed a framework in which many features (perceptual, functional encyclopedic) are distributed throughout a unitary conceptual system. Cree & McRae (2003) conducted a detailed analysis of the relationship between feature distributions and semantic deficits, revealing the factors (e.g., feature distinctiveness, concept familiarity) that can explain how deficits that appear category-specific can emerge from a distributed conceptual system based on modality-specific features. These approaches harken back to earlier feature-based conceptual theories (e.g. Smith et al., 1974).

1.1.2 Cognitive models

After Tulving (1972) coined the term “semantic memory,” the field was dominated by “semantic network” models for quite some time. These models included hierarchical network theory (Collins & Quillian, 1969) and the spreading activation theory (Collins & Loftus, 1975); both represent individual basic-level concepts (e.g., PENGUIN, PARROT) as distinct units in memory, connected to each other within a larger semantic network. The hierarchical network model (Collins & Quillian, 1969) includes higher-level category nodes (e.g., BIRD, ANIMAL) and uses these to encode superordinate relationships (e.g., PENGUIN is a BIRD) in an explicitly hierarchical system. The spreading activation model (Collins & Loftus, 1975) represents both concepts and features as distinct units, and connections between them are based on strength of their relationship. In this framework, concepts are represented as stand-alone units rather than sets of features.

The relationship between concepts and their features has been a point of contention. Research on judgments of category membership and exemplar typicality suggest that boundaries between concepts are fuzzy rather than clear-cut (McCloskey & Glucksberg, 1978), and exhibit a graded structure (Mervis & Rosch, 1981). However, the classical network models cannot account for these findings. Representing category membership in terms of propositions (e.g., by linking exemplars to superordinate categories via “isa” links) rather than sets of features cannot provide a satisfying explanation for graded category structure. Similarly, feature-based models in which concepts are represented as sets of necessary and sufficient features would suggest well-defined, rather than fuzzy, category boundaries (see Smith et al., 1974). Instead, the continuous variation in degree of category membership is best explained by a conceptual organization in which concepts are composed of characteristic, but not defining, features.
(McCloskey & Glucksberg, 1978). That is, some properties may be more central
to the concept, or more reliably associated with the concept, than others.

A fully compositional, feature-based approach emerged in the “conceptual
structure” framework (Tyler & Moss, 2001). In this account, many kinds of
features are distributed throughout a single conceptual system. The properties in
this model can be perceptual (e.g., BLUE, HAS-WINGS), functional (e.g., EATS,
CUTS), encyclopedic (e.g., LIVES IN ANTARCTICA), and category label (e.g., IS A
VEHICLE). Concepts (e.g., PENGUIN, CAR) are represented as vectors indicating
the presence or absence of these features, and semantic domains differ in terms
of how these features are statistically distributed across its members. This is a
compositional and distributed theory of concepts, in that conceptual meaning
emerges from the distributed pattern of properties spanning multiple kinds of
knowledge.

In particular, the conceptual structure account emphasizes property
distinctiveness and property correlations in characterizing the unique structures
of distinct domains. Property correlations refer to the extent to which certain
properties co-occur across exemplars within a domain: for example, the
properties HAS-WINGS and FLIES tend to co-occur with each other within the
domain of living things. Property distinctiveness refers to cue validity: highly
distinctive properties are only present in a small number of concepts, making
these properties useful cues for establishing concept identity. For example, HAS-
TENTACLES is a distinctive property since it only applies to a handful of concepts
(e.g. OCTOPUS, SQUID), whereas HAS-WINGS is far less distinctive. This cognitive
architecture makes certain predictions about human performance on behavioral
tasks, which can be tested against human performance (Randall et al., 2004).
Importantly, the correlational structure used to characterize semantic domains
(e.g., ARTIFACTS) might also be used to characterize the structure of individual
concepts within those domains (e.g., KNIFE), as I explore in Chapter 2.
Additionally, one of the benefits of a compositional conceptual theory is the ability
to define property-level measures, such as distinctiveness, that relate to
conceptual processing. Chapters 3 and 4 examine a property-level measure of
uncertainty and explore whether it influences how combined concepts are
processed.

The “feature-correlation” account of conceptual knowledge, proposed by McRae
and colleagues, is similar to the conceptual structure account but differs in a few
important ways. First, they argue that conceptual features are processed in
modality-specific regions, as opposed to within a unitary conceptual system: this
has implications for how features are integrated to compose concepts (see
McNorgan et al, 2011). Second, instead of making assumptions regarding
features and their distributional statistics across domains, they empirically
derived these statistics using concept-feature norming data (McRae et al., 1997; McRae et al., 2004). These data enabled the discovery of feature correlations across a large set of basic-level concepts, and the exploration of how these statistics influenced human behavior on semantic tasks (McRae et al., 1997; 1999; Cree et al., 1999, 2006). Third, they implemented their cognitive architecture using a type of connectionist model called an “attractor network”: this pairing of cognitive and computational architectures resulted in additional claims regarding how word meaning is computed and provided a rich computational framework within which to explore and test their theory. Behavioral results and model simulations suggest that feature co-occurrence statistics across basic-level concepts contribute to conceptual structure and can explain human judgments of conceptual similarity (Cree et al., 1999; Tyler et al., 2000), typicality (Cree et al., 1999), and property-verification (McRae et al., 1997; 1999). Cree et al. (1999) also pit their theory against classical network models, and provide evidence that their feature-based, non-hierarchical model better explains human performance in semantic tasks. In Chapter 2, I also empirically derive feature statistics across concepts and analyze these structures using a computational approach — however, instead of characterizing the structure of domains of conceptual knowledge using a connectionist architecture, I characterize the structure of individual concepts using the mathematical tools of network science.

1.1.3 Computational models
Whereas cognitive theories of conceptual knowledge are concerned with the types of representations and processes that characterize the conceptual system, the goal of computational models is to define a particular problem and characterize the relevant input-output relationships (see Love, 2015). For example, models can be used to simulate semantic priming tasks: in this case the input would be the concepts and/or properties, and the output would correspond to reaction time or accuracy. For researchers studying semantic memory, connectionist models are the most popular computational framework. These models can naturally capture a compositional conceptual structure and have been implemented in a variety of forms.

Connectionist models became popular in psychology and cognitive science with the proposal of parallel distributed processing (PDP) models (Rumelhart et al., 1986). The PDP model is an artificial neural network in which simultaneous participation of multiple units, and the propagation of information through their weighted connections, gives rise to semantic cognition (for a review see McClelland & Rogers, 2003). Units and links can represent various kind of information — in original PDP models, units represent item-labels (e.g., “canary”), relations (e.g., HAS, IS-A), and features (e.g., WINGS, YELLOW); links represent the strength of association between units. By training these models on item-labels, relations, and a vector of features, the model adjusts the weights.
between units such that representations for each item are developed during a gradual learning process. The fact that some input units represent features suggests a certain degree of compositionality. However, this information is transformed in a hidden layer, which McClelland & Rogers (2003) refers to as the “representational layer” of the network; it is within this layer that semantic representations emerge, along with a similarity space that corresponds to the similarity of their real-world referents. Random damage to this hidden layer results in model performance that mirrors semantic deficits that appear category-specific (Tyler et al., 2000). As noted above, computational architectures sometimes have theoretical implications: hidden layers in connectionist networks are often interpreted to reflect a level of representation that is more abstract than that of the input layer. Whereas Damasio’s model (Damasio, 1989; Meyer & Damasio, 2009) posits that semantic content comprises activation across sensorimotor cortices (representing features) and convergence zones, certain interpretations of the PDP model posit that semantic content emerges in the separate “representational layer”, in which units can no longer be interpreted to represent individual features. There are many versions of this model, however. A recent connectionist model of semantic cognition (Hoffman et al., 2018) includes sets of units that represent sensorimotor features in addition to sets of units that represent verbal inputs and their co-occurrence statistics: separate “hub” and “context” layers integrate this information in order to capture conceptual representations that are influenced by both feature statistics and word co-occurrence statistics. Note that this model deviates theoretically from purely compositional models in which semantic knowledge is captured by direct links between feature representations. Semantic models which represent meaning in terms of word co-occurrence or association statistics will be discussed in more detail below (see section 1.4).

As mentioned above, McRae and colleagues (McRae et al., 1999; Cree et al., 1999, 2006) implemented their feature-correlation theory using a class of connectionist architectures called attractor networks. Attractor networks differ from standard PDP models because they cannot be considered hierarchical: feature-feature correlations are explicitly encoded in weights between feature units, without the addition of a hidden layer (c.f. Cree et al., 1999). These models capitalize on feature associations within semantic memory in order to learn the feature-based representation of individual concepts. More specifically, attractor networks carve a multidimensional feature space into regions corresponding to individual concepts (i.e., attractor basins); the stable state within each of these attractor basins is the learned representation for each concept. Thus, the attractor networks of McRae and colleagues can be considered more purely compositional than the standard PDP models of semantic memory.
In this section I discussed how compositional theories of conceptual knowledge have been implemented in cognitive, neural, and computational frameworks. I will now summarize the evidence for conceptual flexibility, before discussing how this flexibility has been modeled in compositional and non-compositional systems.

1.2 Conceptual flexibility

Intuitively, it makes sense that concepts are flexible. Any given concept, whether domain-level (e.g., LIVING THINGS), superordinate-level (e.g., BIRD), or basic-level (e.g., PENGUIN), can be activated and represented in a variety of ways. Living things can be four-legged or winged, birds can be vibrantly colored or monochrome, penguins can be fuzzy newborns or fully feathered adults. Theories typically adopt a “static” view of concepts, in which conceptual information is stable across instances. But this framework makes it hard to model the conceptual shifts over long and short time-scales that occur during context-dependent concept activation and learning (Casasanto & Lupyan, 2015; Yee & Thompson-Schill, 2016). Flexibility of meaning is also a challenge in the language domain, and is called enriched lexical processing, type-shifting, or coercion (e.g., Pustejovsky, 1998; McElree et al., 2001; Traxler et al., 2005).

General knowledge of a concept can generate many specific conceptual subtypes (e.g., birds can be penguins, pelicans, and ostriches), and can also generate specific exemplars that are bound in space and time (e.g., this particular penguin). Conceptual flexibility is more apparent when we consider the flexible feature fluctuations of a single exemplar in different contexts. For example, while looking for grapes in a supermarket, visual features may be strongly active (e.g., GREEN, ROUND), whereas a cooking context might activate a different set of features (e.g., SWEET, JUICY). The information activated for a concept can also depend on event context: thinking about an egg before and after you crack it to make an omelet results in very distinct representations. Finally, a striking form of conceptual flexibility is the use of non-literal language, when a concept is used to refer to a non-typical referent. The concept BUTTERFLY typically refers to an insect but can also be used to refer to a non-insect (e.g., “The ballerina is a butterfly”) with a different set of features. These forms of flexibility need not be categorically distinct, but making these distinctions helps us map out the conceptual terrain. I will summarize the evidence for property-flexibility, state-flexibility, and referent-flexibility and then will discuss ways in which conceptual models have or have not accounted for this flexibility.

1.2.1 Property flexibility

The crux of conceptual flexibility is that the features activated for a given concept will not be identical across instances: context determines the informational content of a concept. Early cognitive accounts made the distinction between
“context-dependent” and “context-independent” properties. Barsalou (1982) claimed that context-independent properties form the core of a concept and are always active when a concept is processed; context-dependent properties are only activated some of the time. Barsalou (1982) found that participants were always able to verify highly accessible properties (e.g., HAS A SMELL for SKUNK), but were only able to verify less-accessible properties (e.g., CAN BE WALKED UPON for ROOF) when a relevant context was provided (e.g., “The roof creaked under the weight of the repairman”). This work suggested that word-evoked information is not stable across instances but is context-dependent. Casasanto & Lupyan (2015) have taken this further and argued that there is no such thing as a stable conceptual core, and all semantic information is context-dependent. In Chapter 2, this idea of a conceptual “core” will be translated into a network science framework.

More recently, Yee et al. (2012) found that an object’s conceptual color (e.g., GREEN for CUCUMBER) only influenced subjects’ living/non-living judgments if they had previously performed a Stroop color-word interference task, which established color as a relevant feature. This is evidence that task-context influences the extent to which conceptual color information is spontaneously activated. In a neuroimaging study, Hsu et al. (2011) observed increased neural response in left fusiform gyrus (a region sensitive to conceptual color) when participants had to use more detailed color knowledge (e.g., comparing the colors of BUTTER vs. EGG YOLK) relative to when only basic color knowledge was required (e.g., comparing PAPRIKA vs. PENCIL). Taken together, these behavioral and neuroimaging results suggest that the cognitive and neural activation of conceptual features varies across task-contexts.

Eye-tracking studies also provide evidence for flexible conceptual representations. Using a visual world paradigm, Kalénine et al. (2012) revealed that providing an event context modulated the activation of context-relevant functional features. For example, the target WHISK competed with BLENDER in the context of “mixing ingredients,” but competed with SPATULA in the context of “cooking.” This suggests that the blending- and cooking-contexts resulted in different sets of functional properties activated to represent WHISK. The different contexts did not influence competition with thematically related objects (e.g., EGGS). This is further evidence that conceptual features are differentially activated in different contexts.

1.2.2 State flexibility

Some concepts can occur in multiple states, each corresponding to its own distinct set of conceptual properties. Whereas the cases of flexibility above relate to the context-dependent activation of features that are not mutually exclusive (e.g., the same whisk can be used to blend batter and to whip cream), state-
flexibility relates to mutually exclusive sets of features activated by real or imagined event-contexts. For example, the concept EGG can be instantiated as whole or cracked — these two states imply differences in shape, color, and texture properties. A particular egg, bound in space and time, can either be whole or cracked, but not both. Processing multiple states of a conceptual exemplar as it changes in time thus involves differentiating the states’ associated features and keeping the mutually exclusive states of the token distinct. Neuroimaging research on object-state change suggests that regions of visual cortex are recruited to represent distinct states of an object (Hindy et al., 2015), and that left ventrolateral prefrontal cortex is implicated in selecting between competing state representations (Hindy et al., 2012, 2013; Solomon et al., 2015). Tracking a conceptual exemplar as it changes through time recruits neural regions that represent changing conceptual features, as well as regions involved in semantic control that help shift between these representations. Additional neuroimaging research suggests that learned associations between actions and object-states become incorporated into a multistate object representation (Hindy & Turk-Browne, 2015), highlighting the relationship between state-flexibility and conceptual knowledge.

1.2.3 Referent flexibility

When concepts are used creatively — as in figurative language — a concept can acquire novel or atypical referents. Unlike in the cases of property- and state-flexibility above, referent-flexibility results in representations that do not refer to clear instantiations of the concept. This kind of non-literal language is very common and includes standard figurative forms such as similes and metaphors (e.g., Her eyes are diamonds), and also combined concepts like the ones introduced above (e.g., diamond eyes). This flexible use of concepts has also been explained within compositional frameworks.

Lakoff (1973) proposed that metaphorical meaning involves the highlighting of characteristic — rather than defining — conceptual features. Behavioral research reveals that processing a metaphor (e.g., “The ballerina is a butterfly”) involves the activation of relevant properties (e.g., BEAUTIFUL, DELICATE) and/or the suppression of irrelevant properties (e.g., FLIES, HAS ANTENNAE; Gernsbacher et al., 2001; Glucksberg et al., 2001; Solomon & Thompson-Schill, 2017). A related neuroimaging study reveals that prefrontal control mechanisms are recruited to select between these conceptual properties (Solomon & Thompson-Schill, 2017). Successful comprehension of figurative meaning can be understood as a flexible adjustment of conceptual information; theories of conceptual structure must be able to account for this flexible concept use. In Chapter 2, the relationship between a feature-based network structure and figurative language use is explored.
The challenges discussed here are not unique to figurative language. For example, polysemous concepts — for which there are multiple distinct yet related meanings — might also be considered an example of referential-flexibility. For example, BOOK can refer to a bound stack of papers or to the associated informational content; CHICKEN can refer to a live animal or to a menu item. Another example is provided by Anderson & Ortony (1975), in which it is argued that the CONTAINER used to hold apples (e.g., basket) is much different than the CONTAINER used to hold soda (e.g., bottle): the conceptual system might need to navigate through the CONTAINER concept in order to understand what is meant by “container” in each of these contexts. In this sense, polysemy exposes the extent to which conceptual information must be flexibly activated during comprehension (see Hoffman et al., 2018). This flexible activation of conceptual meaning is arguably not restricted to polysemous and figurative language, but pervasive in natural language use.

Conceptual combination is very similar to figurative language in that it involves the flexible selection and integration of conceptual features across concepts (Wisniewski, 1997; Estes & Gernsbacher, 2000; Coutanche et al., 2019). Chapters 3 and 4 use conceptual combination as a tool to study how conceptual features are flexibly modulated in language. After a discussion of previous models of conceptual flexibility, Section 1.4 will be devoted to a review of conceptual combination and how it relates to flexible concept use.

1.3 Previous models of conceptual flexibility

Do existing cognitive, neural, and computational models of conceptual knowledge incorporate conceptual flexibility? The classical network-based models (Collins & Quillian, 1969; Collins & Loftus, 1975) cannot model within-concept flexibility, because concepts are represented as distinct, atomic units. In general, it is hard for atomic (i.e., non-compositional) theories to model flexibility, because it is not clear what information can be manipulated to engender conceptual change (c.f., Carston, 2010). Classical feature-based, compositional theories do not address flexibility per se, but there are feature-based theories of conceptual combination as I will explain in more detail below (Smith et al., 1988). The conceptual structure account (e.g., Tyler & Moss, 2001) and feature correlation account (e.g., McRae et al., 1997) represent concepts as static, distributed patterns of activity across semantic features, though the need to capture conceptual variation is addressed (O’Connor et al., 2009).

1 The spreading activation model of Collins & Loftus (1975) might be able to capture flexible activation patterns across the semantic system as a whole, but this is not the kind of flexibility under discussion.
Attractor networks like the ones used by McRae and colleagues have been used to capture aspects of conceptual flexibility. For example, O’Connor et al. (2009) trained an attractor network to learn both basic-level concepts (e.g., HAMMER) and superordinate concepts (e.g., TOOL) in a non-hierarchical feature space. The basic-level concepts were trained in a one-to-one mapping of wordform to exemplar (e.g., “hammer” paired with a vector of HAMMER features). The superordinate concepts were trained in a one-to-many mapping, in which a wordform was paired with different feature vectors on different trials (e.g., “tool” separately paired with HAMMER, WRENCH, and SCREWDRIVER features). As a result, the model could capture variability in the representation of superordinate concepts. In another study, Rodd et al. (2004) trained an attractor network on homonyms (e.g. “bank”) and polysemes (e.g. “twist”). After training, it was observed that the attractor basins for homonyms are non-overlapping, whereas attractor basins for polysemes overlap to form wider attractor basins in feature space. This prior work suggests ways in which flexible conceptual meaning can be embedded within computational models.

A cognitive theory that can capture conceptual flexibility is Barsalou’s (1999) perceptual symbols system (PSS) theory. In this framework, the set of modality-specific features activated during perception of a particular item is reactivated during conceptual processing. Perceptual symbols (e.g., YELLOW, SWEET) that are activated during perception are integrated to form a “simulator” for each concept (e.g., BANANA). These simulators can activate subsets of perceptual symbols in order to generate a potentially innumerable number of context-specific “simulations” (Barsalou et al., 2003). The organization of conceptual information within the PSS model can provide an explanation of how some kinds of conceptual flexibility emerge, such as token variability, polysemy, and metaphorical language. However, this theory is not computationally well-specified.

A neural theory that can capture conceptual flexibility is the “hub-and-spoke” model in cognitive neuroscience. Proponents of this model claim that representing conceptual knowledge involves “hubs” in bilateral anterior temporal cortex that integrate information from modality-specific neural regions (the “spokes”). The controlled semantic cognition framework updates the hub-and-spoke view, such that semantic representation and control are mediated through different networks; the ATL hub is involved in representing generalizable conceptual knowledge, and separate regions represent task contexts (Ralph et al., 2017). These ideas were further developed into a connectionist model of semantic cognition that directly accounts for conceptual flexibility (Hoffman et al., 2018). In this model, a hub-and-spoke architecture (Rogers et al., 2004) is enriched with a “buffer” that incorporates prior context. The architecture includes two hidden layers: a “hub” layer integrates information from sensorimotor and
verbal input units, and a “context” layer captures the hub activation patterns elicited by the previous input. The fully recurrent architecture enables the model to generate context-sensitive representations and is thus able to capture the multiple meanings of homonyms as well as the context-dependent graded shifts of word meanings that occur in natural language use. The neural and computational models inspired by a hub-and-spoke architecture are intriguing and are valuable additions to research on conceptual flexibility. However, the inclusion of word co-occurrence statistics and amodal semantic hubs result in representations that are transformed away from conceptual features.

Related approaches in “distributional semantics” are popular in cognitive and computational models of language use — these models capturing the meaning of words in terms of their co-occurrence or association statistics. In this class of models, a semantic space is defined based on word statistics extracted from large text corpora (see Mitchell & Lapata, 2010). Researchers using models of distributional semantics have attempted to predict the meaning of combined concepts. These experiments will be summarized below in the discussion of conceptual combination.

1.4 Conceptual combination

Theories of conceptual knowledge and flexibility can be illuminated by studies of conceptual combination. When concepts combine with each other, we can observe flexibility of properties, event-states, and referents — as in green pumpkin, cracked egg, and diamond eyes, respectively. Conceptual combination involves processes of feature selection and feature integration, in which conceptual information is flexibly modulated to result in an appropriate representation of the combined concept.

Early cognitive theories of conceptual combination include the attribute inheritance model (Hampton, 1987; 1988), the selective-modification model (Smith et al., 1988), and the concept specialization account (Murphy, 1988; Wisniewski & Gentner, 1991). In Hampton’s attribute inheritance model, each concept is represented as a prototype comprising a list of independent features weighted by importance. Comprehending combined concepts involves constructing a composite prototype of the constituents by merging their feature lists, and then adjusting this composite list to satisfy various constraints (e.g. necessity, impossibility, coherence). The predictions of this model were consistent with the features explicitly generated in a conceptual combination task (Hampton, 1987). However, a limitation of this model is that it only applies to conjunctive noun-noun combinations of the form “An X that is also a Y” (e.g., “A pet that is also a bird.”) A satisfying theory of conceptual combination will be able
to explain non-conjunctive noun-noun (e.g., *apartment dog*) and adjective-noun (e.g., *small dog*) combinations.

The selective-modification model (Smith et al., 1988) directly targets the feature modulations observed in adjective-noun combinations. Concepts are represented as sets of slots (e.g., *COLOR*) and their fillers (e.g., *RED*, *GREEN*). Each filler refers to a potential feature and has a salience value that indicates its strength for that concept (e.g., *RED* has higher salience than *BROWN* for *APPLE*). In adjective-noun combinations (e.g., “brown apple”), the adjective selects the appropriate slot (e.g., *COLOR*) and filler (e.g., *BROWN*) to be modified in the noun concept. The salience of that feature and the diagnosticity of the slot is increased in the resulting combined concept. This model is appealing because it makes well-formalized predictions about how concepts combine, and it successfully captures similarity and typicality judgments (Smith & Osherson, 1984; Smith et al., 1988).

A limitation of this model is its inability to capture the interdependence of features. Conceptual features interact with each in ways that influence the comprehension of combined concepts. For example, people interpret “wooden spoons” to be large and “metal spoons” to be small, suggesting that *MATERIAL* and *SIZE* features are associated within the *SPOON* concept (Medin & Shoben, 1988). Feature associations are an important component of conceptual structure and will influence the outputs of conceptual combination (Wisniewski & Gentner, 1991; Sloman et al., 1998). The goal of Chapter 2 is to capture feature associations within concept network models and determine whether these structures can predict aspects of flexible concept use.

A related criticism of the selective-modification model is that it cannot capture the flexibility and context-dependence of adjective meaning. The same adjective can have different effects when paired with different nouns (e.g., *fresh vegetable*, *fresh shirt*, *fresh idea*; Murphy & Andrew, 1993), or modulate features to various degrees (e.g., *red face*, *red truck*, *red fire*; Halff et al., 1976). Proponents of the concept specialization model claim that comprehension of combined concepts cannot be a function of the constituent concepts alone, but that additional knowledge is required (Cohen & Murphy, 1984; Murphy, 1988; Wisniewski & Gentner, 1991). The appeal to “world knowledge” addresses the inherent flexibility in the combination process, but this world knowledge is vaguely defined and therefore hard to incorporate into a well-defined model. The goal of Chapters 3 and 4 is to capture some of the flexibility in the conceptual combination process using well-defined models of feature representations.

Though characterized by a very different theory of semantic knowledge, the field of distributional semantics (referenced above in Section 1.1.1) has proposed well-defined models of how concepts combine in complex phrases. Based on the assumption that words with similar meanings appear in similar verbal contexts,
words are represented as vectors based on their co-occurrences with other words in a large text corpus (e.g., Landauer & Dumais, 1997; Mitchell & Lapata, 2008; 2010; Steyvers & Griffiths, 2007). Vectors representing individual words can then be combined into phrases using different classes of functions (e.g., additive, multiplicative). Mitchell & Lapata (2010) tested many different combinatorial functions within a vector-based semantic space. They compared additive models (e.g., simple, weighted) and multiplicative models (e.g., simple, tensor product) in cases of adjective-noun, noun-noun, and verb-noun combinations, and found the best success for weighted-additive, simple multiplicative, and dilation models (see Mitchell & Lapata, 2010). These authors also tested combinatorial semantic functions using a probabilistic topic model of semantic space (Steyvers & Griffiths, 2007) and similarly report success of a simple multiplicative model. Another model is proposed by Baroni & Zamparelli (2010) in which adjectives are treated as linear functions over nouns. The meaning of a noun is captured in a vector-based representation, and each adjective is represented as a matrix which is used to transform these noun vectors. Baron & Zamparelli (2010) report increased success of their model relative to additive and multiplicative models and claim that their model is able to capture aspects of flexible adjective meaning (e.g., green chair vs. green initiative). In sum, approaches in distributional semantics provide techniques that can be used to predict the representations of combined concepts.

In cognitive neuroscience studies of conceptual combination, there have been few attempts to predict the neural representations of combined concepts. An fMRI study by Chang et al. (2009) analyzed distributed neural responses to adjective-noun phrases (e.g., "soft bear", "sharp knife") using the compositional models of distributed semantics described above, and report that the neural patterns evoked by adjective-noun combinations were best predicted by a multiplicative combination of the individual adjective and noun patterns. Two additional fMRI studies examined multivoxel responses to combined concepts in the left anterior temporal lobe (LATL), a region implicated in conceptual knowledge and combination (Baron et al., 2010; Baron & Osherson, 2011). Both of these studies report that multivoxel responses to combined concepts in LATL can be predicted by combining the patterns of the constituent concepts. For example, Baron & Osherson (2010) report that the GIRL-pattern is predicted by

\[ \text{GIRL-pattern} = \text{GIRL} \times \text{GIRL} \]

In these probabilistic topic models, words are not represented as points in a high-dimensional space but as probability distributions over a set of topics. In a related approach, Chapter 4 will explore the possibility of representing concepts as probability distributions over features.
additive and multiplicative combinations of the YOUNG- and WOMAN-patterns in LATL.

Rather than trying to predict the neural representations of combined concepts, much work in cognitive neuroscience has revealed the particular brain regions that are recruited in conceptual combination tasks. The LATL has been implicated in conceptual combination tasks using both MEG (Bemis & Pylkkänen, 2011; 2012; 2013) and fMRI methods (Baron et al., 2010; Baron & Osherson, 2011; Boylan et al., 2017), and its role has been characterized as relating to feature attribution or object-concept specification (Westerlund & Pylkkänen, 2014; Boylan et al., 2015; 2017). Another important region is the angular gyrus, which appears sensitive to the plausibility of adjective-noun combinations (Price et al., 2015) and combinations describing events or relations (Boylan et al., 2015; 2017). The goal of Chapter 4 is to explore which neural regions are specifically involved in the flexible modulation of features during comprehension of combined concepts.

1.5 Current approach

The current approach is to explore global and local aspects of conceptual structure that relate to flexible concept use. In Chapter 2, I construct novel feature-based concept networks and determine which aspects of global network structure might relate to various measures of flexible concept use (e.g., semantic diversity, figurative language). I zoom in to examine local structure of individual features in Chapters 3 and 4, and explore the kinds of feature representations that predict the flexible modulation of conceptual information during comprehension of combined concepts. The fMRI study in Chapter 3 assumes a distributed neural representation of conceptual features, and tests whether the flexible neural activation of features during comprehension of adjective-noun phrases can be predicted by either feature surprisal or feature uncertainty. Following up on this study, Chapter 4 embeds feature uncertainty in probabilistic models of feature modulation to explore adjective-noun comprehension in behavior and in the brain. Neural regions previously implicated in conceptual combination are explored in order to determine which regions are specifically sensitive to the flexible modulation of conceptual information.
2: IMPLEMENTING A CONCEPT NETWORK MODEL

This chapter was previously published as an article in Behavior Research Methods: Solomon, S.H., Medaglia, J.D., Thompson-Schill, S.L. (2019). Implementing a concept network model. Behav Res.

2.1 Introduction

The APPLE information evoked by “apple pie” is considerably different from that evoked by “apple picking”: the former is soft, warm, and wedge-shaped, whereas the latter is firm, cool, and spherical. If you scour your conceptual space for APPLE information, you will uncover the knowledge that apples can be red, green, yellow, or brown when old; that they can be sweet or tart; that they are crunchy when fresh and soft when baked; that they are naturally round but can be cut into slices; that they are firm, but mushy if blended; that they can be found in bowls, in jars, and on trees. Despite the complexity of this conceptual knowledge, we can generate an appropriate APPLE instance, with the appropriate features, based on the context we are in at the time. In other words, the APPLE concept can be flexibly adjusted in order to enable a near-infinite number of specific and appropriate APPLE exemplars. This flexibility enables concepts to be represented in varied and fluid ways, a central characteristic of the semantic system.

The concept APPLE can be instantiated as a Granny Smith or as a Macintosh, and either one can easily be brought to mind. The fact that a single conceptual category has many distinct subordinate types that differ from each other is a basic form of conceptual variation that has been embedded within hierarchical semantic models (e.g., Collins & Quillian, 1969). But even a representation of a single instance of APPLE can be flexibly adjusted: activated properties might be RED and ROUND while shopping, whereas they might be SWEET and CRISPY while eating. A concept can also be represented in varied states, each with their own distinct features: the representation of an APPLE is FIRM versus SOFT before and after baking, and SOLID versus LIQUID before and after juicing. Conceptual flexibility is further evidenced in the frequent non-literal use of concepts: one should stay away from “bad apples” and should not “compare apples with oranges;” and, one can use concepts fluidly in novel analogies and metaphors.

Typically, theories assume a “static” view of concepts, in which conceptual information is stable across instances. But this framework makes it hard to model the conceptual shifts over long and short time-scales that occur during context-dependent concept activation and learning (Casasanto & Lupyan, 2015; Yee & Thompson-Schill, 2016). By “context” we refer to events or situations, whether in the physical environment or in language, that could influence the ways in which conceptual information is activated and represented. The flexibility of meaning is also a challenge in the language domain, and is referred to as enriched lexical processing, type-shifting, or coercion (e.g., Pustejovsky, 1998; McElree et al.,
2001; Traxler et al., 2005). Though conceptual flexibility is a pervasive phenomenon, it poses a formidable challenge: how is conceptual information organized to enable this flexibility?

We are particularly interested in the structure of individual concepts (e.g., APPLE, SNOW), rather than the structure of semantic space more broadly. This latter pursuit — the modeling of semantic space — has already been approached from various theoretical orientations and methodologies. Some theoretical approaches claim that the meaning of a concept can be decomposed into features and their relationships with each other (e.g., Smith et al., 1974; Tversky, 1977; McRae et al., 1997; Sloman et al., 1998; Tyler & Moss, 2001). For example, the “conceptual structure” account (Tyler & Moss, 2001) represents concepts as binary vectors indicating the presence or absence of features, and argues that broad semantic domains (e.g., ANIMALS, TOOLS) differ in their characteristic properties and in their patterns of property-correlations (e.g., HAS-WINGS and FLIES tend to co-occur within the ANIMAL domain). Models that represent basic-level concepts in terms of their constituent features are valuable because they can be implemented in computational architectures such as parallel distributed processing models and other connectionist models. For example, the “feature-correlation” account (e.g., McRae et al., 1997; 1999; McRae, 2004) pairs empirically-derived conceptual feature statistics with a type of connectionist model called an attractor network: property statistics characterize the structure of semantic space, and the model can leverage these statistics to settle on an appropriate conceptual representation given the current inputs (Cree et al., 1999; Cree et al., 2006).

However, most instantiations of feature-based models represent individual concepts with sets of features that are static and unchanging — a clear limitation if one aims to incorporate flexibility into conceptual structure. There are some recent connectionist models that aim to incorporate context-dependent meaning (Hoffman et al., 2018), and the flexibility and context-dependence of individual features has been addressed in prior work (Barsalou, 1982; Sloman et al., 1998). For example, Sloman et al. (1998) modeled the pairwise dependencies between features in order to ascertain the mutability or immutability of features. The authors claim that a feature is immutable if it central to a concept’s structure: it is harder to imagine a concept missing an immutable feature (e.g., a robin without bones), than a mutable feature (e.g., a jacket without buttons). The authors argue that modeling a concept in terms of formal pairwise relationships makes it possible for concepts to be structured as well as flexible. While the goal of these and other researchers has been to characterize the role of individual conceptual features (Sloman et al., 1998; Devlin et al., 1998; Cree et al., 2006; Tyler & Moss, 2001; Sedivy, 2003), our present goal is to examine whether a feature-based conceptual structure can shed light on the flexibility of a concept as a whole.
Another way to model conceptual knowledge is to use a network to capture the relationship *between* concepts in language. The use of networks to model semantic knowledge has a well-established history. The early “semantic network” models (Collins & Quillian, 1969; Collins & Loftus, 1975) represent concepts as nodes in a network; links between these nodes signify associations between concepts in semantic memory. These networks capture the extent to which concepts are related to other concepts and features, and can model the putatively hierarchical nature of conceptual knowledge. Though these models are “network-based”, they are so in a rather informal way. On the other hand, network science, a mathematical descendent of graph theory, has developed a rich set of tools to study networks in a formal, quantitative framework (Barabási, 2016). For example, current network approaches characterize relationships between concepts in terms of their word-association strengths or corpus-based co-occurrence statistics. Word co-occurrence statistics can be extracted from text corpora and have been used to create probabilistic models of word meanings (Griffiths et al., 2007), to represent semantic similarity (Landauer & Dumais, 1997), and to characterize the structure of the entire lexicon (e.g., WordNet; Miller & Fellbaum, 2007). In a similar approach, word association data has been used to capture and analyze the structure of semantic space (Steyvers & Tenenbaum, 2005; Van Rensbergen et al., 2015; De Deyne et al., 2016).

In network science, units and links are referred to as “nodes” and “edges”, respectively, and the pattern of connections between nodes can be precisely described, revealing patterns of network organization. Nodes can represent any number of things (e.g., cities, people, neurons), depending on the system being modeled; edges can likewise represent a range of connection types (e.g., roads, friendship, synapses). Many diverse systems have been described in network science terms, including the world wide web (e.g., Cunha et al., 1995), social communities (Wellman, 1926), the nervous system of *Caenorhabditis elegans* (Watts & Strogatz, 1998), and many others (see Boccaletti et al., 2006). Here we will summarize how network science has been applied, and can be further extended, to the study of conceptual knowledge.

Network structure can be characterized at different levels of organization. For example, the large-scale organization of a network (i.e., topology) can be characterized as a “regular”, “random”, or “small-world” structure (Fig. 1 A-C; Watts & Strogatz, 1998). In regular networks, each node is connected to its *k* nearest neighbors; in random networks, nodes are randomly connected to each other. Regular networks result in long path lengths and high local clustering (modular processing), whereas random networks result in short path lengths and minimal clustering (integrated processing). Between these two extremes is the small-world network, which contains high-clustering as well as a few random, long-range connections: this results in the “small world” phenomenon in which each node is connected to all other nodes with relatively few degrees of separation. A small-world topology thus maximizes efficient spread of
information, enables both modular and integrated processing, and supports network complexity (Watts & Strogatz, 1998; Bassett & Bullmore, 2006). Much work has revealed that naturally evolving networks have small-world topology (Bassett & Bullmore, 2006), including functional brain networks (Salvador et al., 2005; Bassett et al., 2011) and language networks (e.g., Steyvers & Tenenbaum, 2005). These systems exhibit small-world topologies presumably because this structure facilitates local “modular” processing as well as easy communication via a few long range connections.

The semantic network approaches described above use nodes to represent individual words, and edges to represent their co-occurrence or associations. Once modeled in this way, network structure can be quantitatively analyzed and related to other phenomena. As mentioned above, it has been suggested that human language networks exhibit small-world properties (Steyvers & Tenenbaum, 2005; i Cancho & Sole, 2001). Additionally, semantic networks appear to exhibit an “assortative” structure, meaning that semantic nodes tend to have connections to other semantic nodes with similar characteristics (e.g., valence, arousal, concreteness; Van Rensbergen et al., 2015). A spreading activation model applied to these word-association networks makes accurate predictions of weak similarity judgments; for example, between the unrelated concepts of “teacher” and “cup” (De Deyne et al., 2016). Further, Steyvers & Tenenbaum (2005) report that the degree of a word in a language network (i.e., how many links it has to other word nodes) predicts aspects of language development and processing — a high degree word is likely learned at a younger age and engenders faster reaction times on a lexical decision task. These network models are valuable because semantic structure can be analyzed using a rich set of network science tools. However, current network-based implementations do not provide the internal conceptual structure that is necessary — we argue — to model conceptual flexibility. In other words, it is hard to provide a model of conceptual flexibility (in the sense described above) when the features that are being flexibly adjusted are not explicitly represented.

We believe that a feature-based conceptual framework paired with network science techniques provides a platform on which conceptual flexibility can be quantified and explored. Here we introduce a new approach in which concepts are represented as their own feature-based networks, and we work through an example as a proof-of-concept. We create concept-specific networks for 15 concepts (e.g., CHOCOLATE, GRASS, KNIFE), in which nodes represent conceptual features (e.g., BROWN, GREEN, METAL, SHARP, SWEET) and edges represent how those features co-occur with each other within each concept. The creation of such networks thus requires the calculation of within-concept feature statistics, which describe how a concept’s information may be appropriately adjusted to form valid, yet varied, concept representations. Though here we are interested in analyzing the structure of basic-level concepts, these concept network methods could theoretically be applied at any level of the conceptual hierarchy. Our
specific goals here are to (1) show that creation of such networks is possible, (2) confirm that these networks contain concept-specific information, and (3) demonstrate that these networks permit the extraction of measures that relate to conceptual flexibility.

We have hand-picked a selection of measures to extract and analyze from our concept networks. As mentioned above, small-world networks (Fig. 1B) are characterized by high network clustering such that a node’s neighbors also tend to be neighbors with each other. Previous work has found a relationship between the clustering within semantic networks and individual differences in creativity (Kenett et al., 2014) — because creativity relates to flexible conceptual processing, the *clustering coefficient* is one of our measures of interest. In small-world networks, this clustering paired with random connections results in network modules, which are communities of nodes with dense connections between them. *Modularity* is a formal measure that captures the extent to which a given network can be partitioned in this way (Fig. 1D). A network with a modular structure is able to activate distinct, specialized sets of nodes; because this might translate into a concept’s ability to activate distinct sets of features, modularity was another measure of interest. In modular networks, each node can also be characterized in terms of its *diversity* of connections across network modules (Fig. 1E). Some nodes may have links within only one module, whereas others may have links that are highly distributed across different network modules. Because related measures are often used to define network hubs that support flexible network processing (van den Heuvel & Sporns, 2013; Sporns, 2104), we were interested in exploring the relation between network diversity and flexible concept use. Another kind of network topology is *core-periphery* structure (Fig. 1F), in which a network is characterized by one highly-connected core and a sparsely connected periphery. Core-periphery organization, originally observed in social networks (Borgatti & Everett, 2000), has recently been applied to functional networks in neuroimaging data (Bassett et al., 2013). A core-periphery structure in a concept network would reflect one set of highly associated features (i.e., core) but also a substantial collection of features that are weakly associated with one another (i.e., periphery). We included core-periphery structure as a measure of interest because we hypothesized that the “stiff” core and/or “flexible” periphery of a concept network (Bassett et al., 2013) might relate to flexible concept use.
In this proof-of-concept, we will extract measures of network organization from concept-specific networks (i.e., clustering, modularity, core-periphery, diversity) and explore what these structural characteristics might predict about how a concept is used. Conceptual flexibility manifests when a concept is recruited to represent varied subordinate exemplars, when a concept word is used in a variety of language contexts, and when a concept is differentially activated depending on task context. We therefore aimed to relate our network measures of interest to a measure of semantic diversity (SemD; Hoffman et al., 2013) calculated from text-based statistics. We also collected data on two tasks related to conceptual flexibility — a figurative language task (comprehension of novel similes) and a widely used measure of creative cognition, the Alternative Uses Task (AUT) — to explore whether network structure relates to how a concept is flexibly used in different task contexts. Here we present one variation of the concept network approach, implementing a particular set of methodological decisions on a particular set of concepts, in order to show the potential of this framework to provide new ways to characterize the structure and flexibility of conceptual knowledge.

Figure 1: A visualization of network topologies and measures. Networks are defined in terms of nodes (circles) and edges (lines). Network topologies fall into three main categories: (A) regular, (B) small-world, and (C) random (Watts & Strogatz, 1998). Most naturally evolving networks exhibit small-world topology, including neural networks and language networks. Regular and small-world networks have high clustering. (D) Modularity reflects the extent to which a network can be partitioned into a set of densely-connected “modules”, represented here in distinct colors. (E) Some nodes participate in multiple modules, reflecting a diversity of connections: this is captured in a “diversity coefficient.” A diverse node (yellow) participates in multiple modules (green, purple), whereas other nodes (grey) do not exhibit these diverse connections. (F) A network has strong core-periphery structure if it can be characterized in terms of a single densely-connected “core” (yellow) and a sparsely-connected “periphery” (grey).
2.2 Methods

2.2.1 Network Methods
Our goal is to construct feature-based networks that capture each concept’s specific constellation of features and the ways those features relate to each other. There are, however, many ways one could create such networks. Here we walk through one possible instantiation of this method to reveal the feasibility of this approach, and to suggest the kinds of analyses that could be used to examine the relationship between concept network topology and conceptual flexibility. We hope that future researchers interested in conceptual knowledge will be able to use, improve, and expand upon these methods. Our data and code are available online.

We collected data in two rounds, and we refer to these data as Set 1 and Set 2. We collected data for five concepts in Set 1 as a first attempt to construct concept networks. Once we established the success of these methods, we collected data for another 10 concepts in Set 2. The concepts in Set 1 included: CHOCOLATE, BANANA, BOTTLE, TABLE, and PAPER. The concepts in Set 2 included: KEY, PUMPKIN, GRASS, COOKIE, PICKLE, KNIFE, PILLOW, WOOD, PHONE, and CAR. When statistics are reported separately for the two sets, we report Set 1 followed by Set 2. Once the networks are constructed and we analyze network measures and their relation to other conceptual measures, the sets are no longer treated separately, and each concept is treated as an item (N=15). We use Spearman’s rank correlation in all correlational analyses due to this small sample size.

2.2.1.1 Network Construction
The first step was to define our nodes. Since our nodes represent individual conceptual properties, we compiled a list of properties that could be applied to all of our concepts within each set. Participants (N=66, N=60) were recruited from Amazon Mechanical Turk (AMT) and were asked to list all of the properties that must be true or can be true for each concept. It was emphasized that the properties do not have to be true of all types of the concept. Participants were required to report at least 10 properties per concept, but there was no limit on the number of responses they could provide. Once these data were collected, we organized the data as follows. For each concept, we collapsed across different forms of the same property (e.g., “sugar”, “sugary”, “tastes sugary”), and removed responses that were too general (e.g., “taste”, “color”). This was a highly data-driven approach; however, see Bootstrap Analysis for an analysis of robustness across properties. For each concept, we only included properties that were given by more than one participant. We then combined properties across all concepts to create our final list of N properties (N=129, N=276) that were represented as nodes in our concept networks.
The same AMT participants that provided conceptual properties also provided subordinate concepts (from now on referred to as “subordinates”) for each of the concepts. For each concept, participants were asked to think about that object and all the different kinds, forms, types, or states in which that object can be found. Participants were required to make at least five responses, and could make up to 15 responses. For each concept, we removed subordinates that corresponded to a property for that concept (e.g., “sweet chocolate”), subordinates that were highly similar to other subordinates (e.g., “white chocolate chip cookie”, “chocolate chip cookie”) and responses that were too specific, including some brand names (e.g., “Chiquita banana”). Though this was a data-driven approach, there was some degree of subjectivity in the final subordinate lists; see Bootstrap Analysis for an analysis of robustness across subordinates. In Set 1, we only included responses that were given by more than one participant; due to the increased number of participants and responses in Set 2, we included responses that were given by more than two participants. In both sets, we ended up with a set of $K$ subordinates for each concept ($K: M=17$, $SD=3.14$). The included and excluded subordinates for all concepts are presented in Supplementary Table 1.

A separate set of AMT participants ($N=198$, $N=108$) was presented with one subordinate of each of the concepts in random order (e.g., “dark chocolate”, “frozen banana”) and was asked to select the properties that are true of the

![Figure 2: Visualizing the chocolate network. (A) The chocolate concept can be broken down into a range of subordinates, which can each be defined as a property vector (columns). Each property can also be defined as a vector (rows), which can be used to calculate within-concept property relationships. Only a small set of subordinates and properties are shown here for simplicity. (B) A simple schematic of the chocolate network that reveals a selection of potential property relationships. Certain properties might cluster together in the chocolate network, for example EDIBLE, SWEET, BROWN, CREAMY, and MESSY, LIQUID, HOT. (C) The actual chocolate network we constructed based on the empirical property statistics. Our constructed chocolate network was binarized (threshold=90%) in order to reduce the number of properties to ease visualization. Properties are arranged in order of degree (number of links), from low degree (white) to high degree (blue). Image generated using cytoscape (Shannon et al., 2003).](image-url)
specific subordinates (see Fig. 2A). The full list of \( N \) properties was displayed in a multiple-choice format. For each subordinate, responses were combined across participants; we thus know, for each subordinate, how many participants reported each of the \( N \) properties. In the networks we report here, we have used weighted subordinate vectors in which values indicate the percentage of subjects that reported each property. However, in order to reduce noise, we only included a weight for a subordinate-property if it was reported by more than one participant; if only one participant reported a particular property for a particular subordinate, the weight \( \approx 0 \). For each concept we excluded properties that were not present in any of its subordinates, resulting in a smaller set of \( Nc \) properties that were present in \( \leq 3 \) subordinates \((Nc: M=126, SD=32.2)\). Each concept’s data thus included a set of \( K \) subordinates, each of which corresponds to a \( Nc \)-length vector that indicates each property’s weight in that subordinate.

Importantly, these data for each concept can also be considered a set of \( Nc \) properties, each corresponding to a vector indicating that property’s weight in each of the subordinates. For example, if a concept was described by 10 subordinates \((K)\) and 100 meaningful properties \((Nc)\), we have 100 10-element vectors, each of which represents the contribution of a single property across subordinates of the concept. The premise behind this concept-network construction is that the ways in which these patterns of property contributions relate to each other, within a single concept, may be an important aspect of conceptual structure. Our networks will thus capture the pairwise similarities between properties, that is, between the \( Nc \) \( K \)-element vectors. In order to do this, many different distance metrics can be used (e.g., Euclidean, Mahalanobis, cosine); we used the \( pdist() \) function in MATLAB which includes many distance-measure alternatives. In the analysis and results we report here, we constructed our networks based on Mahalanobis distance, a measure suited for high-
dimensional data and which takes the variance between subordinates and correlations between subordinates into account. However, there are many other options, the choice of which might depend on other analysis decisions. For example, if the subordinate concepts are binary instead of weighted, a “matching” measure such as Jaccard distance might be more appropriate.

First, the distance between each of the $Nc \times K$-element vectors was calculated. This results in a square, symmetrical $Nc \times Nc$ matrix which contains the distance between each pair of properties. These values were scaled between 0 and 1, and converted to a similarity measure by subtracting these values from 1. We thus created a network for a single concept that captures pairwise property-property similarities; this represents the patterns of property relationships across subordinates within a given concept (see Fig. 2). Here we used weighted networks, where edges represent similarity measures between 0 and 1, though it is also possible to use unweighted networks by binarizing these similarity values according to a given threshold. We repeated this (weighted) network construction process for each of the 15 concepts. These final networks were then analyzed using standard network science methods (see Network Analysis).

A simple measure of concept stability. The subordinate property data for each concept enabled us to calculate a simple measure of conceptual stability that did not involve treating concepts as networks. For each concept, we counted the number of properties that were represented across all of that concept’s subordinates (weighted value greater than zero). We then divided this number by $Nc$ in order to calculate the proportion of possible properties that were universally consistent for that concept. We interpreted this measure as a measure of conceptual stability, because higher values indicate that a large number of properties are not variable across conceptual instances. We refer to this measure as simple stability, and consider it to be an inverse measure of conceptual flexibility.

2.2.1.2 Classification Analysis

Our goal is to extract concept-specific measures from our networks, and this goal is only justified if the network structures themselves are concept-specific. Even though different sets of data contributed to the different concept networks, it is not necessarily the case that the resulting networks would differ from each other. It could theoretically be the case that property-relationships are consistent across the entire semantic domain; indeed, this is the premise of the neural network models of semantic knowledge created thus far (e.g., McClelland & Rogers, 2003; Cree et al., 1999; 2006). However, our goal is to capture concept-specific property-relationships, and so our first task was to test whether we succeeded in this goal.

If our concept network models capture concept-specific information, the networks should be able to successfully discriminate between new concept exemplars.
Exemplar data were generated from sets of photographs for each concept (see Fig. 3); all subordinates were represented. AMT participants ($N=60$, $N=30$) were shown one image per concept, were asked to imagine interacting with this object in the real world, and to consider what properties it has. The full list of $N$ properties was displayed in multiple-choice format, and participants were asked to select the properties that they believed applied to the object in the image. Individual participants’ responses to each exemplar image were represented as $N$-length property vectors and were used as test data in the classification analysis. Test data comprised 300 property-vectors (Set 1: 60/concept, Set 2: 30/concept); classification analyses were run separately for Set 1 and Set 2.

By performing eigendecomposition on each adjacency matrix (i.e., concept network) we can assess the extent to which a vector is expected given an underlying network structure (e.g., Medaglia et al., 2017; Huang et al., 2018). For each adjacency matrix $A$, $V$ is the set of $Nc$ eigenvectors, ordered by eigenvalue. $M$ is the number of ordered eigenvectors to include in analysis, and designates a subset of $V$. For each eigenvector $v$, we find the dot product with signal vector $x$, which gives us the projection of $x$ on that dimension in the eigenspace of $A$. That is, it gives us an “alignment” value for that particular signal and that particular eigenvector. We can include all eigenvectors in $M$ by taking the sum of squares of the dot products for each eigenvector. The alignment value for each signal is defined as:

$$\tilde{x} = \sum_{i=1}^{M} (v_i \cdot x)^2$$

where $x$ is a property vector, $M$ is the number of eigenvectors to include in alignment (sorted by eigenvalue), $v_i$ is one of $M$ eigenvectors of the adjacency matrix, and $\tilde{x}$ is the scalar alignment value for signal $x$ with adjacency matrix $A$, given the eigenvectors $1$-$M$. In our case, signal $x$ is a property vector corresponding to a particular exemplar image (e.g., Fig. 3), which we align with each of the concept networks. Each exemplar was restricted to the properties included in each concept model before transformation; that is, exemplar data ($x$) were reduced to $Nc$-length vectors. The concept network that resulted in the highest alignment value ($\tilde{x}$) was taken as the “guess” of the classifier; each exemplar was either classified correctly (1), or incorrectly (0). We averaged these data across all exemplars to calculate the average classifier accuracy.

To calculate a baseline measure of classification accuracy, we created traditional vector models for each concept. These models were similar to those used elsewhere in the literature (Tyler & Moss, 2001; McRae et al. 1997; 1999; 2004). For each concept, we averaged the $K$ subordinate vectors resulting in an $Nc$-length vector containing mean property strength values. Each concept’s traditional vector model and network model contained the same conceptual properties. We ran a separate classification analysis using these traditional models and a correlational classifier. Each exemplar property-vector was
correlated with each of the traditional concept vector models; the concept model that resulted in the highest correlation value was taken as the guess of the classifier. We calculated average measures of classifier performance using the same methods described above.

2.2.1.3 Network Analysis

We extracted network metrics from our concept networks using the Brain Connectivity Toolbox (Rubinov & Sporns, 2010). The set of nodes in each network is designated as \( N \), and \( n \) is the number of nodes. The set of links is \( L \), and \( l \) is the number of links. The existence of a link between nodes \((i,j)\) is captured in \( \delta_{ij} = 1 \) if a link is present and \( \delta_{ij} = 0 \) if a link is absent. The weight of a link is represented as \( w_{ij} \) and is normalized such that \( 0 \leq w_{ij} \leq 1 \). \( w \) is the sum of all weights in the network. The network metrics we extracted included clustering coefficients, modularity \( \omega \), core-periphery structure, and diversity coefficients (Fig. 1), for the reasons described above.

The clustering coefficient captures the “cliquishness” of a network, that is, the extent to which a node’s neighbors are also neighbors of each other. The clustering coefficient is calculated for each node individually \( C_i \), by calculating the percentage of potential pairwise connections among the neighbors of node \( i \). A “triangle” is formed when node \( i \) is linked to \( j \) and \( h \), and \( j \) and \( h \) are also linked to each other; the number of existing triangles can be calculated for each node \( t_i \) which is used to calculate the proportion of possible triangles that exist for each node. This proportion is averaged across nodes to result in the clustering coefficient \( C \) for a network (Rubinov & Sporns, 2010; this can also be calculated for weighted networks):

\[
t_i = \sum_{h,j \in N} a_{ij} a_{ih} a_{jh} \quad (2)
\]

\[
C = \frac{1}{n} \sum_{i \in N} C_i = \frac{1}{n} \sum_{i \in N} \frac{2t_i}{k_i(k_i-1)} \quad (3)
\]

Modularity \( \omega \) is a metric that describes a network’s community structure. We can attempt to partition a weighted network into sets of non-overlapping nodes (i.e., modules) such that within-module connections are maximized and between-module connections are minimized. Some networks exhibit more of a modular structure than others; \( \omega \) is a quantitative measure of modularity for each weighted network (Eq. 4; Rubinov & Sporns, 2010), which is defined as

\[
Q^w = \frac{1}{1w} \sum_{i,j \in N} \left[ w_{ij} - \frac{k_i w_i}{1w} \right] \delta_{m_i,m_j} \quad (4)
\]

where \( m_i \) is the module containing node \( i \), and \( \delta_{m_i,m_j} = 1 \) if \( m_i = m_j \) and 0 otherwise. The modularity calculation is stochastic; in our analysis we performed a modularity partition 10,000 times and averaged across these iterations to calculate a mean \( Q \) coefficient for each concept. Nodes may have connections to many different
modules, or have very few such connections. The diversity coefficient \( h_i \) is a measure ascribed to individual nodes that reflects the diversity of connections that each node has to modules in the network. This is a version of the participation coefficient, and is calculated using normalized Shannon entropy; we have previously used entropy to model property flexibility, and so predicted that diversity would be a good candidate for a network-based measure of conceptual flexibility. The diversity coefficient (Eq. 5; Rubinov & Sporns, 2011) for each node is defined as

\[
h_i^\pm = \frac{1}{\log m} \sum_{u \in M} p_i^\pm (u) \log p_i^\pm (u),
\]

where \( p_i^\pm (u) = \frac{s_i^\pm (u)}{s_i^\pm} \), \( s_i^\pm (u) \) is the strength of node \( i \) within module \( u \), and \( m \) is the number of modules in modularity partition \( M \). We averaged diversity coefficients across nodes in a network to obtain a mean measure of diversity for each concept network. The diversity coefficient is based on \( Q \), which is stochastic; we thus calculated a diversity coefficient for each of the 10,000 modularity partitions, and averaged across these iterations for each concept.

Core-periphery structure is another way to describe the structure of a network. Here, we attempt to partition a network into two non-overlapping sets of nodes such that connections within one set are maximized (i.e., the “core”) and connections in the other are minimized (i.e., the “periphery”). Core-periphery fit \( (Q_C) \) is a quantitative measure of how well each network can be partitioned in this way (Eq. 6), and for weighted networks is defined as

\[
Q_C = \frac{1}{v_c} \left( \sum_{i,j \in C_c} (w_{ij} - \gamma_c \bar{w}) - \sum_{i,j \in C_p} (w_{ij} - \gamma_c \bar{w}) \right)
\]

where \( C_c \) is the set of all nodes in the core, \( C_p \) is the set of nodes in the periphery, \( \bar{w} \) is the average edge weight, \( \gamma_c \) is a parameter controlling the size of the core, and \( v_c \) is a normalization constant (Rubinov et al., 2015).

2.2.1.4 Bootstrap Analysis

The properties and subordinates used to create these networks were chosen by participants, not experimenters. However, there was a certain degree of subjectivity in how the final lists were constructed, and the properties and subordinates reported by the participants are unlikely to fully represent the total possible sets. We thus ran bootstrap analyses in order to explore whether relationships between network and non-network measures were dependent on the particular sets of properties and subordinates used in our network construction.

**Bootstrapping over Subordinates:** The goal of this analysis was to generate a distribution of correlation values for a specific pair of measures. In each iteration
of the analysis, new networks were constructed: for each concept, a random subordinate was removed before network construction. Network measures were extracted from these 15 concept networks, and correlated with another measure of choice: this correlation value was recorded. We performed 1000 iterations of this analysis, resulting in a distribution of 1000 correlation values along with a 95% confidence interval.

Bootstrapping over Properties: The goal of this analysis was to generate a distribution of correlation values for a specific pair of measures. In each iteration of the analysis, new networks were constructed: for each concept, a random 10% of its meaningful properties were removed \( (N_C) \) before network construction. Network measures were extracted from these 15 concept networks, and correlated with another measure of choice: this correlation value was recorded. We performed 1000 iterations of this analysis, resulting in a distribution of 1000 correlation values along with a 95% confidence interval.

2.2.2 Figurative Language Task
The structure and flexibility of a concept likely has implications for how the concept can be used in creative contexts, such as in figurative language comprehension (e.g., Sloman et al., 1998). We therefore set out to collect data reflecting the extent to which a given concept is easily interpreted in a figurative context, and to explore the characteristics of conceptual structure that may facilitate this creative process.

2.2.2.1 Participants
300 subjects recruited from Amazon Mechanical Turk contributed data to this study and were compensated according to current standard rates. Consent was obtained for all participants in accordance with the University of Pennsylvania IRB.

2.2.2.2 Stimuli
Experimental similes of the form “X is like a Y” were constructed using the 15 target concepts in the “vehicle” (Y) position. Fifteen additional concepts were used in the “tenor” (X) position of the similes: TRUTH, TIME, CONVERSATION, SADNESS, CITY, LIFE, DREAM, CAREER, FAMILY, FRIENDSHIP, GOVERNMENT, SCHOOL, HAPPINESS, CELEBRATION, BOREDOM. These tenor concepts were chosen to minimize sensorimotor content, but otherwise were chosen randomly. The 15 target concepts and 15 tenor concepts were fully crossed, resulting in 225 novel similes (e.g., “Truth is like a key”, “Happiness is like chocolate”, “Boredom is like a bottle”). The experimental similes were split into 15 lists consisting of 15 similes each; each target concept and each tenor concept occurred once within each list. An additional 10 similes were taken from Blasko & Connine (1993) and were used as control similes; five were high-apt similes (e.g., “A book is like a treasure chest”) and five were moderate-apt similes (e.g., “Stars are like signposts”). Each
list (and therefore each of the experimental similes) was seen by 20 participants. The control similes were seen by all 300 participants.

2.2.2.3 Task
Each participant read 25 total similes (10 control similes, 15 experimental similes) presented in a randomized order. On each trial, the simile was presented at the top of the screen, with two sliding-scale questions beneath. To assess simile meaningfulness, participants were asked: “How meaningful is this figurative sentence?” ranging from “It does not have any meaning at all” to “It has a very strong meaning.” To assess simile familiarity, participants were asked: “How familiar is this figurative sentence?” ranging from “It is not familiar at all” to “It is very familiar.” These questions were motivated by prior work on simile comprehension (Blasko & Connine, 1993). For both questions, participants were asked to slide a bar in order to make their desired response. Values on both questions ranged from 0-100, but these values were not displayed to participants.

2.2.2.4 Analysis
Meaningfulness and familiarity ratings for each simile were separately averaged across participants (experimental: \(N=20\), control: \(N=300\)). Inspection of the control similes borrowed from Blasko & Connine (1993) revealed that our measures were sensitive to simile characteristics reported elsewhere: the five high-apt similes were judged more meaningful than the five moderate-apt similes \((t(8)=7.20, p<0.0001)\); they were also judged as more familiar \((t(8)=5.60, p<0.001)\).

The meaningfulness and familiarity ratings across all experimental and control similes are shown in Fig. 4. Here it is clear that the meaningfulness and familiarity ratings of our experimental similes are not categorically different from the similes reported in outside literature, suggesting that our constructed similes — though novel and pseudo-randomly generated — were meaningful enough for further analysis. The control similes were not included in any subsequent analysis.
Across the 225 experimental similes, meaningfulness and familiarity ratings were highly correlated ($r(223)=0.83$, $p<0.0001$); since this tight correspondence made it difficult to tease apart the separate measures, we averaged meaningfulness and familiarity measures within each simile to construct a composite measure we refer to as simile “goodness”. These simile goodness measures were then averaged with respect to each target concept; that is, the goodness ratings for the 15 similes that contained the same target concept (e.g., “chocolate”) were averaged together (e.g., “Truth is like chocolate”, “Happiness is like chocolate”). This resulted in a single simile goodness measure for each of our 15 target concepts.

2.2.2 Alternative Uses Task

The Alternative Uses Task (AUT) is a widely used measure of creative cognition in which participants generate novel uses for common objects. In order to further explore the relationships between conceptual structure and flexible concept use, we set out to collect data reflecting the extent to which a given concept can be re-imagined in creative ways.

2.2.3.1 Participants

28 participants recruited from Amazon Mechanical Turk generated novel uses in the AUT, and an additional 25 participants provided ratings on these responses. All participants were compensated according to current standard rates. Consent
was obtained for all participants in accordance with the University of Pennsylvania IRB.

2.2.3.2 Alternative Uses Task

Participants (N=28) generated alternative uses for the 15 concepts. They were instructed to think of as many novel uses of each object as they could, that responses should be plausible but significantly different from the common use of the object, and that there were no right or wrong answers. On each trial, the concept label appeared above blank response boxes. Participants had 60 seconds to answer with as many alternative uses as they could. After 60 seconds passed, the next trial immediately began. The presentation order of the 15 concepts was randomized.

After these data were collected, we removed responses that were not task-relevant (e.g., “I can’t think of anything”), and terminal responses that were incomplete (due to the strict 60-second time limit). The number of responses given by each participant for each concept was recorded, and these values were averaged across participants: this resulted in a measure that reflected the mean number of alternative uses generated for each of the 15 concepts (M=3.24, SD=0.39).

From the full set of responses, we selected the first response from each participant for each of the 15 concepts: this resulted in a set of 420 alternative uses (28 per concept). These responses were edited such that they began with a verb (e.g., “Use as a hat”, “Use as a bowling ball”, “Make pie”). This set of responses was rated by an additional set of participants in the next stage of this study.

Figure 5: Classification results. We ran a range of classification analyses using different numbers of eigen-dimensions from our concept networks. Classification was successful using 1 dimensions in both Set 1 and Set 2. Classification performance increased as more dimensions were added, such that performance of the network-models reached performance of the vector-based models (single data points).
2.2.3.3 Alternative Uses Ratings

An additional set of participants \((N=25)\) provided ratings on the alternative uses data described above. Participants were told that they would be judging other participants’ responses on an alternative uses task, which was used to study creative thinking. On each trial, they were presented with a concept label (e.g., “PUMPKIN”) and one alternative-use response beneath (e.g., “Use as a hat”). Participants were asked to make three multiple-choice responses on each trial. The first two questions were: “Would this work?” (yes/maybe/no) and “Have you seen someone use this object to do this before?” (yes/no). These questions were not central to our aims of this study, and the corresponding data will not be reported here. The third question, which provided our main measure of creativity, asked participants to rate each response as one of the following: (1) Very obvious/ordinary use, (2) Somewhat obvious use, (3) Non-obvious use, (4) Somewhat imaginative use, and (5) Very imaginative/re-contextualized use. The design of this question was motivated by Hass et al. (2018), in which good reliability was obtained for ratings on a similar alternative uses task. Each participant rated 80 total responses, which were approximately evenly distributed across the 15 concepts; trials were presented in a randomized order. Each alternative-use response was rated by 5 different participants.

For the main creativity measure, ratings (1-5) for each alternative-use response were averaged across the 5 participants. The mean ratings of creativity for each alternative-use were then averaged within each concept, resulting in a measure that reflects the mean creativity score of alternative-use responses for each of the 15 concepts \((M=2.91, SD=0.45)\).

2.3 Results

2.3.1 Classification Results

In order to determine whether our concept networks contained concept-specific information, we ran a classification analysis using eigendecomposition for both Set 1 and Set 2. We ran multiple analyses using different ranges of eigenvectors, which were sorted by eigenvalue (positive to negative). We started by only using the first eigenvector in each of the concept networks and determined whether this dimension alone could be used to classify the property vector as one of the 5 concepts in Set 1 or 10 concepts in Set 2. One dimension was enough to classify exemplars in both Set 1 (Mean Accuracy=0.31; SE=0.03; Chance=0.20) and Set 2 (Mean Accuracy=0.53; SE=0.03; Chance=0.10). Classification accuracy in both Set 1 and Set 2 continued to increase as more eigenvectors were included in the analysis (Fig. 5), with performance leveling off around 22-25 eigenvectors. The network-based classification accuracy reached the performance of a more traditional vector-based classifier (rightmost point on graph), which was successful at classifying exemplars in Set 1 (Mean
Accuracy=0.85; SD=0.06; Chance=.20) and Set 2 (Mean Accuracy=0.84; SD=0.10; Chance=0.10). The successful classification of conceptual exemplars using our concept network models suggests that the structures of these networks are concept-specific. We can now extract and analyze traditional network science measures from these concept-specific networks in order to examine the relationships between network topology and flexible concept use.

2.3.2 Network Measures of Conceptual Structure

We extracted network measures from 15 concept networks and explored how they relate to text-based and empirical measures of conceptual flexibility. Hoffman et al. (2013) use word co-occurrence statistics to quantify the context-dependent variations in word meanings found in language. In order to capture “semantic ambiguity” and “flexibility of word usage” in a computational framework, the authors provide a measure of semantic diversity (SemD) based on latent semantic analysis (LSA; Hoffman et al., 2013). A high SemD item is a word that occurs in diverse language-based contexts — that is, the verbal context surrounding instances of the word are relatively dissimilar in meaning. Based on the assumption that flexibility of word usage reflects flexibility of meaning, we extracted SemD values for our 15 concepts to determine whether SemD predicts any of our network measures of interest. The correlations between all network measures, along with means and standard deviations, are shown in Table 1.

<table>
<thead>
<tr>
<th></th>
<th>Modularity</th>
<th>Diversity</th>
<th>Core-Periphery</th>
<th>Clustering</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (SD)</td>
<td>0.01 (0.002)</td>
<td>0.96 (0.007)</td>
<td>0.17 (0.025)</td>
<td>0.48 (0.042)</td>
</tr>
<tr>
<td>Modularity</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Diversity</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Core-Periphery</td>
<td>-0.089</td>
<td>0.446†</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Clustering</td>
<td>-0.461†</td>
<td>-0.186</td>
<td>-0.561*</td>
<td>—</td>
</tr>
</tbody>
</table>

* p < 0.10   † p < 0.05

Table 1: Statistics and correlations between network measures. The mean and standard deviation across 15 concepts is shown for each network measure of interest in the top row. Beneath are the spearman correlation values between the four network measures. The only significant relationship is between core-periphery structure and network clustering.
We hypothesized that network modularity, network diversity, core-periphery structure, and network clustering might relate to conceptual flexibility, and we tested whether our network measures correlated with SemD (Hoffman et al., 2013) across our 15 concepts. First we examined modularity and diversity — these measures capture the extent to which a network can be partitioned into distinct clusters of nodes (“modules”), and the extent to which individual nodes participate in these modules, respectively; SemD was not predicted by network modularity ($p>0.2$) nor network diversity ($p>0.5$).

Next, we examined core-periphery structure — this measure reflects the extent to which a network can be partitioned into one densely connected core and a sparsely connected periphery; SemD was positively predicted by core-periphery fit ($r=0.71$, $p=0.003$; Fig 6A). Core-periphery fit was not significantly related to $K$ ($p=0.16$), $Nc$ ($p>0.3$), or simple stability ($p>0.3$). Further, core-periphery fit predicted SemD when separately controlling for each of these measures in a general linear model (all $p$’s <0.035). Our interpretation of the positive relationship between SemD and core-periphery fit is that the presence of a “periphery” — that is, a set of weakly associated features — relates to increased variation of potential word meaning.

The last network measure we explored was the clustering coefficient, which reflects the overall connectivity of a network; SemD was negatively predicted by clustering ($r=-0.70$, $p=0.004$; Fig 6B). Clustering was not significantly related to $K$ ($p=0.18$), $Nc$ ($p>0.4$), or simple stability ($p>0.7$). Further, clustering predicted SemD when separately controlling for each of these measures in a general linear model (all $p$’s <0.04), except for $K$ ($p=0.12$). We interpret this negative
The positive relationship between core-periphery structure and SemD is robust to variations in subordinates and properties; the negative relationship between network clustering and SemD is robust to variations in subordinates and properties.

Core-periphery fit and network clustering were significantly negatively related to each other ($r=-0.56, p=0.03$). No other measures of interest were significantly correlated, though there was a trend towards a negative relationship between modularity and clustering ($r=-0.46, p=0.08$), and diversity and core-periphery fit were marginally positively correlated ($r=0.45, p=0.095$).

Finally, we explored whether SemD related to our simple stability measure which reflects the extent to which properties in a concept are represented across all of that concept’s subordinates. We observed a significant negative relationship between SemD and simple stability ($r=-0.52, p=0.046$). Simple stability did not significantly relate to any of the network measures of interest ($p$'s > 0.3).

### 2.3.3 Bootstrap Analysis

We ran bootstrap analyses to test the robustness of network-SemD relationships when only a subset of subordinates or a subset of properties was used to create the concept networks. These tests were to confirm that the network measures we report in this paper are not dependent on the exact subordinates and the exact properties that went into network construction.
Results of the leave-one-out subordinate analysis are shown in Fig 7A. For each of the four network measures of interest, the bootstrap analysis resulted in a distribution of correlations with SemD (blue histograms), along with a mean and 95% confidence intervals (pink bars). The distribution of core-periphery SemD-correlations was significantly greater than zero ($p=0$), and the distribution of connectivity SemD-correlations was significantly less than zero ($p=0$); these results confirm that the correlations reported above are robust to variations in the set of subordinates for each concept. Though modularity was not significantly related to SemD, the distribution of SemD-correlations was significantly above zero ($p=0.01$). On the other hand, the distribution of diversity SemD-correlations was not significantly different from zero ($p=0.23$).

Results of the leave-10%-out property analysis are shown in Fig 7B. For each of the four network measures of interest, the bootstrap analysis resulted in a distribution of correlations with SemD (blue histograms), along with a mean and 95% confidence intervals (pink bars). The distribution of core-periphery SemD-correlations was significantly greater than zero ($p=0$), and the distribution of connectivity SemD-correlations was significantly less than zero ($p=0$); these results confirm that the correlations reported above are robust to variations in the set of properties for each concept. The distributions of SemD-correlations were not different from zero for either modularity ($p=0.12$) or diversity ($p=0.32$).

In both of these bootstrap analyses, the variability in correlations between network diversity and SemD is striking. This instability of network diversity across
bootsrap might indicate that network diversity is driven by a small number of nodes with highly-varied connections whose presence varies over each iteration. However, the particular local and global structures of a concept network that contribute to network diversity — and the stability of this measure across different concepts and methods of network construction — is an open question for future work.

2.3.4 Similes
Interpreting a simile (e.g., "X is like a Y") involves context-dependent activation of conceptual meaning: the explicit comparisons contained in a simile implies that X is similar to Y with respect to a certain dimension(s) or subset of properties of Y, thus requiring the interpreter to select a likely subset of Y’s properties that is being asserted for X. For example, to interpret the simile “Truth is like a knife”, one must decide which properties of KNIFE can also apply to TRUTH. Since this process involves within-concept property structures and flexible conceptual meaning, we asked which of our measures, if any, predict the simile-goodness measure we constructed for each of our 15 target concepts.

SemD did not predict simile goodness (r=−0.14, p>0.6). The two network measures that were significantly related to SemD — core-periphery and clustering — did not predict simile goodness either (Core-periphery: r=0.04, p>0.8; Clustering: r=0.25, p>0.3); neither did network modularity (r=−0.11, p>0.7). Interestingly, network diversity positively predicted simile goodness across the 15 target concepts (r=0.54, p=0.04; Fig 8A). Network diversity still predicted simile goodness when controlling for K (p=0.04) and Nc (p=0.03) separately as well as simultaneously (p=0.04). However, the bootstrapped distributions of network-diversity and simile goodness correlations were not significantly greater than zero (Supplemental Figs. 1 and 2), suggesting that this relationship might not be robust to variations in subordinates and properties used during network construction. The relationships between concept network structure and figurative language comprehension should be further explored in future work.

2.3.5 Alternative Uses
Generating alternative uses for common objects involves thinking about those objects in new, creative ways (Chryssikou & Thompson-Schill, 2011). Theories posit that the generation of novel uses requires one to suppress the typical function of the object and to pay attention to its sensorimotor properties. For example, the common use of CHOCOLATE is “to eat” — in order to realize that it can also be used “as paint” one must activate the CHOCOLATE properties of MELTABLE and BROWN. Since this process involves consideration of a concept’s properties and property relationships, we asked which of our measures, if any, predict performance on an alternative uses task.
First, we analyzed the average number of responses given for each of the 15 concepts in the alternative uses task. SemD positively predicted the number of responses ($r=0.67$, $p=0.006$), such that if a concept-label (“key”) occurs in a diverse range of textual contexts, participants will generate more potential alternative uses for that concept (e.g., will give more responses for what novel things to do with a key). None of the network measures predicted number of alternative use responses (all $p$'s $>0.1$).

Second, we analyzed the average creativity ratings for the first response given for each of the 15 concepts. Mean creativity was strongly negatively predicted by mean number of responses ($r=-0.77$, $p<0.001$), suggesting that concepts that inspire more responses also tend to inspire less creative (initial) responses. SemD did not predict mean creativity ($r=-0.41$, $p=0.13$). The two network measures that were significantly related to SemD — core-periphery and clustering — did not predict creativity either (Core-periphery: $r=-0.32$, $p>0.2$; Clustering: $r=0.40$, $p=0.14$). A negative relationship between network modularity and creativity was observed, but did not reach statistical significance ($r=-0.50$, $p=0.056$). Interestingly, network diversity was a negative predictor of creativity ($r=-0.56$, $p=0.03$; Fig. 8B). Network diversity still predicted creativity when controlling for $K$ ($p=0.041$) and $Nc$ ($p=0.036$) separately, and was marginally reliable when both were controlled for simultaneously ($p=0.053$). As in the previous analysis, the bootstrapped distributions of network-diversity and creativity correlations were not significantly less than zero (Supplemental Figs. 3 and 4), suggesting that this relationship might not be robust to variations in subordinates and properties used during network construction. However, the other network measures were reliably related to creativity in the bootstrap analyses, suggesting that concept network structure may relate to people’s ability to use the concept in creative ways.

2.4 Discussion

Here our goal was to model basic-level concepts using graph-theoretical networks. A model structured using within-concept feature statistics provides a framework in which varied yet appropriate instantiations of a concept may be flexibly activated. An apple network may contain a strong connection between crunchy + fresh and between soft + baked, enabling the conceptual system to know what sets of properties should be activated in a particular apple instance — for example, in the representations evoked by “apple picking” versus “apple pie.” The property-covariation statistics for a given concept will determine which sets of properties tend to co-occur, and how individual properties relate to those sets and to each other. Here we have demonstrated (1) how to create these concept network models, (2) that these models are concept-specific, and (3)
The structural characteristics of these networks can predict other measures of conceptual processing.

The concept network approach we describe here is a general one, and there are many different ways in which feature-based concept networks can be constructed. We have walked through one potential way to do so, in a proof-of-concept that reveals the feasibility and potential utility of this approach. The specific methodological decisions used in this worked through example are described above, and we hope that other researchers interested in modeling and capturing conceptual flexibility will use variations of these methods — for example, different concepts, distance metrics, or network measures — in order to further explore how conceptual structure relates to flexible concept use.

There would be no point in attempting to extract concept-specific measures of conceptual flexibility if the networks themselves did not contain concept-specific feature relationships. Our approach assumes that this is the case, though this was not a theoretical certainty. In fact, many feature-based models of conceptual knowledge rely on feature correlations across the entirety of semantic space or within a large semantic domain, and represent many concepts within this single correlational feature space. For example, it could have been the case that the property BLACK relates to SOFT and ROTTEN across all concepts. However, our analyses suggest that properties relate to each other in different ways across basic-level concepts. For example, BLACK might relate to SOFT and ROTTEN in BANANA, but with FIRM and BITTER in CHOCOLATE. We found that our 15 concept networks could successfully discriminate between new conceptual exemplars, suggesting that within-concept feature statistics differ reliably between basic-level concepts. These results emerged out of a classification analysis based on eigendecomposition of our concept networks. Eigendecomposition of graphs has previously been used to assess the correspondences between anatomical brain network structure and patterns of functional activation (Medaglia et al., 2017); here we adapted this method to assess the correspondences between conceptual structure and feature-vectors for individual conceptual exemplars. Empirically demonstrating that networks contain concept-specific feature statistics enabled us to analyze each concept’s network structure and relate structural characteristics to aspects of conceptual processing.

The kinds of structure we analyzed here included network clustering, modularity, core-periphery, and diversity. In order to explore whether these network structures could predict interesting aspects of conceptual processing, we examined three external measures: a text-based measure of semantic diversity (SemD; Hoffman et al., 2013), empirical measures of simile goodness, and empirical measures of creativity on an alternative uses task. We found reliable relationships between network measures and each of these three external data
sets, highlighting the potential of this approach to capture aspects of flexible concept use.

Network clustering, quantified in a *clustering coefficient*, captures the extent to which nodes are linked to its nearest neighbors. A network characterized by high clustering is one in which network nodes form “cliques” in which nearby nodes are linked to each other (Watts & Strogatz, 1998). This is intuitive in social networks, in which friends of one person tend to be friends with each other. High network clustering has been observed in text-based semantic networks (Steyvers & Tenenbaum, 2005), and semantic networks exhibit greater clustering in high-versus low-creative individuals (Kenett et al., 2014). Here, we observed that feature-based concept networks with greater clustering exhibit less text-based semantic diversity (SemD). That is, words that do not occur in many text-based contexts correspond with concepts whose features exhibit strong clustering. This result was robust to variations in properties and subordinates used in network construction. This finding suggests that dense local feature associations within a concept network reduce the extent to which word meaning can vary across instances.

Network modularity, quantified in coefficient Q, captures the extent to which a network can be partitioned into densely connected modules (i.e., sets of nodes) with sparse connections between them. Modularity is a defining characteristic of “small-world” networks (Bassett & Bullmore, 2006) and has been observed in semantic networks (Kenett et al., 2015) as well as functional brain networks (Bassett et al., 2011). Here, we did not observe direct relationships between network modularity and our other measures, though our additional bootstrap analyses suggest that stronger relationships between modularity and conceptual processing may be found when a larger set of concept networks are analyzed.

Core-periphery structure, quantified in a *core-fit* measure, reflects the extent to which a network can be portioned into one set of densely connected nodes (core), with sparse connections between all other nodes (periphery). This kind of network structure, originally observed in social networks (Borgatti & Everett, 2000t), has also been observed in functional brain networks (Bassett et al., 2013). Here, we found that concept networks with stronger core-periphery structures exhibit greater text-based semantic diversity (SemD); this result was robust to variations in properties and subordinates used in network construction. This finding suggests that the presence of one set of highly-associated features (“core”) in addition to a substantial set of weakly-associated features (“periphery”) is predictive of conceptual flexibility. In particular, this structure might enable substantial variation in the activation of individual periphery features across instances of concept representation.
Network diversity, quantified in a diversity coefficient, reflects the extent to which nodes in a network participate in few or many network modules. This a version of a “participation” coefficient calculated using Shannon entropy (Rubinov & Sporns, 2011). In functional brain networks, these measures are typically used to define network “hubs” (Sporns, 2014), which are particularly important for transitioning between network states (i.e. patterns of activity in a network). Here, we observed that network diversity positively predicted simile goodness judgments and negatively predicted creativity of responses in an alternative uses task. However, these results were not robust to variations in properties and subordinates used in network construction, so these specific relationships should be interpreted with caution until replicated in a larger set of concept networks.

We observed relationships among non-network measures that are interesting in their own right. First, we found that the distributional, corpus-based measure of SemD (Hoffman et al., 2013) was negatively correlated with our simple stability measure, which is the proportion of properties that are present in all of a concept’s subordinates. As discussed in Landauer & Dumais (1997), two important aspects of word meaning are usage and reference; measures of distributional semantics (e.g., LSA, SemD) are constructed based on usage only, and do not contain nor point to information in the world to which a word refers. Feature-based measures (e.g., McRae et al., 1997; Tyler & Moss, 2001), on the other hand, do incorporate reference into word meaning by pointing to the sets of features contained in each concept. Though distributional semantic approaches have their benefits, it is often difficult to know what their respective measures relate to from a cognitive or psychological standpoint. Our finding that the distributional, corpus-based statistic of semantic diversity was negatively related to a feature-based statistic of conceptual stability provides some insight into how usage- and reference-based measures of conceptual diversity and stability might converge.

The alternative uses task (AUT) also resulted in findings that warrant further investigation. First, we found that the SemD of a word is a positive predictor for the number of alternative use responses participants can generate for that item. One interpretation of this finding is that if a word is found in more diverse text-based contexts it is easier to think of alternative uses of the object to which the word refers. We additionally found that the mean number of alternative use responses for a concept is negatively correlated with the creativity of the initial alternative use response. Though further analysis of these findings is beyond the scope of the current report, this might be a relevant finding for those interested in the AUT task and creativity more generally.

Taken together, these results reveal the ability of feature-based concept networks to capture meaningful aspects of conceptual structure and use. The
analyses reported here were exploratory; we did not have any strong a priori predictions of which network measures would relate to each additional conceptual measure, and in what direction. However, we did predict that network measures— which capture different kinds of conceptual structure— would predict the ways in which concepts are flexibly used in language and thought. We have thus demonstrated how to create concept-specific networks, and that the structures of these networks can be related to other concept-specific measures. The external measures we report here (SemD, simile goodness, alternative uses task) are intended to serve as examples of measures that could be related to concept network structure. We look forward to future work further exploring the utility of this concept network framework in the study of conceptual knowledge.

Linking back to cognitive theories of conceptual knowledge, this concept network approach has similarities to theories that aim to characterize the flexibility of individual features. Sloman et al. (1998) captured pairwise relations between features in order to model the feature-based structure of individual concepts. These authors were interested in the role of individual features with respect to conceptual coherence, which relates to notions of centrality in the “intuitive theory” view of concepts (e.g., Keil, 1989; Murphy & Medlin, 1995). Sloman et al. (1998) simplified this previous notion of centrality by basing conceptual structure on asymmetrical dependency relationships between features; this structure captures the concept-specific “mutability” of a feature, and offers a framework in which concepts can be structured yet flexible. In the current paper, our concept networks are defined by symmetrical feature co-occurrence statistics rather than asymmetrical dependency relationships. However, it would be possible to capture feature dependencies in directed concept networks (i.e., with asymmetrical links). The use of network science tools enables us to analyze not only a concept’s global structure, but also the characteristics of individual feature nodes (e.g., mutability, centrality). These kinds of structures have implications for flexible concept use such as analogies, metaphors, conceptual combination (Sloman et al., 1998).
Though we believe that a feature-based concept network approach will provide a new set of useful tools with which to study conceptual flexibility, it is not the only way to do so. Other frameworks have the potential to capture the flexibility of the conceptual system, including attractor networks (e.g., Cree et al., 1999; 2006; Rodd et al., 2004) and recent updates of the hub-and-spoke model (Ralph et al., 2017; Hoffman et al., 2018). The concept network framework proposed here is not in opposition with these other approaches; the development and implementation of all of these methods will greatly benefit our understanding of the semantic system. However, we do believe that a network science approach to conceptual knowledge has its unique advantages. The ability of graph theoretical network science to model a vast range of systems enables us to examine conceptual structure across cognitive, linguistic, and neural levels of analysis. The structure of behavioral, feature-based networks (as discussed here) can be analyzed and compared with the structure of functional brain
networks within specific cortical sites (e.g., ATL) or across the brain as a whole. There is an additional possibility of analyzing “informational” brain networks, in which networks are constructed on the basis of simultaneous pattern-discriminability across cortical sites (Informational Connectivity; Coutanche & Thompson-Schill, 2013). Network neuroscientists have previously forged links to cognitive processes such as motor-sequence learning (Bassett et al., 2011) and cognitive control (Medaglia et al., 2018), setting a precedent for the application of networks to cognitive neuroscience.

Recent work exploring the intersection of network science with control theory suggests another possible future direction. Network controllability refers to the ability to move a network into different network states, and has been applied to structural brain networks in order to shed insights into how the brain may guide itself into easy- and difficult-to-reach functional states (Gu et al., 2015). There have been additional attempts to link brain network controllability to cognitive control (Medaglia, 2018). The application of control theory to concept networks may provide an additional way to quantify conceptual flexibility by identifying nodes that are well-positioned to drive the brain into diverse, specific, or integrated states. Perhaps concept networks that are more controllable overall — that is, networks in which it is easier to reach varied network states — correspond to concepts that are more cognitively flexible.

So far, we have discussed mainly event context — a BANANA representation will be slightly different while one is painting as opposed to eating, and it will be different before and after a peeling-event has occurred. However, language itself can provide a context: language is inherently interactive, and the meaning of a word (i.e., the corresponding conceptual content) will depend on the words surrounding it (e.g., Pustejovsky, 1998; McElree et al., 2001; Traxler et al., 2005). Researchers interested in conceptual combination aim to understand how the meaning of a combined concept (e.g., “butterfly ballerina”) can be predicted based on the meaning of its individual constituents (Coutanche et al., in press). This is not a unique challenge for noun-noun compounds, but also for adjective-noun compounds: even the (putatively) simple concept RED has different effects when combined with the concepts TRUCK, HAIR, and CHEEKS (see Halff, 1976). Combining a noun-concept with an adjective-concept might not simply involve the reweighting of a single property node, but a more complex interaction governed by within-concept statistics.

These predictions could be generated using models of signal propagation, such as spreading activation (e.g., Collins & Loftus, 1975; De Deyne et al., 2016) or information diffusion (e.g., Bakshy et al., 2012). On graph-theoretical networks, spreading activation can be formalized by simulating “random walks” over the network (Abbott et al., 2015; De Deyne et al., 2016). The edges in the network
can be used to define a transition-probability matrix (e.g. Fig. 9A), which contains the probabilities of transitioning from one node to another node. A random walk begins at an initial node, and then continues to a new node by following one of that node’s edges at random. Eventually the random walk will end (e.g., determined by a maximum path length), and the resulting path and activated nodes can be observed. This approach, which has been used to analyze word association data over large-scale semantic networks (Abbott et al., 2015; De Deyne et al., 2016), can also be used to analyze the concept networks proposed here. Random walks over concept networks could be used to predict the properties activated for a concept in a given context — starting with an initial property node, we could trace the activation of associated properties to predict the cluster of properties that are likely to be activated in that instance. For example, a random walk over the BANANA network starting at node FIRM would likely include GREEN, SWEET, and YELLOW, whereas a random walk starting at SOFT would be more likely to include BROWN, SWEET, and YELLOW (Fig. 9). These methods could thus be used to predict interpretations of conceptual combinations.

Finally, interesting differences might exist in the flexibility of concept networks across individuals, or in the neural networks that support their processing. It has previously been suggested that individual differences in functional neural networks relate to differences in psychopathology (Lynall et al., 2010; van den Heuvel, 2013), and that the structure of semantic networks differ between low- and high-creative persons (Kenett et al., 2014). Though these networks are different in kind than the ones proposed here, the same kinds of comparisons could be explored. The ability to shift a concept network from one state to another, that is, the ability to flexibly modify the activation of certain properties, could relate to an individual’s ability to generate or comprehend novel metaphors, or to generate novel uses for common objects. The flexibility of person-specific concept networks and neural networks could be explored in relation to performance on these kinds of tasks. More generally, the structure and flexibility of individuals’ concept networks may differ in meaningful ways, and this could be a fruitful avenue for future research.

We acknowledge that our proposed concept network model framework has some limitations. A large amount of data had to be collected such that we could calculate within-concept statistics and run a classification analysis. Here we were only able to construct and analyze 15 concept networks, a small sample with which to work. However, the primary bottlenecks were related to the classification analysis, which required that the concepts within a set were initially defined by the same set of properties, and required other tasks related to stimulus design and data analysis. Here we have provided evidence that concept networks are concept-specific, and thus classification analyses are not an essential part of the
pipeline moving forward. This will greatly reduce the effort and time needed to construct concept networks in future work.

Additionally, there is a certain degree of experimenter subjectivity in the final selection of properties and subordinate concepts. Both our properties and subordinates were reported by subjects, and were not determined by the experimenters. However, some decisions had to be made, such as deciding whether a response should be considered a subordinate or a property (e.g., “nuts” with respect to CHOCOLATE), and deciding an appropriate level of specificity of subordinates (e.g., excluding some brand names). In order to mitigate these concerns, we have reported bootstrap analyses in which subsets of subordinates or properties are excluded in order to assess the robustness of network measures and their relationships. These kinds of analyses can be used in the future in order to flag any potentially idiosyncratic effects in a given data set. Despite these limitations, we believe that the construction and analysis of concept networks will provide useful insights on the relationship between concept structure and flexible concept use.

In this proof-of-concept, we have constructed concept network models, confirmed their ability to capture concept-specific information, and extracted network measures that relate to external test-based and behavioral measures. We believe the application of network science to conceptual knowledge will provide a set of tools that will enable the intrinsic flexibility of the conceptual system to be explored and quantified. We hope that other researchers will be able to use these tools to further our understanding of conceptual flexibility and the conceptual system more broadly.
2.5 Supplemental Material

Figure S1: Subordinate bootstrap analysis of network measures and simile goodness. Distributions of correlations between network measures and simile goodness when multiple networks are constructed using subsets of subordinates. Though we observed a significant positive correlation between network diversity and simile goodness, this result is not robust to variations in subordinates. Further work is needed to determine which aspects of network structure relate to comprehension of figurative concept use.
**Fig S2: Property bootstrap analysis of network measures and simile goodness.**

Distributions of correlations between network measures and simile goodness when multiple networks are constructed using subsets of properties. Though we observed a significant positive correlation between network diversity and simile goodness, this result is not robust to variations in properties. Further work is needed to determine which aspects of network structure relate to comprehension of figurative concept use.
Figure S3: Subordinate bootstrap analysis of network measures and AUT creativity. Distributions of correlations between network measures and creativity on the AUT when multiple networks are constructed using subsets of subordinates. Though we observed a significant negative correlation between network diversity and creativity, this result is not robust to variations in subordinates. Further work is needed to determine which aspects of network structure relate to creative concept use.
Figure S4: Property bootstrap analysis of network measures and AUT creativity. Distributions of correlations between network measures and creativity on the AUT when multiple networks are constructed using subsets of properties. Though we observed a significant negative correlation between network diversity and creativity, this result is not robust to variations in properties. Further work is needed to determine which aspects of network structure relate to creative concept use.
<table>
<thead>
<tr>
<th>Concept</th>
<th>Subordinates</th>
<th>Excluded Subordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>KEY</td>
<td>car key, key to a city, door key, encryption key, garage key, gate key, key to my heart, house key, key card, keyboard key, map key, motorcycle key, office key, padlock key, password, piano key, safe key, skeleton key</td>
<td>electronic key, Florida Keys</td>
</tr>
<tr>
<td>PUMPKIN</td>
<td>pumpkin bars, pumpkin bread, pumpkin candy, canned pumpkin, pumpkin cookie, pumpkin field, Halloween pumpkin, Jack O’Lantern, pumpkin latte, pumpkin muffin, pumpkin pie, pumpkin puree, rotten pumpkin, pumpkin seeds, smashed pumpkin, pumpkin soup, pumpkin spice, whole pumpkin, Thanksgiving pumpkin</td>
<td>carved pumpkin, fresh pumpkin, gourd</td>
</tr>
<tr>
<td>GRASS</td>
<td>Astroturf, bamboo, barley, bent grass, grass clippings, crab grass, dead grass, hay, lawn grass, lemon grass, marijuana, oat grass, overgrown grass, grass seeds, sod, wheat grass</td>
<td>artificial grass, Bermuda grass, cut grass, blue grass, fescue, fresh grass, mowed grass, rye grass, St. Augustine grass</td>
</tr>
<tr>
<td>COOKIE</td>
<td>almond cookie, butter cookie, chocolate cookie, chocolate chip cookie, Christmas cookie, cookie cake, cookie dough, gingersnap, Girl Scout cookie, lemon cookie, M&amp;M cookie, macadamia nut cookie, macaroon, mint cookie, no-bake cookie, oatmeal raisin cookie, peanut butter cookie, shortbread cookie, snickerdoodle, sugar cookie, wafer</td>
<td>Chips Ahoy, fresh-baked cookie, frosted cookie, oatmeal cookie, white chocolate chip cookie</td>
</tr>
<tr>
<td>PICKLE</td>
<td>bread &amp; butter pickles, canned pickles, pickle chips, chopped pickles, cucumber pickles, dill pickles, garlic pickles, gherkins, hamburger pickles, homemade pickles, jarred pickles, kosher pickles, relish, sliced pickles, sandwich pickles, pickle spears, whole pickles</td>
<td>spicy pickles, sweet pickles</td>
</tr>
<tr>
<td>PILLOW</td>
<td>airplane pillow, bed pillow, cotton pillow, couch pillow, decorative pillow, down pillow, feather pillow, foam pillow, hypo-allergenic pillow, memory foam pillow, neck pillow, silk pillow, throw pillow, travel pillow</td>
<td>contour pillow, king pillow, standard pillow, Tempurpedic pillow</td>
</tr>
<tr>
<td>KNIFE</td>
<td>bread knife, butcher knife, butter knife, cheese knife, chef knife, dagger, hunting knife, jack knife, machete, paring knife, pocket knife, steak knife, switch blade, sword, throwing knife, utility knife</td>
<td>boning knife, Bowie knife, carving knife, cleaver, filet knife, kitchen knife, pen knife, Santoku knife, survival knife, Swiss Army knife</td>
</tr>
<tr>
<td>WOOD</td>
<td>wood blocks, wood chips, chopped wood, wood fence, firewood, wood floor, wood furniture, log, lumber, wood paneling, paper, wood planks, plywood, wood pulp, sticks, tree, cedar wood, cherry wood, maple wood, oak wood, pine wood, walnut wood</td>
<td>wood boards, balsa wood, birch wood, ebony wood, fir wood, hickory wood, mahogany wood, redwood, rosewood, teak</td>
</tr>
<tr>
<td>PHONE</td>
<td>Android phone, antique phone, broken phone, car phone, cell phone, emergency phone, flip phone, home phone, iPhone, landline phone, payphone, rotary phone, satellite phone, smart phone, wall phone, wireless phone</td>
<td>cored phone, cordless phone, house phone, mobile phone, Motorola phone, push-button phone, Samsung phone, telephone</td>
</tr>
<tr>
<td>CAR</td>
<td>broken down car, compact car, convertible, coupe car, electric car, family car, hatchback car, hybrid car, Jeep, luxury car, pickup truck, race car, rental car, sedan, sports car, station wagon, SUV, toy car, truck, used car, van</td>
<td>crossover car, Audi, BMW, Chevrolet, Dodge, Ferrari, Ford, GMC, Honda, Mazda, Mercedes, Nissan, Toyota</td>
</tr>
</tbody>
</table>
Table S1: Included and excluded subordinates for all concepts. Subordinates were excluded for a given concept if (a) the subordinate was similar or identical to a property included for that concept (e.g., “black paper”), (b) the subordinate was highly similar or identical to another subordinate included for that concept (e.g., “white chocolate chip cookie” compared to “chocolate chip cookie”), (c) the subordinate was a specific brand name that was similar or identical to other subordinates for that concept (e.g., “Chiquita banana”, “Audi”).

<table>
<thead>
<tr>
<th>Concept</th>
<th>Included Subordinates</th>
<th>Excluded Subordinates</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHOCOLATE</td>
<td>bittersweet chocolate, caramel chocolate, chocolate bar, chocolate chips, chocolate syrup, cocoa powder, dark chocolate, chocolate fudge, melted chocolate, milk chocolate, nut chocolate, salted chocolate, white chocolate, hot chocolate (drink)</td>
<td>baking chocolate, semi-sweet chocolate, unsweetened chocolate, smooth chocolate</td>
</tr>
<tr>
<td>BANANA</td>
<td>banana chips, banana pudding, Cavendish banana, fried banana, frozen banana, mashed banana, over-ripe banana, peeled banana, plantain, raw banana, red banana, ripe banana, rotten banana, sliced banana, unripe banana</td>
<td>apple banana, baby banana, Chiquita banana, green banana, old banana, organic banana, big banana, large banana, short banana, small banana, yellow banana</td>
</tr>
<tr>
<td>BOTTLE</td>
<td>baby bottle, beer bottle, broken bottle, juice bottle, liquor bottle, medicine bottle, milk bottle, soda bottle, spray bottle, water bottle, wine bottle</td>
<td>glass bottle, metal bottle, plastic bottle, clear bottle</td>
</tr>
<tr>
<td>TABLE</td>
<td>bedside table, changing table, coffee table, conference table, dining table, drafting table, end table, folding table, kitchen table, play table, poker table, pool table, side table, workbench</td>
<td>rectangular table, oval table, square table, round table</td>
</tr>
<tr>
<td>PAPER</td>
<td>butcher paper, cardboard, cardstock, construction paper, envelope, graph paper, legal paper, newspaper, notebook paper, paper towel, papyrus, poster board, printer paper, sandpaper, scrap paper, sketch paper, stationary, tissue paper, toilet paper, wrapping paper</td>
<td>writing paper, black paper, colored paper, white paper</td>
</tr>
</tbody>
</table>
3: FEATURE UNCERTAINTY PREDICTS THE FLEXIBLE NEURAL ACTIVATION OF CONCEPTUAL FEATURES IN COMBINED CONCEPTS

3.1 Introduction

Individual concepts (e.g., COOKIE, OCEAN) are rich structures comprising many features, and the activation of these features fluctuates across contexts. A COOKIE can be crispy, chewy, fresh, sweet, and sometimes even salty. An OCEAN can be calm, rough, warm, cold, and is always salty. A challenge for researchers interested in conceptual knowledge is to predict what features will be activated in a given instance. One way to navigate the flexibility of conceptual knowledge is to observe what information is activated when concepts combine. The conceptual features that are activated or strengthened in complex concepts (e.g., “warm ocean”, “summer ocean”) can tell us something about how conceptual information is structured and also how it is flexibly used. For example, consider the different effects of “salty” in “salty cookie” and “salty ocean”: salty oceans are much saltier than salty cookies, but comprehension of “salty cookie” involves activating rare and thus potentially more informative features of cookies. Here we explore whether interactions between features and noun concepts can predict how features are represented in combined concepts.

Conceptual combination involves the selection of appropriate conceptual features and the integration of those features into the head concept. This is especially true for noun-noun compounds (e.g., “needle grass”) and figurative phrases (e.g., “The grass feels like needles”). In both of these phrases, comprehension involves selecting appropriate NEEDLE features that are most likely to convey the speaker’s intended meaning (e.g., SHARP) while excluding others (e.g., METAL). These relevant features must then be applied to the GRASS concept. A rigorous scientific exploration of conceptual combination will separately characterize the cognitive processes for selection and integration, especially if the goal is to understand which brain regions contribute to elements of flexible language and concept use. Our current goal is to explore how features are flexibly integrated in combined concepts, and thus want to target the integration component of conceptual combination specifically. To do this, we focus on adjective-noun combinations (e.g., “sharp grass”), which minimizes selection demands and lets us explore how conceptual feature information is flexibly modulated during comprehension. We will review previous cognitive and computational models of adjective-noun combinations before explaining our own approach in more detail. To avoid repetitiveness, we will use “features” and “properties” interchangeably.
A prominent source of divergence between early cognitive theories of combined concepts is whether or not the outputs of a combination can be explained as a function of the individual constituents alone (i.e., in a “closed operation”) or whether knowledge external to the constituents is required. Smith et al.’s (1988) selective-modification theory provides a well-defined model of adjective-noun modification in which the only inputs are the adjective and noun constituents being combined. A concept (e.g., APPLE) is represented as a collection of slots and fillers, which relate to feature dimensions (e.g., color, shape) and specific features (e.g., RED, ROUND), respectively. Each feature has a salience value that indicates its strength for that concept (e.g., RED has higher salience than BROWN for APPLE). Comprehension of adjective-noun phrases involves increasing the salience of a feature and the diagnosticity of the associated slot in an instance of the noun concept. For example, in the phrase “brown apple”, the adjective “brown” increases the salience of BROWN and the diagnosticity of color in the created APPLE instance. The success of this model lies in its ability to make well- formalized predictions about how concepts will combine, as well as in its success at capturing similarity and typicality judgments (Smith & Osherson, 1984; Smith et al., 1988). The simplicity of this model is also appealing: combined concepts are computed in a closed operation, without referring to knowledge external to the concepts being combined.

The simplicity of this model also has its disadvantages. Proponents of the concept specialization theory (Cohen & Murphy, 1984; Murphy, 1988; Wisniewski & Gentner, 1991) argue that Smith et al.’s (1988) model cannot capture the flexibility and context-dependence of adjective meaning. The meaning of an adjective can vary when paired with different nouns in stark or subtle ways. For example, consider the differences between “fresh vegetable”, “fresh shirt”, and “fresh idea” (Murphy & Andrew, 1993), and between “red truck”, “red face”, and “red fire” (Halff et al., 1976). Additionally, conceptual features seem to interact with each other and this aspect of conceptual structure affects how people interpret combined concepts. For example, people assume that “wooden spoons” are large, whereas “metal spoons” are small (Medin & Shoben, 1988). Therefore, critics of the selective modification model argue that its unstructured conceptual representations cannot explain the feature interactions observed in conceptual combination (Wisniewski & Gentner, 1991; Medin & Shoben, 1988), and that the flexibility of adjective meaning cannot be explained in a closed operation but requires the incorporation of additional knowledge (Cohen & Murphy, 1984; Murphy, 1988). On the one hand, it is clear that exploring feature interactions and correlations is critical to understanding conceptual combination and conceptual processing more generally, and conceptual structure has been modeled in various ways (e.g., McClelland & Rogers, 2003; Tyler & Moss, 2001; McRae et al., 1997; 1999; Sloman et al., 1998; Solomon et al., 2019). However, it is unclear how the construct of “world knowledge” increases our understanding of how
concepts combine. The content of this external knowledge is poorly defined and the way it influences the combinatorial process is unspecified. Our goal is to understand and predict flexible adjective meaning within a well-defined framework. More specifically, we want to represent concepts and their features in a way that enables flexible meaning to emerge when concepts combine.

A relevant class of models can be found in distributional semantics, in which word meanings are represented as vectors derived from word co-occurrence statistics in language (e.g., Lund & Burgess, 1996; Landauer & Dumais, 1997; Michell & Lapata, 2010; Baroni & Zamparelli, 2010). In these models, vector representations of adjectives, nouns, and adjective-noun phrases can be extracted from large text corpora. Researchers can compare the success of different compositional models (e.g., additive, multiplicative) in predicting the representations of combined phrases, using the adjective and noun vectors as input. These computational models are well-defined: various functions can be applied to constituent representations in a closed operation in order to make precise, quantitative predictions about the representation of the combined concept. Mitchell & Lapata (2010) contrast different combinatorial functions and provide empirical support for a simple multiplicative model. Baroni & Zamparelli (2010) also use a distributional semantics approach to define adjectives as matrices which are applied to noun representations in order to generate representations of combined phrases that can potentially reflect flexible adjective meaning. Compositional models in distributional semantics thus have the potential to make quantitative predictions of combined concepts; however, representing word and phrasal meanings based on word co-occurrence statistics makes it difficult if not impossible to know what featural information is actually being modified in the combined concept. In other words, it is difficult to associate the text-based representations of word meaning in distributional semantics with the cognitive and neural representations people actually use to understand concepts.

In cognitive neuroscience investigations of conceptual combination, researchers have characterized the neural regions recruited in the combinatorial process. For example, a substantial body of work on conceptual combination implicated the anterior temporal lobe (ATL; Bemis & Pylkkänen, 2011; 2012; 2013; Westerlund et al., 2015; Baron et al., 2010; Baron & Osherson, 2011; Boylan et al., 2017) and angular gyrus (AG; Bemis & Pylkkänen, 2012; Price et al., 2015, 2016; Boylan et al., 2015, 2017; Graves et al., 2010). In most of these studies, neural sensitivity to semantic composition is determined by contrasting different kinds of compositional phrases (e.g., “red meat”, “eats meat”) with non-compositional phrases (e.g., “### meat”). The motivation behind this approach is to characterize the neural regions involved in different kinds of conceptual combination.
Another approach is to analyze the information contained in the neural responses evoked by conceptual combination. Two related fMRI studies found evidence that multivoxel patterns (MVPs) in left ATL represent combined concepts (Baron et al., 2010; Baron & Osherson, 2011). Specifically, Baron & Osherson (2011) determined that MVPs in the left ATL during processing of a “combined” concept (e.g., GIRL) could be predicted by an additive or multiplicative combination of the multivoxel patterns elicited by the simple constituents (e.g., YOUNG + FEMALE). In another multivariate fMRI approach, Chang et al. (2009) analyzed distributed neural responses to combined concepts. A reduced set of voxels (n=120) were selected based on a measure of voxel “stability”, in which the patterns elicited by adjective-noun combinations (e.g., “soft bear”, “plastic bottle”) and their constituent concepts were analyzed. The distributed patterns of activity in these voxels were more successfully predicted using a multiplicative model of feature-based constituent representations (see Mitchell et al., 2008), rather than an additive model. The relative success of the multiplicative model implies that these voxels captured the modified noun meaning, rather than a superimposition of the constituent adjective and noun concepts (Chang et al., 2009). Taken together, these findings suggest that meaningful representations of adjective-noun phrases can be captured in patterns of activity in fMRI data.

We are interested in how conceptual information is modulated in combined concepts. The processes involved in conceptual combination, and the resulting modulation of conceptual features, will be constrained by conceptual structure. That is, the way in which information is structured within individual concepts will influence the ways in which that information will combine. A broad, theoretical goal is thus to propose a theory of conceptual structure that can explain previously reported empirical phenomena relating to flexible concept use (e.g., Halff et al., 1976; Medin & Shoben, 1988; Springer & Murphy, 1992; Yee & Thompson-Schill, 2016). Here we use ideas from information theory (Shannon, 1948) to propose elements of a feature-based conceptual structure that might relate to flexible modulation of information. Specifically, we use the information theoretic measures of surprisal and entropy (e.g., Hume & Mailhot, 2013) to represent the surprisal and uncertainty, respectively, of features given a particular concept (Fig. 12A-B). Surprisal is defined as the negative log probability of a feature given a concept: a high surprisal feature-noun pair indicates that the feature very rarely applies to the noun concept. For example, the features METAL and ORANGE have high and low surprisal for PUMPKIN, respectively (Fig. 12A).

Entropy has previously been adopted in the fields of behavior theory (Berlyne, 1957), psycholinguistics (Hale, 2006), phonology (Hume & Mailhot, 2013) and natural language processing (Berger et al., 1996). Here, we use entropy as a measure of feature uncertainty, such that the highest entropy values are
observed when a feature applies to a noun concept about half of the time. Such features are highly uncertain for a given noun concept, because they could be true or false of any given noun instance. For example, **GREEN** has high uncertainty for **PUMPKIN** because pumpkins are green only some of the time (Fig. 12B). On the other hand, there is low uncertainty for **ORANGE** (because pumpkins are very likely to be orange), and also low uncertainty for **METAL** (because pumpkins are very unlikely to be metal). Importantly, this measure of uncertainty diverges from our measure of surprisal, enabling us to determine which of these is a better predictor of feature modulation in adjective-noun combinations.

Previous research suggests that word meaning is flexible (e.g., Yee & Thompson-Schill, 2016; Casasanto & Lupyan, 2015; McElree et al., 2001) and this extends to the flexibility of adjective meaning in combined concepts (Halff et al., 1976; Springer & Murphy, 1992; Baroni & Zamparelli, 2010). Representations of adjective-noun phrases can be analyzed across distributed neural patterns using fMRI (Chang et al., 2009; Baron et al., 2010; Baron & Osherson, 2011), and these patterns appear to reflect a modification of the noun concept (Chang et al., 2009). Our current goal is to predict the fluctuations of conceptual feature information during comprehension of adjective-noun phrases, as reflected in distributed patterns of neural activity, using the measures of surprisal and uncertainty described above.

We selected a set of eight noun concepts (**KEY**, **TABLE**, **PUMPKIN**, **GRASS**, **COOKIE**, **PICKLE**, **PILLOW**, **KNIFE**) and eight adjective concepts (**METAL**, **WOODEN**, **ORANGE**, **GREEN**, **SWEET**, **SALTY**, **SOFT**, **SHARP**), and fully crossed them to create 64 adjective-noun combinations (see Fig. 10). Analyzing all possible combinations eliminates bias on the part of the experimenter (Murphy, 1988), and enables a wide range of feature-noun interactions to emerge. We observed distributed neural patterns evoked by the individual concepts (e.g., “cookie”) as well as the combinations (e.g., “salty cookie”), and derived a quantitative measure reflecting the extent to which the adjective increased the adjective-denoted feature information in the combination. For example, we determined the degree to which **SALTY** information increases in “salty cookie” relative to “cookie” alone. We then explored whether feature surprisal or feature uncertainty could predict the observed degree of feature modulation observed in neural activity. That is, we set out to explore the feature-noun relationships that contribute to the neural representations of combined concepts. Understanding how individual features are modulated across contexts is one important step in understanding the suite of processes involved in conceptual combination and flexible concept use.
3.2 Methods

3.2.1 Participants
Eighteen right-handed participants (10 female) from the University of Pennsylvania community completed the fMRI study and were compensated $20/hour for their time. An additional 46 participants completed online surveys on Amazon Mechanical Turk (AMT) and were compensated according to standard rates. Consent was obtained for all participants in accordance with the University of Pennsylvania IRB.

3.2.2 Adjective and noun stimuli
We explored the pairwise relationships between eight adjective (property) concepts (METAL, WOODEN, ORANGE, GREEN, SWEET, SALTY, SOFT, SHARP) and eight noun concepts (KEY, TABLE, PUMPKIN, GRASS, COOKIE, PICKLE, PILLOW, KNIFE). These adjective and noun concepts were fully crossed, resulting in 64 adjective-noun combinations. These particular stimuli were selected such that each noun (e.g., “pumpkin”) was matched with at least one adjective (e.g., “orange”) whose corresponding property was highly associated with the noun concept. We also selected adjectives whose corresponding properties spanned four property dimensions (i.e., material, color, taste, texture), such that our experiment would not be restricted to one type of information but would apply to property integration in conceptual combination more generally.

Adjective concepts (e.g., “metal”), noun concepts (e.g., “knife”), and adjective-noun combinations (e.g., “metal knife”) were displayed in verbal format. The words or phrases were presented in black, capitalized, Arial font in the center of the screen on a white background. In the fMRI study, participants viewed noun-words in Run 1, adjective words in Run 4, and adjective-noun combinations in Runs 5, 6, and 7. The order of presentation was pseudo-randomized; eight participants were presented with stimuli in reverse order within each run to minimize order effects.
Adjective concepts and noun concepts were also displayed in an image-based format. The image stimuli comprised real-world photographs collected from Google that represented the appropriate concept. For noun concepts, this resulted in a collection of 20 images/noun that represented examples of each object (e.g., a set of images depicting kinds of cookies); for adjective concepts, this resulted in a collection of 30 images/adjective in which the corresponding property was represented (e.g., a set of images depicting metal things). None of the adjective concept images contained any of the eight object concepts of interest. These images were presented in the center of the screen on a white background. In the fMRI study, participants viewed noun-image stimuli in Run 2 (these data were not analyzed in the current study) and adjective-image stimuli in Run 3. Images were presented in a blocked design by concept. The order of blocks was pseudorandomized; eight participants were presented with blocks in a reverse order within each run to minimize order effects. No image was presented more than once.

### 3.2.3 Feature surprisal and feature uncertainty

We were interested in characterizing unique interactions between the targeted adjectives and nouns. We specifically determined the extent to which each adjective-denoted property is surprising or uncertain for each noun concept using the information theory measures of surprisal and entropy (Shannon, 1948). Both of these measures can be calculated based on a measure of probability. In our

---

**Figure 10: Adjective and noun stimuli.** We collected multivoxel patterns evoked by the verbal labels of eight noun and eight adjective concepts. We also collected multivoxel patterns elicited by visual depictions of the adjective concepts. These patterns were used in our calculation of neural property modulation.
case, we needed to know each feature’s probability with respect to each noun concept. We collected subjective ratings of probability in an online AMT survey (N=46), in which participants were asked whether each noun concept has each property (e.g., “Is a pumpkin orange?”) and responded on a sliding scale from 1-100 ranging from “Definitely No” to “Definitely Yes”. These responses were averaged across participants and scaled between 0 - 1.

Surprisal is calculated as negative log probability (Fig. 12A). If the probability (P) of a feature given a noun concept is low, then feature surprisal will be high (and vice versa). We calculated feature surprisal (S) using the equation:

\[ S = - \log_2(P) \]

Entropy is a measure of uncertainty, and can similarly be calculated from probability values (Fig. 12B). If the entropy (E) of a feature given a noun concept is high, then feature uncertainty will be high. We calculated feature uncertainty (E) using the equation:

\[ E = [-P \times \log_2(P)] + [-(1-P) \times \log_2(1-P)] \]

These methods resulted in a feature surprisal and feature uncertainty value for each of the 64 adjective-noun pairs. The relationships between probability and each of these two measures is visualized in Fig. 12A-B, along with the specific surprisal and uncertainty values for each of the 64 adjective-noun combinations.

3.2.4 Analyzing concept network core-periphery structure

We also characterized unique relationships between features and noun concepts using the concept network models introduced and reported in Solomon et al. (2019). In this approach, individual concepts are represented as networks, in which nodes represent conceptual properties and edges capture within-concept property associations. A range of global (concept-level) and local (property-level) measures can be extracted from these networks, and in Solomon et al. (2019) we observed that the global core-periphery network structure relates to the semantic diversity of the concept label (Hoffman et al., 2013). Here, we examined local core-periphery structure — whether individual properties are in the “core” or “periphery” of a given concept network — and asked whether adjective-noun combinations differentially modulate property activations when the relevant property is in the concept’s core vs. periphery. Peripheral properties are more loosely associated with other properties for a given concept; these properties might therefore be more uncertain for the associated noun. Thus, our prediction is that peripheral properties will exhibit increased property modulation relative to core properties during comprehension of adjective-noun combinations.
Networks for seven of our noun concepts (all except TABLE) were constructed on the basis of the same set of property nodes, and therefore are comparable. Within each of these seven concept networks, we determined which of our current adjective-denoted properties (i.e., METAL, WOODEN, ORANGE, GREEN, SWEET, SALTY, SOFT, SHARP) were represented as nodes. The network-construction methods in Solomon et al. (2019) resulted in some properties being excluded from each of the final network models, so not all properties are represented for all concepts. Out of the 64 total adjective-noun combinations tested in the current study, there were 35 corresponding feature-noun relationships that were captured in the network models. For each of these 35 feature-noun pairs, we determined whether the property was located in the concept’s core or periphery (Rubinov et al., 2015). There was an approximately even distribution of core (N=16) and periphery (N=19) feature-noun pairs.

3.2.5 Task order and design

Participants in the fMRI study completed seven scanning runs; these runs contained different stimuli and tasks. Participants were exposed to noun-words (Run 1), noun-images (Run 2), adjective-images (Run 3), adjective-words (Run 4), and adjective-noun combinations in verbal form (Runs 5, 6, 7). We did not use the noun-image data in the analyses reported here, so will not discuss it further.

In the noun-word (Run 1), adjective-word (Run 4), and adjective-noun combination (Runs 5-7) scans, participants performed a memory task. Lists of six words (or eight adjective-noun phrases) were presented in an event-related design. Each stimulus was presented in the center of the screen for 2 seconds followed by a jittered fixation ISI (2-12 seconds). After eight stimuli were presented, participants were prompted with a word (or phrase) and were asked to respond whether the cued stimulus was on the previous list, by pressing a button on a hand-held response box. Each run contained eight lists; each individual word or phrase was seen by the participant three separate times. In the adjective-noun combination runs, each combination occurred once per run. This task design enabled us to collect verbally-evoked neural patterns of activity for each noun, adjective, and adjective-noun combination.

In the adjective-image scan (Run 3), participants performed an attention task. Stimuli were presented in a blocked design, with blocks corresponding to adjective concepts. Each block began with a cue specifying the dimension to which they should attend (e.g., “color”), displayed in the center of the screen. After the cue, a series of images were presented (6-24 images/block) that represented one of the eight target properties (e.g., GREEN). In half of the blocks there was an additional image added at the end of the block that depicted a property different than the target property (e.g., the last image in a GREEN block would depict a blue object). After participants viewed the series of images in
each block, they were asked whether the last image was different from the others with respect to the cued property-dimension. Each property concept was seen in two separate blocks. Three 12-second fixation blocks were distributed throughout the run. This task design enabled us to collect visually-evoked neural patterns of activity for each adjective concept.

3.2.6 Functional MRI acquisition and analysis

fMRI data were collected on a 3-T Siemens Trio System and 32-channel array head coil. Structural data included axial T1-weighted localizer images with 160 slices and 1 mm isotropic voxels (TR = 1620 ms, TE = 3.87 ms, TI = 950 ms, FOV = 187 x 250 mm, Flip Angle = 15°). Functional data included seven acquisitions of echo-planar FMRI using a multiband sequence performed in 42 axial slices and 2 mm isotropic voxels (TR = 2000 ms, TE = 30 ms, FOV = 234 x 234 mm, Flip Angle = 90°).

Data were pre-processed and analyzed using FSL. Pre-processing included motion-correction using MCFLIRT, spatial smoothing with a Gaussian kernel of FWHM 5 mm, and high-pass temporal filtering. Motion outliers were modeled as covariates of no interest. The runs with verbal stimuli, which were presented in an event-based design, were analyzed with a GLM including item-level regressors modeling the individual TRs for each concept or combination contrasted against the fixation baseline; this resulted in whole-brain beta maps for the 8 adjectives, 8 nouns, and 64 adjective-noun combinations for each subject. Since the adjective-noun combinations were presented across three different runs, beta values were averaged across runs to result in a single beta map for each of the 64 combinations for each subject. The runs with image stimuli, which were presented in a blocked design, were analyzed with a GLM including item-level regressors modeling the two blocks for each concept (6-24 seconds/block) contrasted against the fixation blocks. Individual subjects’ data were transformed to MNI standard space.

3.2.7 Searchlight voxel selection

Our goal was to observe the amount of property modulation evoked by each adjective-noun pair. To do so, we needed to capture the multivoxel patterns (MVPs) that corresponded to each of our target adjective concepts. Our adjectives denoted properties in different dimensions (i.e., material, color, taste, texture), and it is likely that these kinds of information are represented across different populations of voxels. We therefore employed four different searchlight analyses to find voxels that are sensitive to each of these four dimensions (Fig. 11A).

Our searchlight analyses are based on the theoretical assumption that both words and images should evoke conceptually-relevant information. For example,
there should be overlap in neural activity when people read the word “green” and when they visually perceive green things: this overlap in neural representation is amodal with respect to stimulus presentation and is conceptually meaningful (Fairhall & Caramazza, 2013). We therefore sought out to find the voxels that responded similarly to the word-evoked and image-evoked presentations of the same adjective concept compared to presentations of two different adjective concepts.

For each dimension, we created a representational similarity analysis model (Kriegeskorte et al., 2006, 2008) that predicted high similarity for within-adjective correlations (r=+1; e.g., GREEN-word, GREEN-image), high dissimilarity for between-adjective correlations within the appropriate dimension (r=-1; e.g., GREEN-word, ORANGE-image), and baseline similarity for between-dimension property correlations (r=0; e.g., GREEN-word, METAL-image). Within each 123-voxel searchlight, we correlated the eight adjective-word patterns with the eight adjective-image patterns (spearman’s correlation), resulting in an 8 x 8 MVP similarity matrix. We correlated these 64 neural correlation values with the 64 model-predicted values to assess how well the neural patterns captured stimulus-independent property information for each dimension. For each dimension, the correlation between the MVP similarity matrix and the dimension-specific model was assigned to the voxel at the center of the searchlight; the searchlight was swept over all voxels in the brain and the analysis was repeated in each one. This resulted in four whole-brain maps for each subject, indicating which voxels best captured information within each of the four property dimensions. These maps were averaged across subjects, resulting in four group-level dimension maps. We selected our four voxels of interest (VOIs) by extracting the top 100 voxels within each dimension. Because this was a voxel-selection technique and not a statistical test, we did not assess voxel-wise significance. In the subsequent analyses, we computed property activation and modulation within the appropriate VOI; for example, we calculated the activation of GREEN during comprehension of “green pumpkin” within the color VOI, whereas we calculated the activation of METAL during comprehension of “metal pumpkin” within the material VOI.

### 3.2.8 Property modulation in adjective-noun combinations

We wanted to calculate the extent to which the activation of the adjective-denoted property increased in an adjective-noun combination, relative to the noun concept. For example, how much more GREEN is contained in the representation of “green pumpkin” than in “pumpkin” alone? In order to calculate this difference in property activation — which we refer to as “neural property modulation” — we first needed to calculate the degree of property activation in the noun and adjective-noun items. Our approach was based the assumption...
stated above — that correlations between word-evoked and image-evoked MVPs can capture conceptual similarity.

Specifically, our measure of property activation in unmodified noun concepts is calculated by correlating image-evoked adjective MVPs and word-evoked noun MVPs (Fig. 11B). High correlation values indicate high similarity between neural patterns; we interpret a high similarity between noun- and adjective- patterns to reflect that the adjective-denoted property is represented in the noun-MVP. For example, to calculate the degree of GREEN activation in “pumpkin”, we correlate the GREEN-image MVP with the “pumpkin”-word MVP. This results in 64 property-activation values for each subject (i.e., the activation of each of the properties in each of the noun concepts). These values are averaged across subjects to result in group-level data reflecting the degree of property activation in each unmodified noun concept. Note that these are correlations between word- and image-evoked neural patterns; the similarity cannot be driven by confounding variables such as orthographic similarity.

We use the same method to calculate the activation of properties in adjective-noun combinations (e.g., correlate GREEN-image MVP with “green pumpkin” MVP). These data are similarly averaged across subjects, resulting in group-level data reflecting the degree of property activation in each adjective-noun combination. We can then calculate, for each of the 64 combinations, the extent to which the adjective-denoted property increased in the combination relative to the noun; this is calculated by subtracting the property activation in the noun from the property activation in the combination. This our measure of neural property modulation. For example, to calculate the amount of green modulation during comprehension of “green pumpkin”, we subtracted the level of GREEN activation in “pumpkin” from the level of GREEN activation in “green pumpkin.” This reflects the extent to which the property GREEN is modulated during comprehension of “green pumpkin.” Each of these analyses are carried out in the appropriate VOI: modulations of ORANGE and GREEN are assessed in the color VOI, modulations of METAL and WOODEN are assessed in the material VOI, etc. We thus have a neural measure of property modulation for each of the 64 adjective-noun combinations, and aim to determine whether specific interactions between features and noun concepts can predict this observed neural property modulation.
3.2.9 Validating VOIs and analysis methods

We defined our VOIs using the searchlight analysis described above, in which we assessed the sensitivity of voxels to our property dimensions (i.e., material, color, taste, texture). Our main goal was to observe the amount of property modulation within the corresponding VOIs, and specifically to determine whether feature-noun relationships can predict the amount of property modulation during adjective-noun comprehension. However, we first wanted to validate our voxel-selection technique by establishing that the VOIs represent the kinds of information we expect. We therefore used the property modulation analysis described above to calculate the degree of property modulation for all 64...
combinations within each VOI. The expectation was that a given VOI (e.g., color-sensitive voxels) would reveal greater property modulation for properties within the VOI’s dimension (i.e., ORANGE and GREEN) than for properties outside of that dimension (e.g., METAL, SWEET, SHARP). Within each VOI we calculated the average degree of property modulation during adjective-noun comprehension for within-dimension and between-dimension properties. Our VOIs did exhibit greater property modulation for within-dimension properties ($t(3)=4.26$, $p=0.024$). This confirms that our VOIs are differentially sensitive to our property dimensions, and validates our voxel-selection and analysis methods. In all other analyses, we only calculated property modulation for within-dimension properties (e.g., we only analyzed modulation of ORANGE and GREEN in the color-VOI, SWEET and SALTY in the taste-VOI, etc.).

3.2 Results

3.3.1 Feature uncertainty predicts degree of property modulation

Our main question is whether either feature surprisal or feature uncertainty can predict the amount of neural property modulation during comprehension of adjective-noun phrases. We capture these feature characteristics in the information theory measures of surprisal and entropy (Shannon, 1948) and define these values for the feature-noun pairs corresponding to our 64 adjective-noun combinations. For each combination, we used multivoxel pattern analysis of fMRI data to calculate the degree of neural property modulation within dimension-appropriate VOIs (see Methods). For example, we calculated the extent to which the property SALTY is increased during comprehension of “salty cookie” relative to “cookie.” We subsequently tested whether feature surprisal or uncertainty predicted increased property modulation. We observed a positive relationship between feature uncertainty and the degree to which property activation increases during comprehension of adjective-noun combinations ($r=0.26$, $p=0.035$; Fig. 12D). This result was observed in our original 100-voxel VOIs but was robust between VOIs of 50-190 voxels. We did not observe a positive relationship between feature surprisal and property modulation; if anything, the relationship was negative ($r=-0.22$, $p=0.08$; Fig. 12C). This is in part explained by the negative relationship between uncertainty and surprisal ($r=-0.78$, $p<0.001$); we therefore wanted to additionally test whether the relationship between uncertainty and property modulation was significantly stronger than the relationship between surprisal and property modulation. We confirmed that the correlation between uncertainty and property modulation was stronger than the
correlation between surprisal and property modulation in a Steiger’s Z test of dependent correlations ($Z_{H}=2.05, p=0.04$). These results suggest that feature uncertainty, and not feature surprisal, predicts neural property modulation during comprehension of adjective-noun combinations.

We ran an additional analysis to confirm that feature uncertainty influences the specific, targeted activation of the adjective-denoted property. An alternative possibility is that feature uncertainty more generally influences the activation of all conceptual features in a combined concept. In each of the 64 combinations, we calculated the property modulation for all eight properties (e.g., calculated the modulation of METAL, WOODEN, ORANGE, GREEN, SWEET, SALTY, SOFT, and SHARP
for “salty cookie”). We averaged these eight values to calculate a general modulation value for each combination. These general modulation values were not predicted by either feature uncertainty ($r = -0.13$, $p = 0.29$) or surprisal ($r = 0.20$, $p = 0.12$), suggesting that our main observed relationship between feature uncertainty and property modulation reflects a targeted modulation of the adjective-denoted property, and not the modulation of the noun concept’s information more broadly.

3.3.2 Concept network structure predicts degree of property modulation

For 35 of the adjective-noun combinations, we determined whether the adjective-denoted property is in the concept network’s core ($n=16$) or periphery ($n=19$). We compared mean levels of neural property modulation between core- and periphery-properties, and found a greater increase in adjective-denoted information for periphery-properties vs. core-properties ($t(33)=2.19$, $p=0.036$).

Figure 13: Concept network analysis. We exploited the networks reported by Solomon et al. (2019) to assess adjective-noun relationships in a network-based framework. For 35 of our combinations, we could determine whether the adjective denoted a property in the concept network’s core ($n=16$) or periphery ($n=19$). We compared mean levels of neural property modulation between core- and periphery-properties, and found a greater increase in adjective-denoted information for periphery-properties vs. core-properties ($t(33)=2.19$, $p=0.036$).
during comprehension of adjective-noun combinations. More generally, it implies that the structure of feature representations within noun concepts has consequences for how that feature information will be flexibly activated in different language contexts.

3.3.3 Neural representation of properties in unmodified noun concepts

While our main empirical question was addressed above, we ran additional analyses to observe the degree of property activation in the unmodified noun concepts. For example, we calculated how strongly we could detect the activation of GREEN, SWEET, and SOFT during comprehension of the noun “pickle” in isolation. First, we predicted that there would be more activation of core properties than periphery properties, since core properties are more certain and consistent for a given concept. Using the same MVP analysis described above, we did observe that properties present in a concept’s network core are spontaneously more active during comprehension of the unmodified noun, as opposed to properties found in the concept network’s periphery ($t(33)=2.31$, $p=0.026$).

We had no a priori predictions regarding relationships between our information theory measures (i.e., surprisal, uncertainty) and property activation in unmodified nouns. However, if they do exist, negative relationships would be more interpretable than positive ones: we might expect decreased activation of uncertain and/or surprising features when the noun concept is presented alone. We did observe a strong negative relationship between feature uncertainty and property activation in the noun concept ($r=-0.43$, $p<0.001$) and a significant positive relationship between surprisal and property activation in the noun concept ($r=0.32$, $p=0.01$). As above, we ran a Steiger’s Z test on the difference between these dependent correlations: the relationship between uncertainty and property activation in the noun concept was significantly stronger than the relationship observed with feature surprisal ($Z_H=3.28$, $p=0.001$). We thus interpret the significant effect of feature surprisal to be explained by its negative relationship with feature uncertainty. These results suggest that when a noun concept is comprehended, features that are possible but uncertain for that noun concept are suppressed. This would in turn enable activation of this feature to substantially increase when the uncertainty of the feature is resolved, such as in adjective-noun combinations when the feature is directly asserted. This is an interesting line of inquiry to explore in future work.

3.2 Discussion

We tested whether feature surprisal or uncertainty predicted the degree of neural property modulation elicited by comprehension of adjective-noun phrases. Both
of these measures are derived from information theory (Shannon, 1948); in the present study, both characterize relationships between features and noun concepts. Surprisal indicates the extent to which an adjective-denoted property is unexpected given a noun concept, whereas entropy indicates the extent to which an adjective-denoted property is uncertain given a noun concept. We found that uncertainty in the feature-noun relationship positively predicted the increased neural activation of conceptual properties in the adjective-noun combination. The unexpectedness of a feature-noun pair did not predict these neural effects. These results suggest that, at least in the property-sensitive voxels we analyzed, the uncertainty of the adjective-denoted property, rather than surprisal, is informative for the combinatorial process. If there is uncertainty reflecting whether a certain property (e.g., GREEN) applies to an instance of a noun (e.g., PUMPKIN), then asserting this property in an adjective-noun phrase (e.g., “green pumpkin”) reduces this uncertainty and is therefore informative.

In this study, we compared our measures of surprisal and entropy with a neural measure of property modulation. Our measure of property modulation reflected the degree to which a property (e.g., SWEET) is more activated in a combination (e.g., “sweet pickle”) than in the noun alone (e.g., “pickle”). We calculated property activation in both the nouns and adjective-noun phrases by assessing the similarity between each of their corresponding multivoxel patterns (MVPs) with the MVPs representing the adjective-denoted properties (e.g., MVPs elicited by images of sweet things). We then calculated the extent to which, relative to the unmodified noun, property activation increased in the adjective-noun combination.

If properties are increased in adjective-noun combinations in relation to their unexpectedness or novelty, then our measure of surprisal would have positively predicted our measure of neural property modulation. For example, METAL would be a very surprising property for the noun concept PUMPKIN, and so we might have observed a substantial increase in METAL activation during processing of “metal pumpkin.” In line with this hypothesis, we would expect less property modulation in combinations asserting less-surprising properties (e.g., “green pumpkin”, “orange pumpkin”). This might be the prediction stemming from early feature-based models of conceptual combination, in which the salience of the adjective-denoted property was increased in the noun concept (Smith et al., 1988). However, we did not observe a positive relationship between surprisal and property modulation; if anything, the relationship between these two measures was negative. Thus, in the property-sensitive voxels we analyzed, increased property modulation was not explained by an increase in property-object surprisal, or unexpectedness.
We did find that feature uncertainty positively predicted neural property modulation. Even though METAL is a surprising property for PUMPKIN, it has low uncertainty because pumpkins are never metal (or if they are metal, then they might not be considered to be pumpkins; c.f. Wisniewski & Gentner, 1991). The property ORANGE is neither surprising nor uncertain for PUMPKIN, because pumpkins are orange most of the time. On the other hand, GREEN has high uncertainty because sometimes pumpkins are green, and sometimes they are not. Our results suggest that the uncertainty present in a feature-noun pair is a meaningful source of information in the combinatorial process and that it may influence the extent to which activation of the adjective-denoted property is increased during comprehension of an adjective-noun phrase.

The observed relationship between uncertainty and property modulation points to a combinatorial mechanism in which properties are flexibly modulated depending on the adjective and noun being combined. The specific relationship between the adjective-denoted property and the noun concept will influence the ways in which the resulting combination is represented. In other words, an adjective does not have the same effect across all nouns, rather the meaning or effect of the adjective will change across noun contexts. This flexibility or context-dependence of adjective meaning relates to theories of conceptual flexibility (Yee & Thompson-Schill, 2016; Casasanto & Lupyan, 2015; Solomon et al., 2019), as well as related theories of type-shifting or coercion in the language domain (e.g., McElree et al., 2001; Pustejovsky, 1998; Traxler et al., 2005). Classic theories of conceptual combination (e.g., Smith et al., 1988; Wisniewski & Gentner, 1991; Murphy, 1998) reference this flexibility inherent in the combination process and rely on notions of “world knowledge” in order to explain these effects. In a theoretical approach similar to our own, Halff et al. (1976) reports varied effects of “red” in the contexts of e.g. “fire”, “face”, and “sunrise” and claims that features are represented as ranges of numerical values, rather than point estimates. In this view, concepts themselves can contain the structure necessary to explain flexible activation of feature information. In the present study, we observed flexible adjective meaning without relying on world knowledge: the relationship between uncertainty and flexible property modulation could be explained within a closed operation whose only inputs are the adjective and noun constituents and their corresponding conceptual structures. That is, we can capture flexibility of adjective meaning without referring to information external to the two concepts being combined.

Though we did not observe a relationship between surprisal and property modulation, this relationship could in theory also imply a flexible combinatorial mechanism. However, a positive relationship between surprisal and property modulation would permit an alternative, non-combinatorial interpretation. If the voxels we analyzed represented only the adjective during comprehension of the
adjective-noun phrase, then we would observe a positive relationship between surprisal and property modulation but not due to a combinatorial process (but because surprisal is an inverse of probability). For example, the property SWEET might be equally represented across all combinations (e.g., “sweet key”, “sweet pumpkin”, “sweet cookie”), but we could observe the greatest SWEET increase in the high-surprisal combinations (e.g., “sweet key”) simply because SWEET was not present at all in the unmodified noun (e.g., “key”). In other words, an increase in SWEET activation might not be a result of the flexible modulation of SWEET in KEY, but rather a result of how property modulation was calculated. The possibility that a representation reflects an adjective concept alone, rather than an integrated adjective-noun combination, is an important consideration in studies of combined concepts. This idea has been dealt with in distributional semantics, in which researchers attempt to predict the representation of a phrase (i.e., combined concept) based on high-dimensional word representations derived from text corpora (e.g., Mitchell & Lapata, 2008, 2010; Baroni & Zamparelli, 2010). The methods used in these approaches can be applied to cognitive neuroscience investigations of conceptual combination in order to ensure that observed results reflect the compositional processes of interest. Importantly, however, these concerns do not apply to our current observed results: a positive relationship between entropy and property modulation cannot result from an adjective-only representation, but implies a combinatorial process influenced by the unique representations of the adjective and noun concepts.

If the uncertainty of a feature-noun pair influences the outputs of the adjective-noun combinatorial process, it follows that this uncertainty is likely embedded in the associated conceptual representation(s). Feature-based models of conceptual structure have incorporated this uncertainty to various extents. Shifting away from classical theories, early probabilistic views of category structure claimed that concepts are not composed of necessary or defining features, but by characteristic or typical features (Wittgenstein, 1953; Rosch, 1975; Rosch & Mervis, 1975): category boundaries are “fuzzy”, and category membership can be uncertain. Rosch & Mervis (1975) found that the frequency with which a feature applies to category members relates to exemplar typicality. More recent feature-based accounts of conceptual knowledge represent concepts as weighted sets of stable features (e.g., McRae et al., 1997; Tyler & Moss, 2001; McClelland & Rogers, 2003); it is acknowledged that these representations are flexible in theory (O’Connor et al., 2009; Rodd et al., 2004), but flexibility is rarely implemented in these models. A recent neural network model incorporates flexibility of word meaning (Hoffman et al., 2018), though the fluctuations of individual feature representations across contexts has not yet been formalized. An alternative approach that might be able to capture feature uncertainty is to embed feature uncertainty in a probabilistic framework. For example, one could construct probabilistic feature-based models, in which
concepts are characterized by probability distributions over features. This idea might be analogous to the probabilistic topic models in distributional semantics (e.g., Griffiths et al., 2007), in which words are represented as a probability distribution over a set of topics. This approach also relates to probabilistic inference models of language comprehension (Goodman & Frank, 2016; Lassiter & Goodman, 2013). In this proposed probabilistic feature model, each feature within a concept would be represented as a probability distribution which captures the most likely feature value as well as the variance, or uncertainty, around that value. Representing feature uncertainty in this way might enable flexible feature modulations during conceptual combination to be directly predicted and tested within a probabilistic framework.

Recently, we proposed a feature-based concept network model in which individual concepts are represented as networks of features which can be selectively activated across different contexts (Solomon et al, 2019). These models reflect aspects of global structure of whole concepts, as well as the local structure of individual conceptual features. Aspects of this network structure might relate to feature uncertainty. In the present study, we used the concept networks described in Solomon et al. (2019) to analyze neural responses to adjective-noun combinations. We looked within comparable (noun) concept networks to see which of the eight adjective-denoted features were represented in those networks. Out of our total list of combinations, we were able to analyze 35 adjective-noun pairs in terms of their local network structure. Specifically, we determined whether the adjective-denoted feature was located in the noun concept’s network core or periphery (Rubinov et al., 2015). We observed increased neural property modulation when adjectives denoted properties in a concept’s periphery, rather than when those properties are found in the concept’s network core. This result highlights another way in which conceptual structure relates to the flexible adjustment of conceptual information in language contexts.

Here our goal was to determine whether the information-theoretic notions of surprisal or entropy could predict the extent to which conceptual information is adjusted in conceptual combination: we wanted to know whether either measure is informative in the conceptual combination process. There are other measures reported in previous work that capture similar relationships between objects and features, though the goal has typically been to characterize the properties that hold a privileged status in distinguishing between basic-level concepts, rather than to predict how property activations will fluctuate within a concept across language contexts. However, some of these related measures include cue validity (Bourne & Restle, 1959), informativeness (Devlin et al., 1998), distinguishingness (Cree & McRae, 2003), distinctiveness (Garrard et al., 2001, Tyler & Moss, 2001), and diagnosticity (Sedivy, 2003). Neural network models of concept representation and learning have highlighted the importance of
distinctive properties in conceptual knowledge and processing (Randall et al., 2004, Cree et al. 2006). In these approaches, distinctive properties are informative because they can be used to discriminate between similar basic-level concepts. In our case, we are interested in informativity as a predictor of flexible concept activation across different contexts. Our measures of surprisal and entropy are simpler than the measures mentioned above, because they reflect feature representations within single concepts, rather than statistics extracted over a large set of concepts. It is likely, however, that flexible concept activation is influenced by conceptual structure at many levels — on the level of features (e.g., sweet), concepts (e.g., cookie), and semantic domains (e.g., food) — and the relationship between conceptual structure and flexible concept use should be explored at all of these levels.

Though cognitive theories of conceptual combination focus on how features are mapped, integrated, or created in combined concepts (Wisniewski & Gentner, 1991; Springer & Murphy, 1992; Wisniewski, 1997; Smith et al., 1988; Estes & Glucksberg, 2000), there has not, to our knowledge, been previous attempts to directly observe the degree to which the neural representation of conceptual information is modulated during comprehension of complex phrases. Many cognitive neuroscience investigations of conceptual combination have illuminated distinct neural regions implicated in some aspect of the combinatorial process. Two regions that appear particularly implicated in conceptual combination tasks are the left anterior temporal lobe (ATL; Bemis & Pylkkänen, 2011; 2012; 2013; Westerlund et al., 2015; Baron et al., 2010; Baron & Osherson, 2011; Boylan et al., 2017) and left angular gyrus (AG; Bemis & Pylkkänen, 2012; Price et al., 2015, 2016; Boylan et al., 2015, 2017). In a MEG study, the left AG was sensitive to semantic composition relative to non-compositional phrases (e.g., “red boat” vs. “xqv boat”; Bemis & Pylkkänen, 2012). More specific characterizations of left AG response to conceptual combination has been offered by fMRI studies revealing that this region is sensitive to the plausibility of adjective-noun combinations (e.g., “plaid jacket” vs. “fast blueberry; Price et al., 2015), and to combinations containing events or relations (e.g., “eats meat” vs. “red meat”; Boylan et al., 2015).

In the left ATL, MEG studies have revealed this region’s sensitivity to general semantic composition (e.g., “red boat” vs. “xqv boat; Bemis & Pylkkänen, 2011; 2012; 2013; Westerlund et al., 2015), and that an increased response in left ATL to adjective-noun combinations is observed when the adjective modifies a low-specificity noun (e.g., “blue boat” vs. “blue canoe”) which thus results in increased specificity of the object concept (Westerlund & Pylkkänen, 2014). This suggests that left ATL may be involved in the combinatorial process of property integration or modulation, at least when the relevant properties relate to visual information. The left ATL also appears to contain representations reflecting the
output of the combination process: in the left ATL, patterns of fMRI activity evoked by combined concepts can be predicted by adding or multiplying the patterns evoked by the constituent concepts (Baron et al., 2010; Baron & Osherson, 2011). These fMRI and MEG results suggest that the left ATL could be involved in the property integration and property modulation targeted in the current paper. Though here we focused on distributed representations across the brain, a direct analysis of the left ATL in future work could further reveal how information is modulated in the representation of combined concepts.

Our decision to observe property modulation across distributed brain regions was motivated by the assumption that conceptual information is distributed across the brain and comprises many information types (e.g., visual, tactile) that are processed in various neural locations (e.g., Binder et al., 2016; Fernandino et al., 2015, Anderson et al., 2016). We did not aim to characterize the functions of specific neural regions (e.g., ATL, AG) during conceptual combination tasks, but to observe how conceptual information fluctuates in meaningful ways during this process. We thus analyzed distributed patterns of neural activity in voxels that were sensitive to our dimensions of interest (i.e., material, color, taste, texture). Previous work has analyzed distributed neural responses to adjective-noun combinations: Chang et al. (2009) use fMRI to analyze distributed patterns of neural responses to combined concepts and report that these patterns reflect a representation of the modified noun concept. Our current results confirm that information evoked by combined concepts can be detected across distributed sets of voxels, and further suggest that these patterns can be probed for meaningful changes in property activation during comprehension of combined concepts.

If feature uncertainty is relevant for feature modulation in adjective-noun phrases, it is likely that these same principles extend to other kinds of conceptual combination (e.g., noun-noun combinations, figurative language). In many combinations whether simple or complex, features are added, mapped, or integrated onto the head concept. For example, the interpretation of the noun-noun concept “zebra cake” is a striped cake (Wisniewski, 1997); in this case, the STRIPED feature is selected from ZEBRA and applied to CAKE. Additionally, in the figurative phrase “The train is a worm”, features are selected from WORM (e.g., ELONGATED, SLITHERS) and applied to TRAIN (Solomon & Thompson-Schill, 2017). Though these more complex combinations require the additional step of selecting the relevant feature(s), the feature(s) must then be integrated with the head concept in an appropriate way and to an appropriate degree. We hypothesize that the same feature integration process is present across combination types, and thus predict that feature uncertainty will be broadly relevant in conceptual combination.
The meaning of a phrase emerges out of the interaction between its constituents. Characterizing meaningful interactions between constituent concepts is an important step in understanding how the brain flexibly modulates information in different language contexts. Here we have shown that the uncertainty of a feature with respect to a given noun concept is an important interaction between constituents, and that it may influence the degree to which feature information is adjusted during comprehension of combined concepts. Representing conceptual features as probabilistic and uncertain may further our understanding of how information is flexibly modulated in complex language, and in conceptual processing more generally.
4: PROBABLISTIC MODELS OF FEATURE UNCERTAINTY PREDICT BEHAVIORAL AND NEURAL RESPONSES TO COMBINED CONCEPTS

4.1 Introduction

Human language relies on a deep reservoir of conceptual knowledge. Words (e.g., “diamond”) refer to concepts (e.g., DIAMOND) that contain information relating to knowledge about things in the world (e.g., diamonds are bright, sparkly, and expensive). It is rare for individual words to be used alone: most utterances comprise many words strung together, and one must combine the meanings of the underlying concepts in order to generate an appropriate interpretation. This is a complicated task because the meaning of a word is often influenced by the words surrounding it (Frege, 1884). That is, the information activated to represent a concept (e.g., DIAMOND) will be flexibly adjusted when the concept combines with other concepts in language (e.g., “dirty diamond”, “baseball diamond”). Here we use conceptual combination to explore (1) aspects of conceptual structure that enable flexible activation of conceptual features, with a focus on feature uncertainty, and (2) the neural regions that are involved in flexible feature modulation in complex language use.

One approach to testing theories of conceptual structure and combination is to embed theoretical assumptions in different computational models and see how well those models predict behavioral and neural responses to combined concepts. These methods have been fully adopted in models of vector-based distributional semantics in which the meaning of a word is represented in terms of its co-occurrences with other words in large text corpora (e.g., Mitchell & Lapata, 2008; 2010; Baroni & Zamparelli, 2010; Steyvers & Griffiths, 2007). For example, Mitchell & Lapata (2010) compared various additive and multiplicative models against non-combinatorial models and reported general success of a multiplicative model — which, compared to additive models, reflects an integration of constituent representations. A few fMRI studies have used a related approach to analyze neural responses to combined concepts (Baron et al., 2010; Baron & Osherson, 2011; Chang et al., 2009). For example, Baron & Osherson (2011) revealed that multivoxel patterns evoked by complex concepts (e.g., GIRL) in the left anterior temporal lobe (LATL) can be predicted by adding or multiplying the patterns evoked by the constituent concepts (e.g., YOUNG and FEMALE). Focusing on adjective-noun combinations, Chang et al. (2009) found a multiplicative model to be successful in predicting distributed neural responses to combined concepts (e.g., “soft bear”, “small cup”). Here we draw from these methods to directly explore how conceptual features are flexibly modulated in
combined concepts in behavioral and fMRI data. Comparing different combinatorial models will tell us something about how concepts are represented and also the processes by which they are combined.

We specifically hypothesized that the uncertainty of a conceptual feature (e.g., BRIGHTNESS) within a concept (e.g., DIAMOND) can explain the flexible activation of that feature in combined concepts (e.g., “dark diamond”). If a feature is present or absent in a concept with high certainty (e.g., DARK for CHARCOAL), then activation of that feature will be less flexible in different contexts (as in “light charcoal”). On the other hand, uncertainty in a conceptual feature (e.g., DARK for PAINT) allows this ambiguity to be resolved in conceptual combination and therefore substantial feature fluctuations may occur (as in “dark paint”). One way we examined the potential influence of feature uncertainty is using information theory’s notion of “entropy”, a measure derived from probability values which reflects uncertainty or the informativity of a signal (Shannon, 1948). We previously used this approach to explore neural responses to combined concepts (Chapter 3) and found a relationship between feature uncertainty and flexible, distributed neural activation of conceptual features during comprehension of combined concepts (e.g., “green pumpkin”, “sweet pickle”). We also embedded feature uncertainty in a new Bayesian model of feature composition and compared its predictions regarding degree of feature modulation to more traditional models. Probabilistic models of language composition have been explored in prior work (Goodman & Frank, 2016; Lassiter & Goodman, 2013); here we extend these ideas to the analysis of feature-based semantic composition.

We used adjective-noun combinations to modulate conceptual brightness in a range of nouns (e.g., “dark diamond”, “light charcoal”). Our goal was to test how well feature uncertainty (i.e., entropy) and a Bayesian model could predict (1) the degree of feature modulation caused by combined concepts, and (2) neural responses to combined concepts observed in fMRI data. We focused our fMRI analysis on a priori regions of interest including the anterior temporal lobe (LATL), left angular gyrus (LAG), left fusiform gyrus (LFUS), and left inferior frontal gyrus (LIFG).

4.2 Methods

4.2.1 Participants
In the behavioral part of the study, 357 participants completed online surveys on Amazon Mechanical Turk (AMT) and were compensated according to standard rates. Twenty-four additional participants from the University of Pennsylvania
community completed the fMRI study and were compensated $20/hour for their time. All fMRI participants were right-handed, fluent speakers of English with no self-reported neurological disorders or damage. Consent was obtained for all participants in accordance with the University of Pennsylvania IRB.  

4.2.2 Adjective and noun stimuli  
We focused on the single dimension of conceptual brightness to enable a tightly-controlled analysis of how brightness information is modulated across verbal contexts. We thus used the adjectives “dark” and “light” to modulate the conceptual brightness of 45 noun concepts. These 45 noun concepts covered the full range of brightness values (e.g., “diamond”, “snow”, “paint”, “shadow”, “charcoal”). A full list of noun stimuli is shown in the Appendix. 

4.2.3 Behavioral measures of conceptual brightness  
Information theory’s measure of entropy (Shannon, 1948) is a measure of uncertainty that can be calculated from probability values. Here we use entropy as a measure of feature “uncertainty” for a particular feature-noun pair. For example, we want to know the uncertainty of BRIGHTNESS in the concepts.

Figure 14: Explicit measures of conceptual brightness. (A) Participants used a visual scale to indicate the brightness of nouns and adjective-noun combinations. (B) For each of the 45 concepts, we calculated the extent to which conceptual brightness was modulated by “dark” and “light” adjectives. Distance from the central black line corresponds to increased modulation of conceptual brightness for both “dark” and “light” adjectives relative to the noun alone. The calculated adjective effect for each concept is the mean absolute distance between combination brightness and noun brightness across the two adjectives.
DIAMOND and PAINT. We collected subjective ratings of brightness probability on AMT in which participants (N=58) were asked “Is/are [noun] typically dark?” on a 5-point scale ranging from “This is always light” to “This is always dark.” Participants made responses to the 45 noun concepts presented in a random order. These ratings were averaged across participants and scaled between 0-1 to reflect brightness probability (p), in which values close to 0 indicate that a concept is likely light in color and values close to 1 indicate that a concept is likely dark. We then calculated brightness entropy (E) using the following equation:

\[ E = -p_{DARK}\log_2(p_{DARK}) + -p_{LIGHT}\log_2(p_{LIGHT}) \]

This was our measure of brightness uncertainty. Entropy is symmetrical around \( p=0.5 \), where \( p=0 \) and \( p=1 \) indicate maximum lightness and darkness respectively; each concept was thus assigned a single entropy value that captured the concept’s brightness uncertainty on the light-dark spectrum. We predict that brightness uncertainty will positively relate to the degree to which conceptual brightness can be modulated across the 45 noun concepts when paired with “dark” and “light” adjectives.

A separate set of participants on AMT provided explicit brightness judgments for the noun and adjective-noun items concepts. For each of the 45 noun concepts, participants (N=100) were asked to rate the darkness of each concept by sliding a bar corresponding to a visually-presented scale transitioning from white to black (e.g., “Rate the darkness of: PAINT”, Fig. 14A). The values of the visual scale ranged from 0 (white) to 50 (black) and were unseen to participants. A different set of participants (N=199) saw each noun modified by either “dark” or “light” and performed the same task (e.g., “Rate the darkness of: DARK PAINT”). Items were presented in a randomized order. Responses were averaged across participants, resulting in ground-truth brightness values for the 45 noun concepts (e.g., “diamond”) and for the combined concepts containing each of these noun concepts modified by “dark” and “light” adjectives (e.g., “dark diamond”, “light diamond”).

We then calculated the “dark” and “light” effects for each noun by calculating the difference in brightness between the unmodified and modified items (Fig. 14B). For example, we took the absolute difference between the brightness values of “diamond” and “dark diamond” to calculate the DIAMOND “dark”-effect, and the absolute difference between the brightness values of “diamond” and “light diamond” to calculate the DIAMOND “light”-effect. The “dark” and “light” effects for each noun were averaged together to result in a single measure that reflects the extent to which brightness information can be modulated within a concept across language contexts. We refer to this as our measure of ground-truth adjective effects.
4.2.4 Predictive models of adjective-noun combination

In order to understand how conceptual features are modulated in combined concepts, we created a set of predictive models that made different predictions about how concepts combine. Each model generates predictions reflecting the conceptual brightness of the adjective-noun combinations ($B_{COMBO}$) based on the conceptual brightness of the adjective ($B_{ADJ}$) and noun ($B_{NOUN}$), and thus also generates a prediction of brightness change ($B_{CHANGE}$), which corresponds to the ground-truth adjective effects described above. We constructed and tested an adjective model, a noun model, a weighted additive model, and a Bayesian model.

The adjective- and noun-models are non-combinatorial and are useful because their outputs can be considered baseline predictions that the combinatorial models should outperform. A similar approach is found in distributional or vector-based semantics (e.g., Mitchell & Lapata, 2008; 2010; Chang et al., 2009). In the adjective model, the predicted brightness of the combined concept (e.g., “dark diamond”) is identical to the brightness of the adjective (e.g., “dark”), which is formalized as:

$$B_{COMBO} = B_{ADJ}$$

Where $B_{ADJ}$ corresponds to the extreme values of the scale (dark=50; light=0). Within a single adjective (e.g., “dark”), the adjective model does not predict differences in $B_{COMBO}$ across the 45 nouns but it does predict differences in $B_{CHANGE}$ based on differences in $B_{NOUN}$ for “dark” and “light” separately. That is, it does make predictions about how observed feature activation will be influenced by each combination. However, when adjective effects are averaged across adjectives — as we did to define our ground-truth behavioral adjective effects — the adjective model does not predict differences in $B_{CHANGE}$. This is an important point: the non-combinatorial adjective model is not able to capture variability in the extent to which brightness can be modulated across our noun concepts.

In the non-combinatorial noun model, the predicted brightness of the combined concept is identical to the brightness of the unmodified noun:

$$B_{COMBO} = B_{NOUN}$$

The noun model predicts differences in $B_{COMBO}$ across items. It does not predict any variance in $B_{CHANGE}$ in the “dark” and “light” combinations separately, nor when averaged together. Thus, like the adjective model, the noun model is also unable to capture variability in the extent to which brightness can be modulated across the 45 noun concepts.
We constructed a combinatorial additive model in which the predicted brightness of the combined concept is a weighted sum of $B_{\text{ADJ}}$ and $B_{\text{NOUN}}$. This model has been proposed as a candidate combinatorial mechanism in both cognitive (e.g., Smith et al., 1988) and computational models of distributional semantics (e.g., Mitchell & Lapata, 2010). The simple form is:

$$B_{\text{COMBO}} = W \cdot B_{\text{ADJ}} + B_{\text{NOUN}}$$

In our case, $B_{\text{ADJ}}$ represents the extreme brightness values (dark=50, light=0). Our implementation of the additive model makes predictions for “dark” and “light” combinations that can be represented as:

$$B_{\text{COMBO-DARK}} = W \cdot 50 + B_{\text{NOUN}}$$

$$B_{\text{COMBO-LIGHT}} = B_{\text{NOUN}} - W \cdot 50$$

We optimized $W$ between $0 \leq W \leq 1$ (intervals of 0.01), separately for “dark” and “light” combinations. This resulted in a value of $W_{\text{DARK}}$ that minimized the mean squared error (MSE) of $B_{\text{COMBO-DARK}}$ predictions relative to the ground-truth dark-combo brightness values, and the value of $W_{\text{LIGHT}}$ that minimized the mean squared error (MSE) of $B_{\text{COMBO-LIGHT}}$ predictions (Fig. 15D).

We also constructed a combinatorial Bayesian model of adjective-noun combinations (Fig. 15A-C). The motivation behind this approach is that feature uncertainty may influence how concepts combine. In addition to using entropy as a measure of feature uncertainty, feature uncertainty can be embedded in probabilistic feature models, like the one described here. We represented conceptual brightness for adjective and noun concepts as probability distributions over possible brightness values. In a Bayesian model, the predicted brightness of a combined concept is the maximum a posteriori (MAP) estimate of the product of the constituent concepts’ brightness distributions. If the brightness probability distributions for adjectives ($P_{\text{ADJ}}$) and nouns ($P_{\text{NOUN}}$) are each captured by a gaussian defined by a mean ($\mu$) and standard deviation ($\sigma$), then the Bayesian predictions of the brightness of combined concepts is the maximum a posteriori (MAP) estimate of the product of these distributions:

$$B_{\text{COMBO}} = \arg \max f \{P_{\text{ADJ}} (\mu, \sigma) \cdot P_{\text{NOUN}} (\mu, \sigma)\}$$

We derived the 45 $P_{\text{NOUN}}$ distributions (e.g., for DIAMOND, PAINT, CHARCOAL) by fitting gaussian distributions to histograms reflecting the frequency of responses in the explicit brightness judgment task (Fig. 15A). We do not have data from which the $P_{\text{ADJ}}$ distributions can be calculated; for simplicity, we assumed that $P_{\text{DARK}\mu}=50$ and $P_{\text{LIGHT}\mu}=0$. We optimized separately for $P_{\text{DARK}\sigma}$ and $P_{\text{LIGHT}\sigma}$ ($0 \leq \sigma \leq 50$; intervals of 0.01); two example $P_{\text{DARK}}$ distributions with $P_{\text{DARK}\sigma} = 8$ and $P_{\text{DARK}\sigma} = 15$ are shown in Fig.2B-C. This procedure resulted in values for $P_{\text{DARK}\sigma}$
and $P_{\text{LIGHT}}$ that minimized the MSE of $B_{\text{COMBO}}$ predictions relative to the ground-truth $B_{\text{COMBO}}$ values (Fig. 15E).

The two combinatorial models (i.e., additive and Bayesian) each had one free parameter. For each model we generated predictions by optimizing the parameter (i.e., $W$ or $P_{\text{ADJ}}$) across the 45 concepts and then used that parameter value to make $B_{\text{COMBO}}$ and $B_{\text{CHANGE}}$ predictions for each concept. We

Figure 15: Predictive combinatorial models. (A) In our Bayesian model, conceptual brightness was represented as a probability distribution over brightness values for each noun concept. Greater values on the scale indicate increased conceptual darkness. Each distribution is defined by a mean and sigma, derived from our behavioral task (see Fig. 14). We defined the means of the “dark” and “light” distributions as the extreme ends of the brightness scale, and optimized for sigma. (B-C) Different “dark” $\sigma$ values result in different $B_{\text{COMBO}}$ predictions for “dark diamond”, and our goal was to find the sigma that generated the most accurate predictions of $B_{\text{COMBO}}$ across the 45 noun concepts. (D) In the additive model, we optimized the adjective-weight for “dark” ($W=0.35$) and “light” ($W=0.33$) separately. (E) In the Bayesian model, we optimized the standard deviation of the “dark” ($\sigma=8.42$) and “light” ($\sigma=10.27$) distributions separately. We averaged the “dark” and “light” parameters within each model to analyze fMRI data.
compared the accuracy of these models in their ability to predict the behavioral data.

In our analysis of fMRI data, we simplified the combinatorial models by averaging the “dark” and “light” parameter estimates within each model, resulting in a single value for $W$ and a single value for $P_{\text{adj}}$. One of the goals of the fMRI analysis was to determine which of any neural regions exhibit responses to combined concepts that reflect a combinatorial feature-based mechanism. The additive model and Bayesian model are both combinatorial, and success of either one in predicting neural response would satisfy our goal. We thus created a “composite combinatorial model” by averaging the $B_{\text{CHANGE}}$ predictions across the additive and Bayesian models. This resulted in $B_{\text{CHANGE}}$ predictions for the 45 dark-combos and the 45 light-combos.

4.2.5 fMRI task and design

Participants in the fMRI study completed 6 scanning runs. Participants viewed the 45 unmodified noun concepts in the first two scans, and the 90 adjective-noun combinations in the final four scans. Participants completed two tasks simultaneously: a conceptual color detection task (“color task”), and a fixation size-change detection task (“fixation task”).

In the unmodified noun scans (1-2), items were presented in an event-related design with 2s stimulus presentation and a fixation ISI of 2-8s. In the color task, participants were asked to press a button on a hand-held response box when an item referred to a cued color; the color cue (i.e., red or green) was presented before each block of trials. We thus interspersed filler items throughout each scan that were either typically red (e.g., “strawberry”, “ruby”), or typically green (e.g., “lettuce”, “frog”). This task was chosen to encourage visual imagery of the items without explicitly asking participants to think about conceptual brightness. Each run comprised one block of red-cued trials and one block of green-cued trials; the order of red/green blocks was pseudorandomized across runs. Each of the 45 target noun concepts was presented once per scan in a pseudorandomized order and was seen once in a red block and once in a green block across the experiment. To increase engagement with the stimuli, we included an additional fixation task in which participants were asked to press a different button on the response box when the fixation cross presented between stimulus items briefly changed in size, which happened at random intervals 8 times per scan (4 per block).

In the combined concept runs (3-6), participants completed the same color task and fixation task described above. The color task here involved fillers that were combined concepts for red (e.g., “dark blood”, “stop sign”) and green (e.g., “light moss”, “football field”). We included fillers that did not include “dark” and “light”
modifiers in order to encourage participants to process the full combined phrases rather than focus on the final word alone. Each of the 45 noun concepts appeared (modified by “dark” or “light”) once per scan, resulting in two presentations of each specific combination across the experiment. Each combination (e.g., “dark diamond”) was seen once in a red block and once in a green block.

4.2.6 fMRI acquisition and analysis

fMRI data were collected on a 3-T Siemens Trio System equipped with a 64-channel array head coil. Structural data included axial T1-weighted localizer images with 160 slices and 1mm isotropic voxels (TR = 1850 ms, TE = 3.91 ms, TI = 1100 ms, FOV = 240 mm, Flip Angle = 8°). Functional data included six acquisitions of echo-planar FMRI using a multiband sequence performed in 78 axial slices and 2 mm isotropic voxels (TR = 2000 ms, TE = 30 ms, FOV = 192 mm, Flip Angle = 75°).

Data were pre-processed and analyzed using FSL. Pre-processing included motion-correction using MCFLIRT, spatial smoothing with a Gaussian kernel of FWHM 5 mm, and high-pass temporal filtering. Motion outliers were modeled as covariates of no interest. All scans were analyzed with a GLM including item-level regressors modeling the individual TRs for each concept or combination contrasted against the fixation baseline. We added regressors for TRs in which filler items or instructions were presented, TRs in which the fixation cross changed size, and TRs in which participants made a response on the button box. These data were averaged across scans, resulting in whole-brain beta maps for the 45 unmodified noun concepts, 45 dark-combos, and 45 light-combos for each participant. We excluded time-points in which participants incorrectly responded to an experimental item from all subsequent analyses. Individual subjects’ data were transformed to and analyzed in MNI standard space.
4.2.7 Defining a priori ROIs

We analyzed neural responses to combined concepts within a priori regions of interest (ROIs; Fig. 16). We defined four ROIs in the left fusiform gyrus (LFUS), left angular gyrus (LAG), left anterior temporal lobe (LATL), and left inferior frontal gyrus (LIFG). Each of these spherical ROIs were comprised of 123 voxels and were centered on the peak voxel previously reported in a related study. The LFUS ROI was drawn around the peak voxel in an analysis reported by Hsu et al. (2011) in which increased activation in left fusiform was observed when the task required more detailed or specific color knowledge. This context-dependent activation of conceptual color information might relate to context-dependent feature modulation during combined concepts. The LAG ROI was drawn around the peak voxel in an analysis reported by Price et al. (2015) in which increased activation in left angular gyrus was observed for more plausible adjective-noun combinations (e.g., “plaid jacket” vs. “fast blueberry”). We created our LATL ROI based on Baron & Osherson’s (2011) finding that multivoxel patterns in left anterior temporal lobe could be predicted by multiplicative and additive combinations of the constituent concepts. Finally, we created our LIFG ROI

Figure 16: A priori neural regions of interest (ROIs). We analyzed responses in four ROIs; each 123-voxel spherical ROI was drawn around the peak voxel reported in a prior study. The left inferior frontal gyrus (LIFG) ROI (blue) was centered at Talairach coordinates x=-52, y=22, z=11 based on an analysis of metaphor processing (Cardillo et al., 2012). The left anterior temporal lobe (LATL) ROI (green) was centered at x=-39, y=10, z=-25 based on a multivoxel analysis of combined concepts (Baron & Osherson, 2011). The left angular gyrus (LAG) ROI (orange) was centered at x=-51, y=-54, z=23 based on an analysis of adjective-noun combinations (Price et al., 2015). The left fusiform gyrus (LFUS) ROI (yellow) was centered at x=-36, y=-47, z=-7 based on an analysis of task-dependent color processing (Hsu et al., 2012). (A) Left lateral view and (B) ventral view, projected onto a Talairach surface.
based on an analysis reported by Cardillo et al. (2012) in which activation of left inferior frontal gyrus was tuned by metaphor familiarity.

4.2.8 Neural adjective effects
We calculated univariate neural adjective effects (from now on “univariate effects”) that reflected the extent to which neural activity was modulated by “dark” and “light” adjectives across the 45 noun concepts. This measure was calculated similarly to the behavioral ground-truth adjective effects described above. The neural response evoked by each of the 45 unmodified noun concepts was averaged across scans and z-scored for each subject, such that the mean univariate response across all items was set to 0 across subjects. Similarly, the neural responses to the 90 combinations were averaged across scans and z-scored for each subject. Standardized univariate responses to the nouns and adjective-noun combinations were then averaged across subjects and analyzed as follows. For each concept, the neural “dark” effect was the absolute value of the difference between the dark-combination (e.g., “dark diamond”) and noun (e.g., “diamond”); the neural “light” effect was the absolute value of the difference between the light-combination (e.g., “light diamond”) and the noun (e.g., “diamond”). These values were averaged to result in the univariate effect for that noun concept. This approach resulted in univariate effects for the 45 noun concepts reflecting the extent to which levels of neural activity were influenced by the adjective-noun combinations.

We also captured the extent to which multivoxel patterns (MVPs) were modulated in adjective-noun combinations. We calculated the Spearman’s distance between the noun (e.g., “diamond”) and dark-combination (e.g., “dark diamond”), as well as the distance between the noun and the light-combination (e.g., “light diamond”), for each noun concept and separately for each subject. These distance values were averaged across subjects, resulting in mean pattern dissimilarity values for each noun compared to its dark- and light-combinations. The dark- and light- distances were summed for each noun concept, resulting in an overall MVP adjective effect for each of our 45 noun concepts. This “multivariate effect” measure captures the extent to which patterns of neural activity are influenced in combined concepts in each of the ROIs.

4.2.9 Analyzing neural effects
We tested whether our behavioral measures and combinatorial models predicted either the univariate or multivariate effects observed in our fMRI data. In addition to running Spearman’s correlations between these measures, we implemented additional bootstrap analyses. In each bootstrap analysis, we randomly sampled 45 concepts with replacement and calculated the spearman’s correlation between two measures of interest (e.g., feature uncertainty and univariate effects in LIFG). We repeated this analysis 10,000 times, resulting in a distribution of
correlation values between the two measures. In all cases we had an a priori prediction regarding the direction of the relationship: we were looking for positive relationships between our behavioral and model predictions and the neural adjective effects. We thus assessed significance in a one-tailed design ($\sigma = 0.05$).

To determine whether our compositional model predictions (i.e., the average of the additive and Bayesian predictions of $B_{\text{CHANGE}}$) related to our univariate and multivariate neural effects, we needed to ensure that the compositional predictions outperformed the non-compositional predictions. Remember that the adjective-model does not make mean $B_{\text{CHANGE}}$ predictions when adjectives are combined but does make $B_{\text{CHANGE}}$ predictions for “dark” and “light” separately. We thus concatenated across “dark”- and “light”-combinations to result in $B_{\text{CHANGE}}$ predictions for all 90 combinations for both the adjective-model and our compositional model; we the “dark” and “light” neural effects were similarly concatenated. To test the success of our combinatorial model, a general linear model was used to determine whether the combinatorial model $B_{\text{CHANGE}}$ predictions explained any unique variance in the neural adjective effects above and beyond that explained by the non-combinatorial adjective model. We used this regression analysis in the same bootstrap procedure described above, resulting in a distribution of 10,000 beta values indicating the extent to which the combinatorial model could predict either univariate or multivariate neural effects. As above, significance was assessed in a one-tailed design ($\sigma = 0.05$).

4.3 Results

4.3.1 Feature uncertainty predicts ground-truth adjective effects

We used entropy to derive a measure of feature uncertainty, which specifically captured the uncertainty of conceptual brightness within each of our 45 noun concepts. We predicted that brightness uncertainty would positively predict the extent to which a concept’s conceptual brightness could be modulated by related adjectives (i.e., “dark” and “light”). We did observe a strong positive relationship between brightness uncertainty and ground-truth adjective effects ($r=0.66$, $p<0.0001$), a result which supports the hypothesis that feature uncertainty is an aspect of conceptual structure that may influence processes of conceptual combination (Fig. 17A).

One reason to combine across “dark” and “light” when calculating adjective effects is to reduce the influence of edge effects. By collapsing across “dark” and “light” modifiers to calculate adjective effects, each concept has the same maximum potential movement within the brightness scale (i.e., 50 brightness
units). For each individual modifier (e.g., “dark”), some concepts are given more of an opportunity to change than others due to the brightness of the unmodified concepts: for example, when modified by “dark”, the darkness of CHARCOAL (\(B_{\text{noun}} = 43\)) can only increase by 7 units, whereas the darkness of SNOW (\(B_{\text{noun}} = 3\)) can increase by 47 units. Combining across “dark” and “light” thus eliminates this particular concern. However, to further confirm that our result was not driven by noun concepts on the extreme edges of the bounded brightness scale, we removed the 15 darkest and 15 lightest noun concepts and ran the same correlation with the remaining 15 concepts. We still observed a positive relationship between brightness uncertainty and ground-truth adjective effects (\(r = 0.78, p = 0.0006\)). In fact, the relationship between uncertainty and adjective effects still held when only the 9 middle-brightness noun concepts (GREY, CAR, ROCK, MARBLE, FUR, SLIPPERS, PAINT, SILVER, COCONUT) were analyzed (\(r = 0.74, p = 0.03\)). These additional analyses support our claim that the relationship between feature uncertainty and feature modulation is not merely due to edge effects but does reflect something meaningful about how concepts are combined.

4.3.2 Bayesian model best predicts ground-truth adjective effects
We tested the success of each of our models (i.e., adjective, noun, additive, Bayesian) at predicting the brightness values of combined concepts. For each model, we calculated the mean squared error (MSE) of dark- and light-
combination predictions for each item, averaged these MSE values across modifiers, and then across items to calculate the overall error of each model (Fig. 17B). As expected, the adjective-model (MSE=258.6) and noun-model (MSE=207.3) performed poorly relative to the combinatorial models. Comparing the two combinatorial models, the Bayesian model (MSE=16.8) outperformed the additive model (MSE=29.6), such that the Bayesian model generated predictions that were closer to the actual brightness ratings of the combined concepts ($t(44)=2.93$, $p=0.005$). These results further suggest that feature uncertainty contributes to feature modulation in conceptual combination.

When the parameters were averaged across modifiers within each model, the Bayesian model still outperformed the additive model ($t(44)=2.29$, $p=0.027$). We averaged the $B_{\text{CHANGE}}$ predictions of these simplified models — separately for “dark” and “light” combinations — in our subsequent analyses of fMRI data in order to test whether activity in the ROIs reflected a combinatorial process.

4.3.3 Neural responses to combined concepts

We measured univariate and multivariate neural responses to combined concepts in LFUS, LAG, LATL, and LIFG, and explored whether the neural responses in these regions related to ground-truth adjective effects, feature uncertainty, or our predictive combinatorial models.

4.3.3.1 Combinatorial effects in LIFG

The difference between mean LIFG response to a combined concept (e.g., “dark shadow”, “light shadow”) relative to the noun alone (e.g., “shadow”) was predicted by the explicit change in conceptual brightness caused by “dark” and “light” modifiers (Fig. 18A). That is, we observed a positive relationship between ground-truth adjective effects and univariate effects in LIFG ($r=0.33$, $p=0.03$). The robustness of this finding was confirmed in our bootstrap analysis, which similarly revealed a significant positive relationship between ground-truth effects and univariate LIFG effects ($p=0.006$).

Additionally, feature uncertainty positively predicted univariate LIFG effects ($r=0.38$, $p=0.01$), suggesting that a concept’s brightness uncertainty influences activity in LIFG when the brightness of that concept is directly modified (Fig. 18B). A bootstrap analysis confirmed this positive relationship ($p=0.001$).
Univariate effects in LIFG were not predicted by the combinatorial model in the main regression analysis \((p=0.20)\) nor in the bootstrap analysis \((p=0.09)\). Multivariate LIFG effects did not relate to ground-truth effects \((p's>0.2)\), feature uncertainty \((p's>0.3)\), or the combinatorial model \((p's>0.15)\).

We thus observed that univariate responses in LIFG during comprehension of combined concepts are predicted both by ground-truth effects and by feature uncertainty. These results suggest that recruitment of LIFG during conceptual combination in part relates to the flexible modulation of conceptual features.

**Figure 18: fMRI results.** Sensitivity to our combinatorial measures of interest were found in LIFG and LATL. (A) Ground-truth adjective effects predicted univariate responses to combined concepts in LIFG. (B) Feature uncertainty, defined using entropy, predicted univariate responses to combined concepts in LIFG. (C) Ground-truth adjective effects predicted multivariate responses to combined concepts in LATL; the greater difference in conceptual brightness when modified by “dark” and “light”, the greater the difference in multivoxel patterns in LATL between noun and adjective-noun patterns. (D) The Bayesian model explained unique variance in univariate LATL response when controlling for the non-combinatorial adjective model. This result also held for the initial analysis using a composite combinatorial model (averaged across additive and Bayesian models).
4.3.3.2 Combinatorial effects in LATL

Univariate effects in LATL were precited by our predictive combinatorial model when controlling for a non-combinatorial baseline. We ran a general linear model to test whether $B_{\text{CHANGE}}$ predictions of the combinatorial model explained unique variance above and beyond that explained by the $B_{\text{CHANGE}}$ predictions of the non-combinatorial adjective model. The result of this initial analysis did not reach statistical significance ($p=0.06$), however our bootstrap analysis revealed that the combinatorial model did explain unique variance over 10,000 iterations ($p=0.04$). These results suggest that univariate responses to combined concepts in LATL may reflect a combinatorial process in which features are flexible modulated. In a post-hoc analysis we tested whether the predictions of the additive model or Bayesian model could individually explain unique variance above and beyond the non-combinatorial baseline. The additive model did not explain a significant amount of unique variance in either the main regression analysis ($p=0.40$) nor in the bootstrap analysis ($p=0.17$). On the other hand, the predictions of the Bayesian model explained unique variance in both the main regression analysis ($p=0.044$) and the bootstrap analysis ($p=0.032$; Fig. 18D). We did not observe positive relationships between univariate LATL effects and either ground-truth adjective effects ($p's>0.2$) or brightness uncertainty ($p's>0.5$).

We calculated the extent to which multivoxel patterns of neural activity differed between processing the noun concepts (e.g., “marble”) and the corresponding combined concepts (e.g., “light marble”, “dark marble”). We found that ground-truth adjective effects positively predicted the extent to which multivoxel patterns were changed in LATL during comprehension of the combined concepts ($r=0.37, p=0.012$; Fig. 18C). This positive relationship was confirmed in the corresponding bootstrap analysis ($p=0.007$). These multivariate effects in LATL suggest that this region represents the output of a conceptual combination process. We did not observe a positive relationship between multivariate LATL effects and brightness uncertainty ($p's>0.5$), and multivariate effects were not predicted by the combinatorial model ($p's>0.2$).

4.3.3.3 Sensitivity to conceptual brightness in LFUS

Neither LFUS or LAG exhibited responses that related to our combinatorial measures (i.e., ground-truth effects, feature uncertainty, combinatorial model). However, we did find that univariate response to unmodified noun concepts (e.g., “charcoal”, “snow”) in LFUS positively correlated with explicit ratings of conceptual brightness ($r=0.31, p=0.036$). This finding is consistent with previous work implicating LFUS in the representation of visual features (e.g., Hsu et al., 2011). Univariate LFUS response to combined concepts did not relate to explicit brightness ratings of the 90 combinations ($p>0.2$). These results suggest that the flexible modulation of conceptual features in combined concepts might not necessarily rely on the
same neural mechanisms involved in representing the features of individual concepts in isolation.

4.4 Discussion

Here we explored how conceptual information is flexibly activated during comprehension of combined concepts. We simplified this challenge by focusing on the feature dimension of conceptual brightness, and we modulated conceptual brightness using adjective modifiers (i.e., “dark” and “light”). We collected explicit ratings of conceptual brightness for both unmodified nouns (e.g., “diamond”, “shadow”) and their combinations (e.g., “dark diamond”, “light diamond”), and used these data to characterize the extent to which conceptual brightness could be flexibly modulated in different noun concepts. We then set out to determine whether brightness uncertainty related to these ground-truth effects, and to determine which brain regions exhibit sensitivity to this feature-based combinatorial process.

We captured feature uncertainty using entropy, a measure from information theory (Shannon, 1948) that reflects the uncertainty of an outcome or the potential informativity of a signal. If \( P \) is the probability of an outcome, then entropy is highest when \( P=0.5 \). For example, consider flipping fair vs. biased coins. A flip of a fair coin has \( P_{\text{heads}}=0.5 \) and \( P_{\text{tails}}=0.5 \); the result of the coin flip is maximally uncertain. On the other hand, if a biased coin has \( P_{\text{heads}}=0.8 \) and \( P_{\text{tails}}=0.2 \) then the result of the flip will be less uncertain, as it is likely to result in heads. We translated these ideas to the realm of conceptual knowledge to explore the flexible activation of features in conceptual combination. If a noun concept is characterized by complementary values of \( P_{\text{dark}} \) and \( P_{\text{light}} \), then conceptual brightness will be most uncertain when both \( P_{\text{dark}} \) and \( P_{\text{light}} = 0.5 \). Consider the concepts DIAMOND and PAINT, which were characterized by \( P_{\text{dark}} \sim 0.2 \) and \( P_{\text{dark}} \sim 0.5 \), respectively. These values reflect the fact that diamonds are unlikely to be dark, whereas paint is equally likely to be dark or light in color. Because \( P_{\text{light}} = 1 - P_{\text{dark}} \), and because entropy is symmetrical around 0.5, each concept can be assigned a single brightness uncertainty value: DIAMOND has a lower brightness uncertainty (0.73) than PAINT, which has very high brightness uncertainty (0.99).

We predicted that brightness uncertainty would positively relate to the extent to which conceptual brightness would be modulated in combined concepts. For example, we predicted greater change in conceptual brightness for PAINT when modified by brightness adjectives (i.e., “dark paint”, “light paint”) than for DIAMOND when paired with the same adjectives. These predictions were supported by our data: we observed a strong positive relationship between brightness uncertainty
and ground-truth adjective effects across the 45 noun concepts. We also embedded feature uncertainty within a predictive Bayesian combinatorial model, in which a concept’s conceptual brightness is represented as a probability distribution defined by a mean and standard deviation (see Fig. 15A-C). We compared the success of this Bayesian model with a more traditional additive model (Smith et al., 1988; Mitchell & Lapata, 2008; 2010) in its ability to predict brightness modulations evoked by our combinations. The Bayesian model outperformed the additive model, highlighting the relevance of feature uncertainty in the conceptual combination process. Our behavioral and model results suggest that conceptual feature uncertainty influences how features are flexibly modulated in complex phrases.

The left inferior frontal gyrus (LIFG) was sensitive to our measures of flexible feature modulation in combined concepts. Univariate effects in LIFG were correlated with ground-truth adjective effects as well as brightness uncertainty. LIFG is not typically associated with conceptual combination, but it is known to play a role in metaphor processing (Solomon & Thompson-Schill, 2017; Cardillo et al, 2012; Bambini et al., 2011; Eviatar & Just, 2006; Lee & Dapretto, 2006; Rapp et al., 2004, 2007; Stringaris et al., 2007). Figurative language and conceptual combination rely on similar conceptual devices (Wisniewski, 1997; Estes & Glucksberg, 2000), in that they both involve the selection and integration of conceptual features. The metaphor “His teeth are pearls” and the noun-noun combination “pearl teeth” both involve selecting the relevant conceptual features from the PEARL concept (e.g., WHITE, SHINY) and mapping or integrating these features into the TEETH concept. The feature is often pre-selected in adjective-noun combinations (e.g., “white teeth”) but integration of feature information is still required. Our current results suggest that LIFG is involved in the feature integration process during complex language comprehension, and we argue that this process is relevant for combined concepts, figurative language, and natural language use in general.

The other neural region sensitive to our measures of flexible feature modulation was the left anterior temporal lobe (LATL). Univariate effects in LATL were predicted by a combinatorial model of conceptual combination, suggesting that activity in LATL reflects a feature-based combinatorial process. There is a substantial body of evidence implicating LATL in conceptual combination (Bemis & Pykkänen, 2011; 2012; 2013; Westerlund & Pykkänen, 2014; Boylan et al., 2015; 2017), and MEG evidence suggests that this region is involved in feature integration and the specification of object concepts (Westerlund & Pykkänen, 2014). We applied predictions of a well-defined combinatorial model to LATL responses and confirmed this region’s sensitivity to a feature modulation and/or integration process during conceptual combination.
Multivariate effects in LATL were predicted by ground-truth adjective effects. That is, increased modulation of conceptual brightness across concepts relates to increased distance between multivoxel patterns evoked by the noun and combined concepts. For example, the brightness of PAINT was substantially modified in combinations (i.e., “dark paint”), and the neural patterns evoked by these combinations were substantially different from those evoked by the noun alone (e.g., “paint”). This finding strengthens the theoretical claim that LATL’s role in conceptual combination relates to the flexible modulation or integration of features. Multivariate responses in LATL were previously explored in two related fMRI studies in which it was found that processing complex concepts evoked multivoxel patterns in LATL that could be predicted by combining the patterns elicited by the constituent concepts (Baron et al., 2010; Baron & Osherson, 2011). These studies revealed the contribution of ATL to conceptual combination but could not provide conclusive evidence of conceptual integration, per se. Here we explored fine-grained effects of combined concepts on multivoxel patterns in LATL and observed sensitivity of this region to the amount of conceptual change evoked by the combinations. These results converge on the hypothesis that LATL is involved in integrating feature representations of constituent concepts during conceptual combination.

The left angular gyrus (LAG) has emerged in multiple neuroimaging studies of conceptual combination (Bemis & Pylkkänen, 2012; Price et al., 2015; 2016; Boylan et al., 2015; 2017, Graves et al., 2010). In particular, it has been argued that LAG is sensitive to the plausibility of adjective-noun combinations (Price et al., 2015), and to “relational” combinations which imply an event or relation between two concepts, rather than feature attribution (Boylan et al., 2015; 2017). Here we specifically targeted the process of feature modulation in adjective-noun combinations and did not observe any LAG involvement. Though we hesitate to strongly interpret these null results, the apparent lack of LAG response to feature-based combination is consistent with the theory that LAG is recruited for “relational” rather than “attributive” conceptual combinations (Boylan et al., 2015; 2017).

We analyzed neural responses in LFUS based on prior work establishing sensitivity in this region to conceptual color (Martin et al., 1995; Simmons et al., 2007; Hsu et al., 2011; 2012). We did not observe sensitivity in this region to any of our feature-based combinatorial measures, though we did find a relationship between conceptual brightness and univariate LFUS response to unmodified nouns (e.g., “charcoal”, “snow”). Our results thus support a role of LFUS in representing visual conceptual features, but do not support a role of LFUS in combining conceptual features.
Previous studies comparing color representations evoked by perception or conception have revealed distinct yet overlapping representations in these tasks (Martin et al., 1994; Simmons et al., 2007; Hsu et al., 2012). It is possible that combined visual features are represented in distinct regions of LFUS that we did not test in the current study (see Supplemental Materials). However, our positive results in LATL and the lack of combinatorial results in LFUS suggests that combined conceptual features may be represented in regions that do not represent those features in the constituent concepts alone. It has been argued that the ATL is a semantic “hub” that integrates sensorimotor (e.g., visual) information represented elsewhere in the cortex, and that the representations in ATL reflect a conceptual similarity space transformed away from sensorimotor features (Patterson et al., 2007; Ralph et al., 2010, 2017). Related work has shown that ATL represents the conjunction of visual features (i.e., color and shape), but not the visual features themselves (Coutanche & Thompson-Schill, 2014). Our current results suggest that visual features modified by conceptual combination are similarly represented in ATL, distinct from the representations of features evoked by individual concepts. Additional work is needed to clarify how and where features are represented in combined concepts.

Here we have characterized the computational and neural mechanisms underlying the flexible modulation of conceptual information during language comprehension. Using methods inspired from information theory and Bayesian modeling, we provided evidence that feature uncertainty plays a role in conceptual combination. Further, our analyses exposed the LIFG and LATL as regions involved in this feature-based combinatorial process. These findings are likely to extend more generally to complex language processing and flexible concept use.
4.5 Appendix

<table>
<thead>
<tr>
<th>Noun Concepts</th>
</tr>
</thead>
<tbody>
<tr>
<td>white</td>
</tr>
<tr>
<td>paper</td>
</tr>
<tr>
<td>snow</td>
</tr>
<tr>
<td>sugar</td>
</tr>
<tr>
<td>diamond</td>
</tr>
<tr>
<td>teeth</td>
</tr>
<tr>
<td>pearls</td>
</tr>
<tr>
<td>rice</td>
</tr>
<tr>
<td>dove</td>
</tr>
<tr>
<td>bone</td>
</tr>
<tr>
<td>ivory</td>
</tr>
<tr>
<td>cloud</td>
</tr>
<tr>
<td>foam</td>
</tr>
<tr>
<td>shell</td>
</tr>
<tr>
<td>bread</td>
</tr>
</tbody>
</table>

**Table A1: Noun concepts.** The 45 noun concepts used in the current study are displayed in order of conceptual brightness (light to dark).
4.6 Supplemental Materials

Whole-brain Searchlight Analysis

In addition to analyzing the a priori regions of interest, we ran an exploratory whole-brain searchlight analysis (Kriegeskorte et al., 2006) to observe what other regions of the brain respond to flexible feature modulation in combined concepts. We used the Princeton MVPA toolbox to define 123-voxel searchlights across the whole brain; within each searchlight we calculated group-level univariate neural effects as reported in the main manuscript. We correlated the univariate effects with our measure of (1) ground-truth adjective effects, (2) feature uncertainty, and (3) the predictive Bayesian model (controlling for a non-combinatorial baseline).

We applied a voxel-wise threshold of \( \sigma = 0.001 \) to each of the resulting whole-brain searchlight maps, and clustered these thresholded maps using a minimum cluster size of 10 voxels (faces or edges touching). These maps were then overlaid to determine whether there was any overlap between the three measures of interest.

Using this conservative thresholding technique, we observed one region in which univariate responses to combined concepts was significantly correlated with both ground-truth adjective effects and feature uncertainty; this cluster had a peak voxel in the right lingual gyrus (Fig. S5). We did not observe any other significant overlap between the measures. However, we did observe one significant cluster emerging from the Bayesian model analysis and several significant clusters reflecting sensitivity to feature uncertainty (Table S2).

These exploratory results reveal that several brain regions exhibit responses to combined concepts correlated with feature uncertainty. These results also highlight the fact that many areas of the brain — besides the a priori ROIs we analyzed in the main study — are likely implicated in the flexible modulation of features during comprehension of complex language.
Fig S5: Overlap between ground-truth effects and feature uncertainty. In a searchlight analysis we found voxels distributed across the brain which exhibited a univariate response (see main Methods) that related to our measures of interest. Whole-brain maps were thresholded at $\sigma \leq 0.001$ and then clusterized with a minimum cluster size of 10 voxels. Significant overlap in sensitivity to ground-truth effects and feature uncertainty was found in right lingual gyrus (peak MNI coordinates: $x=20, y=-50, z=2$).

Table S2: Significant clusters in whole-brain searchlight analysis. Whole-brain searchlight maps were thresholded at $\sigma \leq 0.001$ and clusterized with a minimum cluster size of 10 voxels.
The goal of this thesis was to explore the relationship between conceptual structure and flexible concept use. These two questions — how conceptual information is represented, and how concepts are flexibly used in language and thought — are inextricably linked (Wisniewski & Gentner, 1991). Early debates in cognitive science carved out possible solutions in this theoretical space, focusing on the topic of conceptual combination (Smith et al., 1984; 1988; Hampton, 1988; Murphy, 1988). However, this challenge has not been revisited after subsequent decades of research in the cognitive neuroscience of conceptual knowledge and the accumulation of evidence relating to flexible concept use. Here I attempted to merge cognitive theories of conceptual knowledge with more recent methods in cognitive neuroscience, network science, and computational modeling. The fusion of these disciplines might help us understand how conceptual knowledge is represented and used.

One claim presented in early cognitive theories of conceptual knowledge was that conceptual representations are structured — that is, the features comprising conceptual knowledge are interdependent or associated with one other (Medin & Shoben, 1988; Wisniewski & Gentner, 1991; Sloman et al., 1998). However, no well-defined models of conceptual combination were proposed that included structured feature-based conceptual representations (c.f. Sloman et al., 1998).

The aim of Chapter 2 was to capture a feature-based, structured conceptual representation using a network science model, and to test whether these network models could capture empirical measures of conceptual flexibility. Individual concepts (e.g., banana, bottle) were modeled as graph-theoretical networks, in which network nodes represented conceptual properties (e.g., yellow, sweet) and network edges captured their associations. These networks captured the within-concept feature associations that reflect how properties relate to one another across instances of a concept. Formal measures that relate to different aspects of network structure were extracted in order to determine whether a concept’s network structure relates to how that concept is flexibly used.

Network measures were compared with a text-based measure of semantic diversity and also to empirical data from a figurative language task and an alternative uses task. The semantic diversity (SemD) measure was derived within the framework of distributional semantics described above (e.g., section 1.4) which assumes that “you shall know a word by the company it keeps” (Firth, 1957) — that is, by observing the other words with which it co-occurs. The SemD measure was constructed in Hoffman et al. (2013) by quantifying the diversity of a word’s co-occurring neighbors; this was interpreted to reflect meaningful variability in a word’s semantic content. One of the goals of Chapter 2 was to see
if aspects of concept network structure related to the SemD measure of flexibility derived within distributional semantics. Two aspects of network structure (i.e., core-periphery structure and network clustering) were correlated with SemD across the analyzed concepts. These results reveal that conceptual flexibility — previously only captured in methods of distributional semantics — can also be captured using the compositional, feature-based representations frequently adopted in cognitive theories of conceptual knowledge.

The relationship between concept network structure and flexible concept use was further explored using two additional empirical measures. First, a measure of “simile goodness” was derived for each concept, based on ratings of simile meaningfulness and familiarity. Second, a measure of creativity was derived for each concept based on an alternative uses task. Results revealed possible relationships between aspects of network structure and both of these empirical measures. These results are offered as a proof-of-concept that feature-based network representations of concepts can relate to how those concepts are flexibly used in language and thought.

The methods proposed in Chapter 2 are broad in scope and there are many iterations of this model that have not yet been explored. In particular, the concept network models capture feature relationships, but the kind of relationship can be modified based on methodological or theoretical preferences. For example, Sloman et al. (1999) describe asymmetrical dependences between features that could be captured in directed concept networks, instead of the undirected networks described here. Methodological variations of network construction could reveal further insights into the representation and use of conceptual knowledge. However, using a particular set of methodological decisions, I revealed that variations in concept representation and use can be formally understood in terms of the informational content and topology of concept networks. This approach also suggests a possible way to represent feature interdependencies and associations without appealing to “world knowledge” external to conceptual representations themselves (Murphy, 1988; Wisniewski & Gentner, 1991).

Conceptual structure can be analyzed on global and local scales. In Chapter 2, I analyzed the global structure of each concept by exploring the structural characteristics of a concept as a whole. In Chapters 3 and 4, I zoomed in to examine conceptual structure on a local scale — that is, I explored how the representation of individual features may relate to flexible concept use. These experiments specifically explored whether probabilistic measures of feature representation relate to flexible modulation of features in conceptual combination.

Examining the principles by which individual concepts combine into more complex phrases can illuminate not only how the brain combines concepts but also can reveal the key ingredients of conceptual structure. Concepts interact
with each other in language in interesting ways: the conceptual information
activated to represent a noun concept (e.g., *pumpkin*) can be flexibly modulated
depending on the verbal context. One way to approach this challenge is to
observe patterns of feature activation when concepts combine (e.g., *sweet
pumpkin*). In Chapter 3, I explored the feature-noun relationships that influence
feature activations in combined concepts, and focused on feature surprisal and
feature uncertainty, measures developed in information theory. In an fMRI study,
eight adjective concepts (e.g., *orange, sweet*), eight noun concepts (e.g., *pumpkin, cookie*) and the resulting 64 adjective-noun combinations were
analyzed. A searchlight analysis was used to select voxels distributed across the
brain that were sensitive to the kinds of information targeted by the adjectives. In
these voxels, a correlational multivoxel pattern analysis was used to calculate
property activation (e.g., *sweet*) in nouns (e.g., *pumpkin*) and combi-
nations (e.g., *sweet pumpkin*). These methods provided a way to characterize the degree to
which conceptual property information was influenced by the adjectives.

A measure of neural property modulation reflected the degree to which adjective-
noun combinations increased the activation of the adjective-denoted property.
Feature uncertainty, not feature surprisal, positively predicted the neural
modulation of conceptual features during comprehension of adjective-noun
combinations. These results suggest a more nuanced view of feature adjustment
than those provided in previous theories of combination, in which the uncertainty
of a feature is important — not (only) its strength or salience (e.g., Smith et al.,
1988). Using the network structures described in Chapter 2, it was possible to
examine local aspects of network structure — that is, how the adjective-denoted
feature nodes related to the noun concept network as a whole. The specific
comparison was between combinations in which the adjectives denoted features
in the network “core” versus “periphery”; this analysis revealed increased neural
property modulation for “periphery” properties. Taken together, these results
reveal how local feature representations influence the neural representation of
combined concepts.

The neural representations analyzed in Chapter 3 were distributed feature
representations across a range of property dimensions (i.e., material, color, taste,
texture). This view of conceptual representation is consistent with and adds to
prior work in cognitive neuroscience (Hubel & Wiesel, 1962; Haynes & Rees,
2005; Kanwisher et al., 1997; Kourtzi & Kanwisher, 2001; Martin et al., 1995;
Beauchamp et al., 1999; Konkle & Oliva, 2012; Lederman et al., 2001; Cavina-
Pratesi et al., 2010; Boronat et al., 2005; Buxbaum et al., 2006). This approach
made it possible to observe how neural representations of conceptual features
are influenced in conceptual combination, but this study was not designed to
reveal the specific brain regions that might be involved.
The goal of Chapter 4 was to further explore how the cognitive and neural structure of conceptual knowledge affects how concepts combine in language and thought. Following up on the findings from Chapter 3, the behavioral and fMRI experiments in Chapter 4 tested the role of feature uncertainty in the modulation of conceptual brightness evoked by adjective-noun combinations (e.g., dark diamond). Explicit ratings of conceptual brightness were collected for 45 noun concepts (e.g., diamond, fur, paint, charcoal) and their “dark-” and “light-” combinations, resulting in an explicit measure of conceptual brightness modulation for each noun concept. Feature uncertainty was captured in an entropy measure, as well as in a predictive Bayesian model of feature modulation.

In the behavioral data, feature uncertainty (i.e., entropy) was a strong predictor of the explicit measure of conceptual brightness change evoked by adjective-noun combinations. This was a clear behavioral replication of the fMRI results reported in Chapter 3: feature uncertainty relates to how conceptual information is modulated in combined concepts. To further analyze the behavioral data, I created a set of predictive models and compared their ability to predict behavioral responses to combined concepts. Following approaches in distributional semantics (e.g., Mitchell & Lapata, 2010), non-combinatorial adjective- and noun-models were compared to two combinatorial models that made different assumptions about how concepts combine. The first combinatorial model was a traditional weighted-additive model, in which the predicted brightness of combined concepts (e.g., dark diamond) was a combination of the brightness of the noun (e.g., diamond) and adjective brightness (i.e., dark and light), the latter weighted by a value determined through parameter optimization. The second combinatorial model was a novel Bayesian model of conceptual combination in which the brightness of each adjective and each noun was represented as a probability distribution over brightness values. These distributions captured the uncertainty of conceptual brightness in the variance of these distributions. The predicted brightness of each combination in the Bayesian model was the maximum a posteriori estimate of the product of the adjective and noun distributions. Unsurprisingly, both the weighted-additive and Bayesian model outperformed the non-combinatorial models. More importantly, the Bayesian model outperformed the weighted-additive model in its ability to predict explicit behavioral ratings of conceptual brightness for adjective-noun combinations. The success of the Bayesian model further reveals the role of feature uncertainty in the process of conceptual combination.

The fMRI study in Chapter 4 revealed the neural regions affected by feature uncertainty in the comprehension of adjective-noun combinations. The neural responses evoked by the concepts and combinations were observed in a priori regions of interest based on previous work related to conceptual combination.
Specifically, neural regions in left anterior temporal lobe (LATL), left angular gyrus (LAG), left fusiform gyrus (LFUS), and left inferior frontal gyrus (LIFG) were analyzed. Participants were exposed to the same nouns and adjective-noun combinations while fMRI data were collected, and the univariate and multivariate neural effects were calculated. Univariate neural effects were calculated analogously to the explicit behavioral effects, such that a large univariate effect indicated that mean activity in a neural region was substantially influenced by the adjective-noun combinations, relative to the noun alone. Multivariate neural effects were calculated by analyzing the dissimilarity of multivoxel patterns between the combinations and the noun alone.

Univariate effects in LIFG were predicted by both feature uncertainty and the explicit behavioral measure of brightness change. LIFG is not currently considered an important region for conceptual combination but is known to play an important role in figurative language comprehension (Solomon & Thompson-Schill, 2017; Cardillo et al., 2012; Bambini et al., 2011; Eviatar & Just, 2006; Lee & Dapretto, 2006; Rapp et al., 2004, 2007; Stringaris et al., 2007). The current results reveal this region’s sensitivity to feature uncertainty, and more generally support the role of LIFG in flexible concept use.

Univariate effects in LATL were predicted by a combinatorial model of feature integration (i.e., combined predictions of the additive and Bayesian models). A post hoc analysis revealed that LATL responses were specifically predicted by the Bayesian model. Additionally, multivariate effects in LATL correlated with the explicit behavioral measure of conceptual brightness change. Previous work has revealed the sensitivity of LATL to conceptual combination in general (Bemis & Pylkkänen, 2011; 2012; 2013; Westerlund & Pylkkänen, 2014; Boylan et al., 2015; 2017) and specifically suggests that LATL can capture representations of combined concepts (Baron et al., 2010; Baron & Osherson, 2011). The univariate and multivariate effects observed in LATL in the current study further support the hypothesis that this region represents combined concepts, and additionally suggest it is involved in the flexible modulation of conceptual features. These fMRI findings inform cognitive neuroscience theories of conceptual combination by highlighting the role of LIFG and LATL in the process of flexible feature modulation during comprehension of complex language. In sum, Chapter 4 provides a set of behavioral and neuroimaging results suggesting that feature uncertainty is a key ingredient of conceptual structure.

Questions remain about how conceptual features are represented in combined concepts. The experiment in Chapter 3 suggests that the comprehension of adjective-noun phrases results in modulation of conceptual features in distributed neural regions that represent those features in isolation. However, analysis of fMRI data in Chapter 4 revealed that a region that represented conceptual
brightness in unmodified nouns (i.e., LFUS) did not represent the conceptual brightness of the combinations. Further exploratory analyses suggest that other regions of ventral visual cortex might be involved in combining visual features in combinations (see section 4.6), but additional work exploring how features are represented in combined concepts is needed. One possibility is that combined concepts are represented in the same regions that represent the features in individual concepts; another possibility is that conceptual features are transformed in combinations and are represented in distinct neural regions (e.g., LATL). These questions reflect some of the core questions in the cognitive neuroscience of conceptual knowledge (e.g., Patterson et al., 2007; Martin, 2007; Ralph et al., 2010; Binder & Desai, 2011), further revealing the utility of conceptual combination in the study of conceptual knowledge more broadly.

We can further ask whether the kinds of structured feature-based representations captured in the concept networks in Chapter 2 can be observed in neural data. Chapters 3 and 4 analyzed the modulations of individual features in adjective-noun combinations, though it is likely that a whole suite of features were affected in these combinations. For example, the difference between *pumpkin* and *green pumpkin* is more than a change in color: the age or size of the pumpkin is also likely to be modified. Methodological and theoretical developments are needed such that we can understand how structured conceptual representations are represented and modified in the brain.

Another open question is whether feature uncertainty (which appears to be an important aspect of local structure) can be directly captured in feature-based concept networks. A certain degree of correspondence was suggested in Chapter 3, in which both feature uncertainty and core-periphery structure predicted neural responses to combined concepts. However, there is no clear one-to-one correspondence between feature uncertainty and any individual aspect of network structure. Though the integration of feature uncertainty and network models may not be necessary, linking these local and global theories of conceptual structure would provide a more cohesive view of the relationship between conceptual structure and flexible concept use.

An intriguing line of inquiry regards the extent to which concept network models can predict interpretations of combined concepts. As mentioned in Chapter 2, concept networks could be used to predict the features activated in combined concepts using spreading activation models, or similar methods (e.g., Abbott et al., 2015; De Deyne et al., 2016). For example, imagine a noun-noun combination with concept A in the modifier position and concept B in the head noun position. To generate feature predictions of combination AB, concept B would be represented as a full feature-based network, and concept A would be represented as a compressed vector of feature activations. This A feature-vector
would be used as input in a spreading activation model when applied to the $B$ concept network. The conceptual information contained in $A$ would spread through the $B$ concept network in a pattern determined by the structured feature associations embedded in the $B$ network. This method would likely result in asymmetrical interpretations of combined concepts (i.e., $AB$ would be interpreted differently than $BA$), consistent with prior work (Hampton, 1988; Ortony et al., 1985). One implication of this model is that all of the features activated by combination $AB$ will be present in the representation of concept $B$ — that is, there are no features unique to $A$ and no "emergent" features in the combination (Wisniewski & Gentner, 1991; Hampton, 1997). However, an important difference between concept network models and previous feature-based representations is the complexity of information they can contain. A network-based representation might obviate the need to explain emergent features.

For example, Springer & Murphy (1992) compared the extent to which the combination hard cake activated the properties inflexible, sweet, and stale. Whereas inflexible and sweet were considered properties contained in the adjective and noun concepts respectively, stale was considered to be an emerging property of the combination and not contained in either of the constituent concepts. The authors reported that combination properties (e.g., stale) were accessed more quickly than noun properties (e.g., sweet), and argue that a spreading activation model cannot explain these findings. Concept network models include strong and weak features for a concept and their associations across multiple sub-concepts: it is thus possible that a cake concept network would include stale as a feature, and that this feature would be activated by the hard cake combination, even if it is not typically true of cakes. In this case, an explanation of emergent features would not be necessary, and a spreading activation model might suffice.

In sum, the work contained in this thesis explored the global and local aspects of conceptual structure that relate to how concepts are flexibly used in language and thought. The global structure of feature-based concept network models relates to text-based and empirical measures of conceptual flexibility, and a local measure of feature uncertainty relates to how conceptual information is flexibly modulated in combined concepts in both behavioral and neural data. Integrating classical theories of conceptual knowledge with more recent computational approaches can help illuminate the relationships between conceptual structure and flexible concept use.


models of biomolecular interaction networks. *Genome research*, 13(11), 2498-2504.


van den Heuvel, M. P., Sporns, O., Collin, G., Scheewe, T., Mandl, R. C., Cahn, W., ... & Kahn, R. S. (2013). Abnormal rich club organization and functional brain dynamics in schizophrenia. *JAMA psychiatry, 70*(8), 783-792.


