HARVARD
Kenneth C. Griffin

GRADUATE SCHOOL
OF ARTS AND SCIENCES

THESIS ACCEPTANCE CERTIFICATE

The undersigned, appointed by the

Department of Linguistics

have examined a dissertation entitled

Artificial intelligence and fake reefs: what privative inferences and LLMs tell us about adjective-noun composition

presented by Hayley Ross

candidate for the degree of Doctor of Philosophy and hereby certify that it is worthy of acceptance.

Signature _____

Typed name: Prof. Kathryn Davidson, Chair

Signature _____

Typed name: Prof. Gennaro Chierchia

Signature _____

Typed name: Prof. Najoung Kim, Boston University

Date: April 16, 2025

Artificial intelligence and *fake reefs*: what privative inferences and LLMs tell us about adjective-noun composition

A dissertation presented

by

Hayley Ross

to

The Department of Linguistics

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Linguistics

Harvard University

Cambridge, Massachusetts

April 2025

Dissertation Advisor: Professor Kathryn Davidson Hayley Ross

***Artificial intelligence* and *fake reefs*: what privative inferences and LLMs tell us about adjective-noun composition**

Abstract

The fact that people understand completely novel phrases is often taken as an argument that linguistic meaning is composed from the meaning of its parts. Thus, a central concern for the study of meaning is *how* that meaning is composed, especially for open-class content words like adjectives and nouns. This dissertation studies meaning composition and its interaction with context through the lens of adjective-noun modification and the privative inferences that sometimes result (e.g., a *fake gun* is (usually) not a gun, and a *stone lion* is not a (living) lion). This dissertation shows that privativity is not limited to a particular class of adjectives, which leads to a new, non-intersective semantics for adjective-noun composition which handles potential contradictions as part of composition. Further, we find that humans and modern large language models (LLMs) can generalize to the inferences of adjective-noun combinations that they have not seen before. Working with LLMs foregrounds the possibility that these inferences could be drawn by other means than meaning composition, such as memorization or analogy. In fact, success on this task is not explained by analogical generalization, as a computational analogy model and a human experiment involving analogy do not yield the expected inferences for all of the dataset. More broadly, the necessary adaptation in experiment design as well as reflection on our standards of evidence feeds into the broader, currently emerging discussion about how to study compositionality in humans and language models alike.

Contents

Acknowledgments

It is difficult to do justice in writing to the appreciation I have for all the people who helped me on this path, and so I will keep these acknowledgments brief. Instead, I will do my best to express my gratitude in person.

This dissertation would not have been possible without the guidance of each of my committee members: Kathryn Davidson, Najoung Kim and Gennaro Chierchia. Without Kate's encouragement, willingness to explore beyond her current research areas into LLMs, and insightful advice helping me see the forest for the trees in advising meeting after advising meeting, I do not know how I would have brought this dissertation together. Najoung has helped me find a home from home at BU's tinlab and I am indebted to her for many deep discussions about everything from methods to big picture questions about both humans and LLMs. Both Najoung and Kate have encouraged me not to shy away from the big questions that this work touches on, for which I am particularly grateful. Finally, I must thank Gennaro for sticking with me even after I veered from theoretical semantics (back) into more computational pursuits, and for substantially shaping the theory chapter at relatively short notice.

I would like to thank my co(rona)hort-mates Dasha Bikina, Yvette Wu and Tess Monks for their friendship and support—from our first bonding over phonology problem sets over Zoom in 2020 to now—as well as everyone in the Meaning & Modality lab at Harvard and in tinlab at BU. Thank you for each listening to presentations about areas sometimes far from your own and giving me so much helpful feedback and discussion nonetheless. Special thanks are due to Natasha Thalluri for listening to all my complaints during this long process, and for encouraging me to follow my heart and not be afraid of proposing a theory substantially different from previous work. A second special mention goes to my research assistant Kate Bigley, who is the author of many of the contexts used in this dissertation and who was instrumental in running Qualtrics surveys over hundreds and hundreds of items.

Finally, no acknowledgments section would be complete without thanking my friends and family on both sides of the Atlantic. You know who you are.

Chapter 1

Introduction

1.1 Compositionality and privative adjectives

The fact that people understand completely novel phrases is often taken to provide an argument that linguistic meaning must be governed by some kind of principle of compositionality: the meaning of a phrase arises from the meaning of its parts (Partee, 2009; Szabó, 2012).[1] This is often formulated as C_0 (but see Section 1.3):

(C_0) The meaning of a complex expression is a function of the meanings of its constituents and the way they are combined. (Szabó, 2012)

If compositionality is indeed key to understanding new phrases and a key part of linguistic competence, then a central concern for formal semantics is *how* the meanings of complex expressions are composed from the meanings of their constituent parts, in particular if the wider linguistic context, not just the lexical items and their syntax, is considered one of those constituent parts.

[1]Not all linguists take compositionality to be necessary for language acquisition or generalization. Bybee (2010) and Ambridge (2020), i.a., propose that we interpret novel phrases by analogy to similar "constructions", using an appropriate degree of abstraction. We will discuss analogy in Chapter 5.

Adjective-noun modification provides an excellent window into this question of *how* meaning is composed in the case of open-class content words. Composing two content words yields a rich set of effects typically dependent on both words and the concepts behind them, as well as the context in which those words are interpreted. When composing a function word such as a quantifier with a content word, by comparison, the resulting meaning is often predictable from the semantics of the function word alone. Adjective-noun modification thus bridges semantic composition and conceptual composition (Murphy, 2002; Pelletier, 2012), while also being much more predictable and constrained than noun-noun composition, another case of conceptual composition (ten Hacken, 2016). Despite this, adjective-noun modification is not straightforwardly compositional in meaning: for example, *corporate* seems to mean something different in *corporate stationery* vs. *corporate account* or even *corporate donor* (Murphy, 2002). Moreover, adjective-noun composition results in a membership ("subsective" or "intersective") inference or non-membership ("privative") inference relative to the original noun which illuminates what is happening during composition. Privative inferences yield a second challenge to compositionality, as we will see in Section 1.2: if the resulting composed meaning is not a member of the noun concept we started with, then what meaning did the noun contribute? This dissertation follows in a growing tradition of studying the dynamic interaction of meaning and context during composition through the lens of (privative) adjective modification (del Pinal, 2015, 2018; Guerrini, 2024a; Martin, 2022; Partee, 2009, 2010; Reinert, 2024, i.a.).

Historically, privativity has been defined as an adjective-specific phenomenon where an adjective such as *fake* negates the noun that it combines with. A *fake gun* is said to be precisely *not a gun* (Partee, 2010). Canonical examples of privative adjectives include *fake, false, former, counterfeit, knock-off, mock* and sometimes also *artificial* and *virtual* (Nayak et al., 2014). Noun membership inferences have been used to divide adjectives into classes of intersective, subsective, nonsubsective and privative adjectives, exemplified by the adjectives below.

(1) *Intersective inference*

This is a yellow flower.

∴ This is yellow.

∴ This is a flower.

(2) *Subsective inference*

This is a small elephant.

∴ This is an elephant.

∴ This is small.

(3) *Privative inference*

This is a fake gun.

∴ This is not a gun. (*according to Partee, 2010; judgments vary somewhat, see Chapter 2*)

(4) *Nonsubsective inference*

This is an alleged murderer.

∴ This is a murderer.

∴ This is not a murderer.

Privativity in adjective-modification is interesting to study for the following three reasons: (1) The inference pattern (privative or subsective) associated with these instances of adjective-noun modification can be gathered in offline experiments, providing an empirical window into compositional processes which are otherwise difficult to measure. (How do you measure whether someone *understands* a novel phrase?) (2) Privative inferences pose a challenge for some accounts of compositionality, especially ones which use set intersection. It turns out (Chapter 2; previewed in Section 1.4.1) that privative inferences are not restricted to a small class of particular adjectives, and that all the adjectives in this "class" may also yield subsective inferences, suggesting that distinguishing such a class is not fruitful. Accounting for privative inferences as a natural part of adjective-noun modification, rather than some awkward exception, leads us to a deeper, new understanding of how adjective-noun modification works. (3) As we will see in Chapter 2, privativity inferences are sensitive to discourse

context, as well as which noun the adjective is used with. Moreover, *fake* appears to be able to target different properties of the noun in different contexts. This lets us sharpen the ways in which context is involved in composition as we assess whether this contextual variation can be handled by having the lexical entries be context-sensitive, or whether the composition function itself should be context-sensitive.

1.2 A challenge for theoretical accounts of compositionality

Privativity poses a challenge for some accounts of compositionality, especially ones which use set intersection. Compositionality requires the meaning of a complex expression to be derived solely from the meaning of its constituent parts. If the meanings of words are just sets, then in the case of *fake gun*, that means the set of *guns* combined with whatever function or set *fake* denotes (plus possibly the linguistic context). Modification of nouns by adjectives is classically treated as simple set intersection, shown in (5), as in textbooks like Heim and Kratzer (1998) and Coppock and Champollion (2023).[2]

(5) $[\![\text{yellow flower}]\!] = [\![\text{yellow}]\!] \cap [\![\text{flower}]\!] = \{x : x \text{ is yellow}\} \cap \{x : x \text{ is a flower}\}$

If a *fake gun* is not a *gun*, it is not clear how to derive the meaning of *fake gun* from the set of (real) *guns* through set operations such as subsection. This means that our compositional process cannot arrive at the meaning of *fake gun* simply by having *fake* yield some subset of the meaning of *gun*. This problem is not unique to *fake*, or even to other adjectives called privative like *counterfeit*. It extends to material adjectives like *stone* in *stone lion* (Partee, 2010), and even to color and shape adjectives (see Chapter 2):

(6) This is a stone lion.

∴ This is not a lion.

[2]To their credit, both textbooks note that gradable and/or non-intersective adjectives are not handled by this.

(7) This is a red / hexagonal €50 note.

∴ This is not a €50 note.

(*€50 notes are orange and rectangular (140 × 77 mm) by definition, so this must be a forgery.*)

There are two possible responses to this problem. One, due to Partee (Kamp and Partee, 1995; Partee, 2007, 2009, 2010) is to say that adjectives such as *stone* and *fake* really are subsective, and that in fact the noun meaning gets adjusted so that there is a subset of it that contains the *stone lions* and *fake guns*, even though those are not ordinarily part of the meaning of *lion* and *gun*. The other is to say that adjectives like *fake* do not actually involve a subset of the noun as part of their meaning. Instead, they pick out the relevant properties of the noun that the meaning of the adjective-noun phrase retains by other means. A notable account of *fake* of this form is del Pinal (2015, 2018). We will give a brief overview of these two accounts here; Chapter 3 holds a full discussion of the theory of adjective-noun modification, including other existing accounts of privative adjectives.

Widening accounts Partee proposes in a sequence of papers (Kamp and Partee, 1995; Partee, 2007, 2009, 2010) that *fake* is subsective and that *fake gun* initially yields an empty set, because *fake guns* are precisely not guns. The Non-Vacuity Principle then rescues the derivation by widening the meaning of *gun* to include some non-guns, namely *fake guns*.

(8) **Non-Vacuity Principle**

In any given context, try to interpret any predicate so that both its positive and negative extension are non-empty (Kamp and Partee, 1995, p. 161)

The same happens for *lion* in *stone lion*, which is also expanded to include some non-lions, namely stone statues of lions. Note that Partee's account can be viewed as a contextual modulation account, whereby the linguistic context modulates (widens) the meaning of the noun in order to make it compatible with the adjective (del Pinal, 2018; Recanati, 2010). This idea of widening is carried over in several subsequent accounts, such as Guerrini (2024a) and Reinert (2024).

Dual content accounts Del Pinal (2015, 2018) argues that we can explain the behavior of adjectives like *real, typical* and *fake* without recourse to contextual modulation by moving to a two-dimensional semantics. Del Pinal's Dual Content Semantics enriches lexical entries to have an extensional component (E-structure), which corresponds to the traditional set extension, plus a conceptual component (C-structure). This C-structure is a list or structure of 'dimensions' that essentially captures the concept behind the noun or adjective, leaning on the large body of literature on concepts in psychology to do so. Del Pinal illustrates it using four qualia (Pustejovsky, 1991) as its dimensions: CONSTITUTIVE (what is it made of?), PERCEPTUAL (what does it look/feel like?), TELIC (what is its purpose, if any?) and AGENTIVE (what is its origin?). (See Section 3.2 for a more detailed introduction to C-structure.)

(9) $[\![\text{gun}]\!]^c_M$ =

E-structure: $\lambda x.\, \textsc{gun}(x)$

C-structure:

C: $\lambda x.\, \textsc{parts-gun}(x)$

P: $\lambda x.\, \textsc{perceptual-gun}(x)$

T: $\lambda x.\, \textsc{gen}\, e\, [\textsc{shooting}(e) \wedge \textsc{instrument}(e, x)]$

A: $\lambda x.\, \exists e_1\, [\textsc{making}_w(e_1) \wedge \textsc{goal}(e_1, \textsc{gen}\, e[\textsc{shooting}(e) \wedge \textsc{instrument}(e, x)])]$

Under this account, adjectives like *fake* draw on the C-structure of nouns like *gun* to build the new E-structure for *fake gun*, as shown in (10). Firstly, *fake guns* are not in the extension of guns. Secondly, *fake guns* do not have the origins of guns (the agentive property, accessed by Q_A); instead, they were made to appear as if they were guns (the perceptual property, accessed by Q_P). The C-structure of the composed phrase is built in a similar way (the perceptual properties are preserved, while the origin is replaced with "made to appear as if they were guns", etc.), allowing recursive application of adjectives that draw on the C-structure.

(10) $[\![\text{fake gun}]\!] =$

E-structure: $\lambda x.\ \neg E([\![\text{gun}]\!])(x) \wedge \neg Q_A([\![\text{gun}]\!])(x) \wedge$
$\exists e_2\ [\text{making}(e_2) \wedge \textsc{goal}(e_2, Q_P([\![\text{gun}]\!])(x))]$

C-structure: ...

This account of *fake* in fact encodes that *fake* is necessarily privative, since the E-structure of *fake gun* involves $\neg E([\![\text{gun}]\!])(x)$, meaning it is not in the extension of *gun*. Del Pinal (2018) notes that whether this negation is present is an empirical question. What is important is that this nonsubsective lexical entry for *fake* uses the C-structure of *gun* to describe the meaning of *fake gun* in terms of certain *gun*-properties, but not others. Selecting the relevant properties in this way allows us to avoid taking a subset of *gun* in order to inherit them, and thus sidesteps the intersection problem described above while still being a clearly compositional account.

1.3 What do we mean by compositionality?

The Prinviple of Compositionality is usually phrased as something like C_0, as given at the beginning of this chapter (Partee, 1984; Szabó, 2012):

(C_0) The meaning of a complex expression is a function of the meanings of its constituents and the way they are combined. (Szabó, 2012)

In fact, this phrasing leaves a lot of room for ambiguity. Much rests on the meaning of *function* (Pelletier, 2012; Szabó, 2012), as well as the meaning of *meanings of its constituents*, whether the pronoun *they* refers to how the constituents are combined (syntactically) or how their *meanings* are combined (Szabó, 2012), and how we even identify constituents (Nefdt, 2020). The last is usually the job of syntax, but this is not always trivial, and not all syntactic constituents need be "meaningful parts" for the purposes of semantic composition. Moreover, there is usually assumed to be a tacit "that's all" clause (Szabó, 2012): the meaning of a complex expression is not a function of anything else, such as the context or other associated

meanings or knowledge that we might retrieve from memory.[3] It is only a function of the meanings of the (syntactic) constituents. (The meaning of these constituents is not restricted, and may still depend on context.)

If all we mean by *function* is a mathematical function, then even if this function takes only the constituents as its arguments, it could well introduce new meaning (say, in the form of constants) or dramatically transform the meanings of the constituents (by some deterministic but complex process). Further, a mathematical function need not retain all or even any of the meaning of the constituents (Pelletier, 2012). Intuitively, this is not what linguists mean by composition. Firstly, the function that combines meanings should be *faithful* in the Optimality Theory (OT) sense (Hendriks and de Hoop, 2001; Prince and Smolensky, 2004): it should preserve all the features of the input (OT MAX) and it should not add any additional features (OT DEP). In particular, in our case, it should preserve all or as much as possible of the adjective and noun meaning, and it should not add any additional meaning, not even constants. (This does not preclude that the adjective meaning and noun meaning may be rich and context-sensitive.) While faithfulness is standardly applied as a constraint in OT treatments of adjective-noun semantics (Hogeweg, 2012; Kluck, 2007), it is in fact a concern for any account of linguistic composition and should not be taken for granted (Kim, 2021). For example, Kim (2021) finds that the neural networks she studies appear to rank faithfulness below other biases they have when attempting a compositional task.

Secondly—and this is harder to formalize—there is an intuition that the function combining the meanings should be somehow simple: the function should not overly transform the meaning of the parts and instead there should be a transparent relationship between the meaning of the combined phrase and the meaning of the parts. This can partly captured by a MAX-style faithfulness constraint if we are able to break down the meanings of words into features that should also be present in the combined meaning, but is difficult to formalize if the adjective and noun meaning are taken to be atomic.

[3]This restriction is what rules out analogy, which we discuss at length in Chapter 5, as a type of composition, since analogy makes reference to the meaning of words which are not constituents of the sentence.

One possible reading of C_0 that attempts to restrict C_0 to match (some of) these intuitions is the following, strengthened variant C, as phrased by Szabó, which asserts that once we know the meaning of the individual lexical items and the syntactic structure of the sentence, the meaning of the phrase is completely "determined".

(*C*) The meaning of a complex expression is determined by the meanings its constituents have individually and the way those constituents are combined. (Szabó, 2012)

C uses *is determined by* instead of *is a function of*, which captures the 'that's all' intuition and a notion of causality between the meaning of the the constituents and the meaning of the whole by virtue of the natural language semantics of *determined*. Further, *C* does not allow context to affect the meaning of the whole, either by affecting the meaning of the parts or by affecting the "function" (as phrased in C_0) that combines the meanings of the constituents. This may be too strong for our purposes, since as we will see in Chapter 3, we will want context to be able to affect the meaning of nouns. That said, the word *determined* alone does not seem to fully capture the faithfulness and lack-of-transformation intuitions noted by Pelletier (2012). I will not attempt to phrase my own single-sentence definition of compositionality here, but instead return to these notions on composition and faithfulness in Chapter 3 when we discuss our compositional theory, and again in the discussion of analogy in Chapter 5.

Semantics theories of meaning composition typically implement meaning composition as function composition / application: constituents are composed by virtue of one constituent having a higher, function type and taking the lower-typed constituent as an argument. (An exception to this is predicate modification for adjective-noun composition; as discussed above, predicate intersection does not explain all adjective modification and we will not be using it in this dissertation.) If composition is function application, then all the constraints discussed for C_0 above, such as faithfulness, apply directly to the higher-typed constituent meaning, as function, instead of to some external function that composes the two meanings. All word meanings in the lexicon that are functions of other meanings (i.e. are not type $\langle e, t\rangle$ or $\langle v, t\rangle$) must be faithful to their arguments as well as to whatever notion of "meaning" they implement (in our case, anything that is considered to be the "adjective meaning" that is distinct

from how it combines with the noun). This is hard to make precise if adjectives are opaque functions of nouns, but we will see in Chapter 3 how we can decompose adjective meaning into a "core" meaning component and a separate component that combines this core meaning with the noun, allowing us to speak of faithfulness to both the adjective "core" meaning and the noun meaning.

Szabó (2012) notes that linguists also assume a locality constraint for composition, namely that composition can be done locally at each level of the tree and does not require access to lower levels (previously composed items). For semantic theory, this follows automatically since it uses function application along the syntactic tree; moreover since our tree of adjective-noun composition has depth 1, we will not need to discuss further whether locality applies in other cases.

A final question is what this principle of compositionality applies to. Groenendijk and Stokhof (2005) argue that it is merely intended as a constraint on linguistic theory, but some definition of compositionality is now increasingly applied to other systems such as human language processing and LLMs (Baggio, 2021; McCurdy et al., 2024; Nefdt, 2020 i.a.). Nefdt (2020) distinguishes three levels of compositionality that may be applied to a system: process, state and output (see also Baggio (2021) for a processing-level principle of compositionality). Most expectations of compositionality in language target either the theory or the process level. By studying composition in LLMs, as we will do in Chapter 4, we assume not just that language *can* be accounted for compositionally (in some theory), but that we can look for a process-level implementation of that composition, or at least behavioral signatures of such an implementation. This requires a definition of compositionality that can apply equally to linguistic theory, human processing and LLM processing. Not all the considerations discussed above apply. For example, composition can, but need not be function application. Considerations that can be retained are that the process that combines these parts should be faithful and should not make reference to information not provided by those parts (though the latter is rarely discussed for LLMs, and hard to evaluate). As Nefdt (2020) notes, we need a definition of what the meanings of the parts / constituents are, and potentially some evidence that

these parts are represented in the theory / human / LLM, or behavior that we believe implies this (McCurdy et al., 2024). As we will discuss at the end of Chapter 4, the field is still split on what kind of evidence (behavioral, process-level or otherwise) is sufficient to show that a system is compositional, and by extension, exactly what we mean by compositionality.

1.4 The two strands of this dissertation

We are now in a position to map out the contributions of this dissertation. There are are two strands to this dissertation. The first strand, laid out in Section 1.4.1, investigates what we can learn about (content word) composition from privative inferences, starting with collecting data in Chapter 2 and then developing a theory in Chapter 3. The second strand, laid out in Section 1.4.2, looks into generalization to novel adjective-noun combinations: compositionality is typically argued for on the basis of such generalization, but we need to gather evidence that humans can indeed generalize in this rich landscape (Chapter 2). It is also not clear that composition of symbolic representations is the only way to draw (privative) inferences. Chapter 4 studies whether large language models (LLMs) can draw these inferences and what that tells us about what we should consider evidence for compositionality. Chapter 5 explores whether analogical reasoning is sufficient to draw these inferences (both for humans and LLMs). We find that analogy does not yield the expected inferences in all cases, suggesting that the compositional mechanism proposed in Chapter 3 is indeed justified.

In this section, I will give an overview of the dissertation by main contributions in each of these two strands, which sometimes cross-cuts the division into chapters. Section 1.5 then gives a short overview of the five chapters.

1.4.1 Broad insights from a (seemingly) small phenomenon

Chapter 2 begins by collecting data on a wide range of adjective-noun combinations. Chapter 3 then develops a new theoretical account of adjective-noun modification.

Privativity is not tied to specific adjectives In contrast to the conventional framing of privativity as an adjective-specific phenomenon, Martin (2022) shows in an experiment that inference patterns for so-called privative adjectives vary depending on the noun used, and need not be privative, refining previous corpus-based observations by Pavlick and Callison-Burch (2016b). For example, *counterfeit* may license a privative or subsective inference depending on the noun (and accompanying context).

(11) *Subsective inference*

This is a counterfeit watch.

∴ This is a watch.

(12) *Privative inference*

This is a counterfeit dollar.

∴ This is not a dollar.

This causes issues for our two previous theories, which either do not explain why both inferences are possible (Partee's theory) or force a privative inference (del Pinal's theory). Chapter 2 collects inferences for nearly 800 bigrams, and finds that (1) so-called privative adjectives can yield either subsective or privative inferences for many nouns (substantially extending Martin's dataset), (2) adjectives like *homemade* and *stone*, which are not traditionally classed as privative, can yield privative inferences, (3) the ability to be privative extends even to adjectives like *tiny* and *red*, traditionally classed as intersective.

Insights into adjective-noun modification If nearly any adjective can be privative with the right noun, and no adjective is necessarily privative, this suggests that we need to substantially revise the pathways by which privativity arises and how we should handle adjective-noun modification compositionally. I argue that privativity can be more productively viewed as a case of contradiction arising between the meaning contributed by the adjective and the meaning of the noun. In principle, there can be a noun that stands in contradiction with any adjective's meaning, explaining why any adjective can be privative. Moreover, this allows us to decouple adjective meaning and privativity and instead handle privativity as part of

modification: it is the responsibility of the composition mechanism, or of the modifier in an asymmetric modifier-modified relationship, to compose gracefully and avoid engendering a contradiction that prevents composition. Sometimes, we need to avoid a contradiction by adjusting the meaning of the noun.

Chapter 3 lays out a new theoretical account of adjective-noun modification which embodies these two choices. By using dual content semantics (del Pinal, 2018), we can take advantage of the modifier-modified asymmetry of adjective-noun modification to incorporate this responsibility directly into the adjective, avoiding the need for context-dependence and getting more directly at what it means for an adjective to modify a noun. In this account, the adjective contributes its own meaning and keeps only those properties of the noun which are compatible, keeping as many as it can but discarding any that would cause a contradiction.

Context-sensitivity Further, Chapter 2 finds that inferences for so-called privative adjectives are sensitive to context: the same adjective-noun bigram may yield a privative or subsective inference depending on the discourse context provided. This raises the question of whether the adjectives themselves are context-sensitive, or whether the method of composition is context-sensitive, or whether the context simply influences the inference directly without affecting composition. One account of *fake* (Reinert, 2024) posits that *fake* is as context-sensitive as *skillful*, which takes a contextually specified comparison property that need not have any relation to the noun. In fact, carefully constructed examples show that of all the privative adjectives, only *fake* has any context-sensitivity, and its context-sensitivity is much more constrained than that of *skillful*. Experiment 4 in Chapter 2 digs deep into the conditions under which *fake* is acceptable and concludes that it may target one or more "core" properties of the noun, as does *real*. We take which property of the noun it targets to be contextually supplied, but unlike *skillful*, this must be derived from the noun rather than freely supplied by the context. Chapter 3 shows how we can build this into *fake*'s lexical entry: the lexical entry invokes the context, but only to choose among the properties already provided by the noun.

***Fake* and *real* as windows into conceptual structure** Chapter 2 shows that *fake* and *real* appear to be sensitive to "core" properties of nouns, and not to other statistically correlated properties. Which property they target depends on the context. This places them in a class of their own, distinct from other "privative" adjectives which have more specific, context-insensitive meanings. This gives us an exciting reflex of language on a domain ordinarily considered outside the bounds of theoretical semantics, namely how we structure concepts. Chapter 3 shows how we can link semantics with conceptual structure using dual content semantics to capture this sensitivity, as well as explaining (constrained) contextual modulation of noun meaning (del Pinal, 2018).

Certainly, some linguists take the position that the meaning of content words like *cat* and their associated conceptual structure, if there is any, is outside the domain of linguistics. While this is a reasonable position, this dissertation shows that it has more significant consequences than previously appreciated: to take this position, we must also take the internals of adjective-noun modification to be entirely outside of linguistics: adjectives like *red*, *stone* and *fake* alike are all black boxes which mean whatever they mean and modify the noun in whatever way is appropriate for that meaning. We cannot posit that they use set intersection and we cannot infer anything about their (subsective or privative) inferences a priori. We can also not infer anything about how composition works from these inferences. However, I believe that taking this position misses out on the tight bridge between language and concepts, and moreover on what we can learn from language about how people structure concepts and how they compose meaning, at both the linguistic (formal semantic) and conceptual levels. Further, opening the black box of compositionality lets us think about how humans—and other language-using systems like neural language models—can or might use it to generalize, which will be the focus of the rest of the dissertation.

1.4.2 Compositional generalization

The most common argument for compositionality is that it is the best (or only) way to explain why we can understand the meaning of novel sentences or phrases (Szabó, 2012). This can

be phrased as generalizing from the meaning of known words and known syntactic structures (and possibly known combinations of them) to new combinations of them. The second thread of this dissertation studies generalization in language users: both humans and also large language models (LLMs).

The term "generalization" can take on a wide variety of meanings, especially in the context of LLMs (Hupkes et al., 2023). In the case of adjective-noun composition, generalization is the ability for humans or LLMs to interpret adjectives and nouns that they have seen before in combinations that they have not seen before. In particular, we want to know whether they are able to derive a (consistent) inference for these combinations. For humans, we firstly ask whether they are able to derive an inference at all (as opposed to choosing "Unsure"). Secondly, we measure whether a consistent inference is derived across participants, which would suggest a stable and shared mechanism of interpreting novel adjective-noun bigrams. For LLMs, we measure whether they can derive the inference(s) that humans derive for each bigram.

Generalization for humans While many linguists take it as given that humans can generalize (compositionally) to novel phrases, the data pattern for privativity we discover in Chapter 2 is so nuanced that we should not assume this for our data. This is particularly relevant for *fake*, which we show in Chapter 2 to contextually target different properties, such that *fake handbag* most likely means *counterfeit handbag* (and is a handbag) while *fake flower* most likely means *plastic flower* (and is not a flower in the strict sense). *Counterfeit* and *artificial* also occur predominantly with a small set of nouns, such as *counterfeit money* or *artificial intelligence* (or *artificial light*). It is in principle possible that people prefer to memorize the most common inference for these adjectives for each noun they are commonly used with, and so do not actually generalize well to novel bigrams.

Perhaps more plausibly, humans might generalize to some novel adjective-noun bigrams involving these adjectives, but they might generalize by analogy rather than compositionally, in the spirit of construction grammar analogy (Ambridge, 2020; Bybee, 2010). In this case,

we would expect them to only generalize consistently to adjective-noun bigrams that have a sufficiently similar analogical counterpart they have encountered before.

Chapter 2 incorporates this into our experiments by controlling for bigram frequency in an internet-scale corpus (C4; Radford et al., 2019) and including 180 bigrams that are presumed to be novel to humans, including some for which there is no obvious similar high-frequency bigram (presumed to be known), such as *fake concert*. We find that humans do generalize to novel bigrams, as measured by the fact that our participants derive consistent, clearly privative or subsective inferences for many of them.

Extending the paradigm to LLMs Chapter 4 extends the experiments from Chapter 2 to large language models (LLMs) and studies whether LLMs can draw human-like privativity inferences for bigrams assumed to be novel to them, as well as for bigrams assumed to be known. Since Chapter 2 showed that context affects the inference and narrows the human distribution of ratings, we test LLMs both with and without a provided context. When no context is provided, humans converge for some bigrams but provide a wide distribution of ratings for others, so we ask LLMs to match the whole distribution in order to qualify as "human-like". Surprisingly, we find that LLMs are very good at matching human inferences when a context is provided, and reasonably good at matching the inference distribution in the no-context case. More surprisingly, there is no effect of frequency for this task: LLMs are equally good at matching the human ratings for novel bigrams as they are for known ones, suggesting that whatever method they use to solve this task generalizes.

What can we learn from studying compositional generalization in LLMs? Studying compositional generalization in language-using systems other than humans helps us both design these experiments more rigorously and also set standards for finding evidence of composition in humans. Studying LLMs brings out the possibility that shortcuts like frequency or other types of reasoning like analogy could be used to solve the task, which are not always considered for humans. Controlling for these also leads to more rigor in the corresponding human studies, including ours. Further, previous discussion of composition and symbolic

reasoning more generally has often been defined by behavior on certain tasks (Fodor, 1998; Pavlick, 2025; Quilty-Dunn, Porot, and Mandelbaum, 2023). When experiments show that LLMs can do these same tasks or exhibit these same behaviors, as we will (mostly) find in Chapter 4 for privativity inferences, some people find themselves nonetheless hesitant to ascribe compositionality or symbolic reasoning to LLMs. The resulting debate may lead to more carefully defined tasks that exclude confounding factors, or may lead to the definition or required standards of evidence for the target concept being refined and strengthened. In particular, recent work on compositionality (linguistic and non-linguistic) in LLMs reveals that some researchers believe that evidence of symbolic representations is necessary as evidence for composition, and that behavioral measures are not sufficient (McCurdy et al., 2024). Each of these paths motivates new, more rigorous experiments in humans that meet these new requirements.

LLMs also have a second role to play in an account of human adjective-noun composition: we can use them to model aspects of the problem that are difficult to model in semantic theory, such as the effect of context, complementary to the formal theory. One observation from Chapter 2 is that the inference drawn for a given adjective-noun bigram depends on the context. A natural hypothesis is then that the sometimes widely varying distributions of privativity inferences for the out-of-the-blue setting can be partially explained by assuming that different participants imagine different "contexts" (or underspecified sketches of contexts) when they answer the question. This variation is hard to model in a standard semantic theory and is typically left to some opaque context variable c. In fact, we can use LLMs to model this variation and test this hypothesis. Since LLMs are assumed to capture statistical knowledge about the world based on their pretraining corpora, we can use them to generate a sample of contexts and then have humans or LLMs rate the inference in these contexts. If this setup matches the variation of the out-of-the-blue human rating distributions, then this is plausible evidence that the variation is indeed (partially) derived in this way. In Chapter 4, we do indeed find a relatively good fit to human out-of-the-blue distributions using this method,

although there is still some deviation which suggests room for future improvement in the experiment design.

Privative inferences by analogy Chapter 2 and Chapter 4 find that humans and LLMs can both generalize to inferences for novel adjective-noun bigrams. We would like to conclude that they do this by virtue of composing the meaning of the adjective and the noun. However, as we discussed, generalization can also in principle be accomplished by analogy (Ambridge, 2020; Bybee, 2010 i.a.). Perhaps participants answers that *a counterfeit scarf is a scarf* by analogy to more familiar items like *counterfeit watches*, which are watches. Chapter 5 investigates this hypothesis in detail. We build a computational model of analogy as well as conducting an experiment where we ask people to reason analogically. While we indeed find that much of the dataset can be handled by analogy, we crucially also find that not all bigrams have useful analogies – either participants cannot come up with any analogy at all, or the analogy that the model or participants use leads them "astray". (One example is *homemade currency*, where reasoning by analogy to *homemade jam* leads to a subsective instead of the expected privative inference.) Moreover, our analogy model's detailed behavior correlates poorly with the behavior of humans and with the behavior of our best LLM.

This suggests that humans are not solving this task by analogy, and are indeed using composition. The same conclusion can be drawn for LLMs, if we are using the same standard of evidence as for humans. Even if we are not willing to go that far, it is exciting and interesting that they do not appear to be using analogy over distributional semantics-based embeddings to solve the task, even though this would be an available and plausible path for them. This sets the stage for exciting future research to determine what LLMs are doing, and perhaps establish a better standard of what counts as composition and compositional generalization that can be applied to both humans and LLMs.

1.5 Overview of chapters

Chapter 2 contains three experiments to gather judgments from native English speakers on adjective-noun inferences on a large set of adjective-noun bigrams, including novel bigrams which test generalization. This sets up the data used in the remaining chapters. These three experiments were previously published as a conference proceedings as Ross, Kim, and Davidson (2025), co-authored with my advisors Najoung Kim and Kathryn Davidson. A fourth experiment in Chapter 2 studies the felicity of *fake* and what properties it can target. Chapter 2 concludes with three generalizations about adjective-noun composition and privativity.

Chapter 3 proposes a new theoretical account of "responsible" adjective-noun modification which puts the burden of on the adjective modification directly to handle conflicts and thus privativity, using a dual content semantics which makes reference to the conceptual structure of nouns and relates it to their extension via genericity. It also discusses previous accounts of privative adjectives in more detail, with a focus on the alleged context-sensitivity of *fake*, and finds that such accounts overgenerate.

Chapter 4 shifts gears to LLMs and studies their ability to draw human-like privativity inferences and generalize to novel bigrams. It contains three experiments which evaluate the same data gathered for humans in Chapter 2 for LLMs, as well as a discussion of the best method to use to evaluate LLMs with respect to this kind of high-variance human data. This chapter, minus the discussion about standards for evaluating compositionality, was previously published as a conference proceedings as Ross, Davidson, and Kim (2024), co-authored with my advisors Kathryn Davidson and Najoung Kim.

Finally, Chapter 5 asks whether we can explain the generalization seen in Chapter 2 in humans and in Chapter 4 LLMs by virtue of analogy, or whether we can genuinely conclude that they are each using composition to solve the task. It contains a computational model of analogy as well as one last experiment that gathers human judgments on analogical reasoning. This chapter will be published in the proceedings of SCiL 2025, co-authored with my advisors Kathryn Davidson and Najoung Kim.

Chapter 2

Fake reefs are sometimes reefs and sometimes not [1]

2.1 Introduction

This chapter follows a growing tradition (del Pinal, 2015, 2018; Martin, 2022; Partee, 2009, 2010, i.a.) in studying the dynamic interaction of meaning and context through the lens of (privative) adjective modification and how to account for it compositionally. Privativity poses a potential challenge for compositionality, which requires the meaning of a complex expression to be derived solely from the meaning of its constituent parts (Szabó, 2012): the set of *guns* combined with whatever function or set *fake* denotes. If a *fake gun* is not a *gun*, it is not clear how to derive the meaning of *fake gun* from the set of (real) *guns* through set operations such as subsection. Accounting for privativity thus requires us to posit a more complex account of adjective-noun composition.

As we discussed in Chapter 1, Martin (2022) shows that inference patterns for so-called privative adjectives vary depending on the noun used. For example, *counterfeit* may license a privative or subsective inference depending on the noun (and accompanying context).

[1] Parts of this chapter were previously published as Ross, Kim, and Davidson (2025) ("*Fake reefs* are sometimes reefs and sometimes not, but are always compositional"), co-authored with my advisors Najoung Kim and Kathryn Davidson.

(11) *Subsective inference* This is a counterfeit watch.

∴ This is a watch.

(12) *Privative inference* This is a counterfeit dollar.

∴ This is not a dollar.

This per-adjective and per-noun variation raises an additional question of whether these adjective-noun combinations and their inferences are computed (compositionally) on the fly, based on just the given adjective, noun and context, or whether there is an element of convention or past experience necessary to derive these varying inferences, in which case the inference would be stored (memorized). A significant body of processing work (Arnon and Snider, 2010; Caldwell-Harris, Berant, and Edelman, 2012; O'Donnell, 2015; Tremblay and Baayen, 2010, i.a.) reveals plenty of cases where humans *don't* appear to compose meaning on the fly: chunks of various sizes from multi-morpheme words to entire idiomatic expressions, especially highly frequent words or expressions, can get stored as units and trigger priming effects in experimental studies, whether their meaning is idiomatic or fully compositional from their parts. If the effect of adjectives with privative inferences is stored rather than composed on the fly, then deriving the inference for infrequent adjective-noun bigrams with such adjectives, such as *fake scarf* or *fake reef*, might be difficult or result in widely varying results between people. The same might be true for intermediate, less memorization-heavy approaches such as learning (memorizing) the inferences for some high-frequency bigrams and then reasoning about novel bigrams by analogy where possible.[2] This chapter explores the effect of experience (as measured by corpus frequency of the bigram) and context on adjective-noun combination and inferences, especially for novel (zero corpus frequency) adjective-noun bigrams.

We gather a large quantity of adjective-noun inference judgments for both high-frequency and novel / zero corpus frequency adjective-noun bigrams over three experiments and show that inferences depend not just on the adjective but also on the noun and the context. Fur-

[2]For example, a participant might reason that the novel bigram *counterfeit scarf* is a *scarf* by analogy to other clothing items and accessories such as *watch* or *handbag* which they have seen *counterfeit* occur with subsectively.

ther, we show that novel adjective-noun bigrams and their privativity inferences are handled as productively and consistently by participants as those of high-frequency ones, despite the significant variation by adjective, noun, and context. Thus, any theory of adjective-noun combination must address the context-sensitivity of these inferences and predict the ability to generalize to novel combinations, for example using a compositional approach; it cannot be based on memorized idiomatic meanings. In Section 2.5, we discuss two compositional accounts of adjective-noun modification which satisfy these requirements and can handle the new data presented in this chapter.

2.2 Choice of adjective-noun bigrams

We first establish a set of 798 bigrams to study, 23% of which are zero frequency in a large corpus, thus presumed novel for participants. The full set of bigrams, as well as results for all experiments in this chapter, are available on GitHub.[3]

2.2.1 Selection by corpus frequency

We consider 6 "privative" adjectives of interest: *fake*, *counterfeit*, *false*, *artificial*, *knockoff* and *former*. Since we established in the introduction that such adjectives need not always be privative, from here on out the phrase "privative adjective" will refer to these adjectives such as *fake* that have been discussed in prior literature as (typically) resulting in privative inferences. We select 6 intersective/subsective adjectives as "controls" which each have a similar frequency to one of the privative adjectives in a very large corpus (C4, ca. 130 trillion words; Dodge et al., 2021; Raffel et al., 2020) and which have relatively few selectional restrictions: *useful, tiny, illegal, homemade, unimportant and multicolored*. Frequencies are shown in Table 2.1. We choose *multicolored* as a low-frequency example of a (standardly intersective) colour adjective, *illegal* since it has negative valency while typically being subsective, and *homemade*

[3] `https://github.com/rossh2/artificial-intelligence`

since it targets the manner of manufacture, similar to *counterfeit* and *artificial*, without being obviously privative.

Table 2.1: *Adjective frequencies in the C4 Corpus (130T words)*

Adjective	Tokens	Adjective	Tokens
former	15.8M	useful	13.6M
false	4.6M	tiny	5.8M
artificial	3.9M	illegal	4.5M
fake	3.1M	homemade	2.2M
counterfeit	450K	unimportant	170K
knockoff	57K	multicolored	93K

We algorithmically select 43 nouns from 300 nouns which commonly occur with a wide range of adjectives (Pavlick and Callison-Burch, 2016a), plus the 36 nouns used in Martin (2022), with the goal of generating a high number of zero-frequency bigrams. We then manually select an additional 59 nouns which are semantically similar to these 43 nouns, for a total of 102 nouns. We cross these 102 nouns with the 12 adjectives for a total of 1224 bigrams to use in subsequent experiments. We determine (relative) bigram frequency by counting the frequency of all bigrams involved in this process in C4, for a total of 3979 bigrams (358 unique nouns × 12 adjectives, plus experiment fillers), since calculating the frequency over every possible corpus bigram would be prohibitively expensive. Thus, terms like "high frequency bigram" or "top quartile bigram" in this chapter should be interpreted in relative rather than absolute terms.

2.2.2 Experiment 1

Experiment 1 filters out clearly nonsensical combinations resulting from the blind crossing of adjectives and nouns. Combinations like *counterfeit accusation* or *multicolored effort* must be excluded before we can reasonably ask questions like "Is a counterfeit N still an N?" Participants are presented with a bigram and asked "How easy is it to imagine what this would

mean?", as shown in Figure 2.1.[4] 144 native American English speakers[5] were recruited on Prolific (of which 7 were excluded due to failed attention checks and/or not meeting the criteria for native English speaker); the study was implemented in Qualtrics. Participants were paid pro rata at $12/hour; the experiment took 4 minutes on average. Each participant saw 14 questions (12 target bigrams, 2 fillers), for 3 ratings per bigram in total.

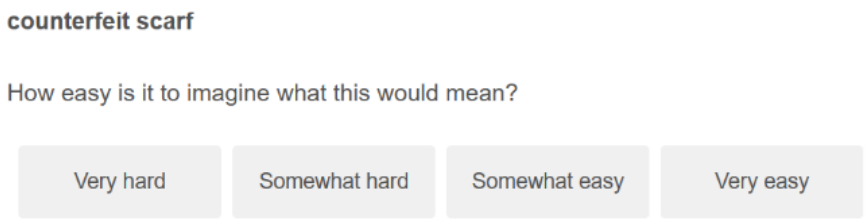


Figure 2.1: *Screenshot of question in Experiment 1*

We categorize bigrams whose ratings were majority "very hard" or "somewhat hard" as nonsensical, and exclude them from subsequent experiments. This leaves 798 bigrams, of which 23% (180) are zero frequency in C4, i.e. almost certainly novel to new participants, and another 21% (170) are low frequency (bottom quartile), so also quite possibly novel to participants.

2.3 Experiment 2: *Is an A N an N?*

2.3.1 Method

Experiment 2 asks participants *Is an A N still an N?* for each of the 798 adjective-noun bigrams left after filtering in Experiment 1. An example question is shown in Figure 2.2. We choose

[4]Previous work studying novel adjective-noun combinations (Vecchi et al., 2017) uses a more complex pairwise ranking approach to precisely measure semantic deviance, but we only need to filter out obviously nonsensical bigrams.

[5]We recruit people on Prolific who self-report English as their first and primary language and are located in the United States. We further ask them at the end of the study whether they learned English before the age of 5 and whether they speak American English as opposed to another dialect of English (if not, they are paid but excluded). This implementation of "native speaker" is merely intended as a practical way to expect shared language experiences among our participant sample (Cheng et al., 2021).

to use the same design as Martin (2022), with the slight modification of adding *still* to make the question more natural.[6]

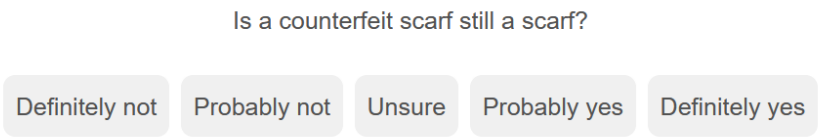


Figure 2.2: *Screenshot of question in Experiment 2 on PCIbex*

This differs from previous privativity studies (Pavlick and Callison-Burch, 2016a,b; Pustejovsky, 2013) which ask participants to rate these inferences given a particular sentence context drawn from a corpus. For example, Pavlick and Callison-Burch (2016b) ask whether (13a) entails (13a). While more realistic than out-of-the-blue judgments, this also creates a much noisier picture, as demonstrated by this example: participants rate (13a) as contradicting (13a), but the verb *denied* and the world knowledge of pharmacists selling medicine also seem to be driving part of this inference. It actually remains unclear whether the counterfeit medicine that the pharmacists were selling qualifies as medicine in this scenario.

(13) a. Pharmacists in Algodones denied selling counterfeit medicine in their stores. Pharmacists in Algodones denied selling medicine in their stores.

This experiment aims to show, for a wider range of adjectives and nouns than Martin, 2022, that privativity varies depending on the noun, and investigates whether participants behave differently for high and zero frequency (assumed to be novel) bigrams. We will study non-out-the-blue judgments (in a more controlled setting than Pavlick and Callison-Burch) later, in Experiment 3.

[6]Using *still* seems to help foreground the idea that adding the adjective might change noun membership: a *fake scarf* or *unimportant sign* might *not* be a scarf or sign, or conversely might be a scarf *despite* being fake.

We ran this experiment in two parts. For the first 305 bigrams, 510 native American English speakers were recruited on Prolific (of which 15 were excluded due to failed attention checks and/or not meeting the criteria for native English speaker); the study was implemented in PCIbex (Zehr and Schwarz, 2018). For the next 498 bigrams, 756 native American English speakers were recruited on Prolific (of which 24 were excluded); this study was implemented on Qualtrics. Participants were paid pro rata at $12/hour and the experiment took 3 minutes on average. Each participant saw 12 questions (4 typically-intersective adjectives, 4 typically-privative adjectives, 4 fillers), for a total of approx. 12 ratings/bigram.[7] Since some bigrams which may not make sense to everyone likely remain after Experiment 1, we explicitly alert participants in Experiment 2 to this possibility and instruct them to use the "Unsure" rating if a combination does not make sense to them.

2.3.2 Results

Mean bigram ratings are shown in Figure 2.3 (organized by adjective, each dot represents the mean rating for one adjective-noun bigram), and individual ratings by participants for a selection of bigrams are shown in Figure 2.4. We find that each so-called "privative" adjective in fact yields graded variation from privative to subsective depending on the noun, with ratings spanning all the way from 1 ("Definitely not [an N]") to 5 ("Definitely yes [an N]"). In Figure 2.4, we can also see that intermediate means are often associated with high variance rather than participants agreeing on "Unsure". We also find that "subsective" adjectives are usually subsective ("Probably yes" or "Definitely yes"), warranting the name, but are nonetheless not so clearly subsective with certain nouns (e.g. *homemade cat* with $\mu = 2.6$, *illegal currency* with $\mu = 2.83$).

Secondly, we find no effect of bigram frequency on rating variance. A linear regression in R (R Core Team, 2023) shows that bigram frequency correlates poorly with the variance in the ratings (typically subsective: $R^2 = 0.003$, typically privative: $R^2 = 0.010$, both $p >$

[7] Due to issues with PCIbex, the first part of experiment did not yield an even number of ratings per bigram. In the analysis of this experiment, we randomly sample and cap the number of ratings at (10-)12 ratings/item.

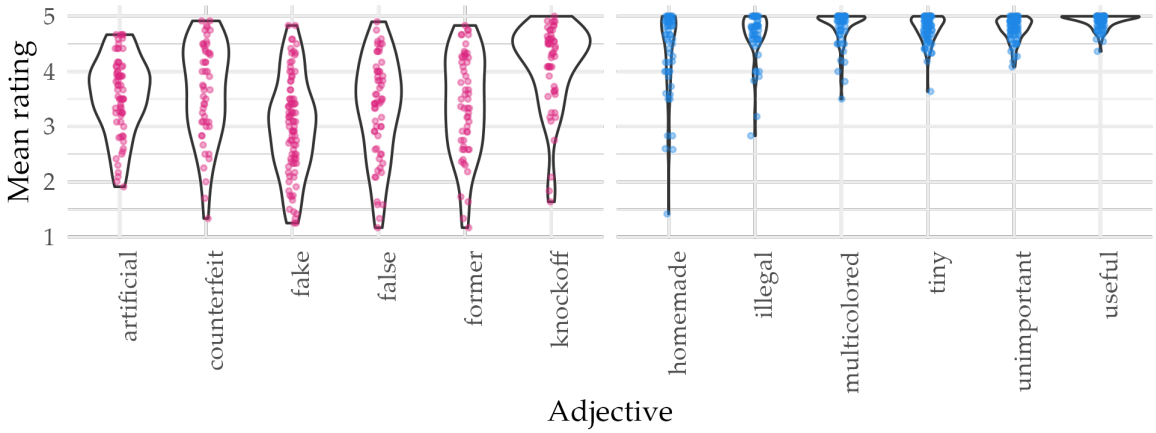


Figure 2.3: *Mean ratings for "Is an AN an N?" for each bigram by adjective in Experiment 2, where 1 is most privative ("Definitely not") and 5 is most subsective ("Definitely yes").*

0.05). Instead, participants agree to a similar degree on the meaning and inferences for high-frequency and zero frequency (novel) adjective-noun bigrams. Some high frequency bigrams such as *artificial tree* or *former house* show high variance in ratings ($\mu = 3.50, \sigma^2 = 1.83$ and $\mu = 3.63, \sigma^2 = 1.76$ respectively), suggesting that these bigrams do not have a conventionalized meaning or inference when presented out of the blue. Moreover, some zero frequency bigrams like *knockoff image* and *counterfeit scarf* have quite low variance ($\mu = 4.90, \sigma^2 = 0.10$ and $\mu = 4.80, \sigma^2 = 0.18$), suggesting that participants compose even novel bigrams systematically.

2.3.3 Discussion

The results from Experiment 2 lend further weight to previous work illustrating that no adjective is unequivocally privative, but rather that privativity depends on the combination of adjective and noun (Martin, 2022). We further see no correlation between rating variance and bigram frequency. We conclude that high frequency need not lead to a fixed conception of bigram meaning or inference, and that previous exposure to a (potentially) privative adjective-noun pair is not required to draw this inference, even for adjectives with relatively broad meanings like *fake*. Instead, we suspect that high variance may be due in part to participants imagining different contexts for the bigrams (which were presented out of the blue

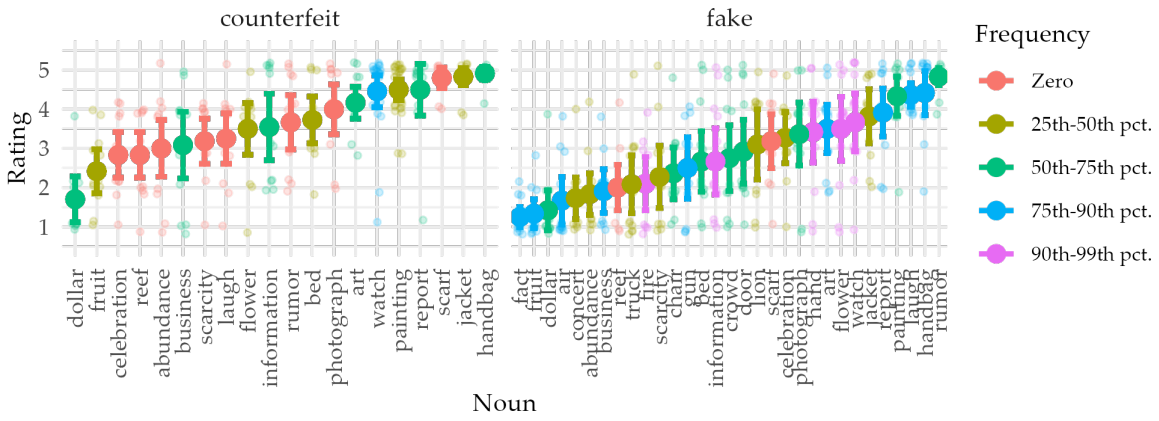


Figure 2.4: *Participant ratings for a selection of bigrams involving fake and counterfeit in Experiment 2, where 1 is most privative and 5 is most subsective. • shows the mean with SE for each bigram.*

in Experiment 2), such that e.g. *fake* might target different aspects of the noun's properties or different properties might be relevant for noun-hood in that context. For example, a *fake crowd* might qualify as a *crowd* if it is made up of paid actors, but less so if it is just painted dummies on a movie set.

2.4 Experiment 3: Context

2.4.1 Method

For 28 adjective-noun bigrams from Experiment 2, we construct two contexts each intended to bias the reader towards a subsective or privative inference respectively. Two example contexts for *fake fire* are shown in Figure 2.5. We targeted 6 pairs of adjective-noun bigrams from Experiment 2 with intermediate mean ratings and high variance, such that one bigram is zero/low frequency and the other is high frequency; we will use these pairs to investigate any effect of frequency. We then selected an additional 16 bigrams with intermediate mean ratings and high variance for which we were able to write convincing example contexts. We select these bigrams which are neither at ceiling or floor precisely because we suspect that there may be more than one context in participants' minds, and thus the specified contexts might split apart these middle-of-the-scale ratings into high (subsective) or low (privative)

ratings respectively, explaining (some of) the variance in Experiment 2. Further, we select pairs of high and zero/low frequency bigrams because it is possible that high frequency bigrams might come with more conventionalized contexts and/or more fixed meanings and inferences in general, and thus might resist manipulation by provided contexts. Conversely, zero/low frequency bigrams might be particularly easy to manipulate, since they lack any preconceived "default" context.

For the first 12 bigrams, 40 native American English speakers were recruited on Prolific (of which 1 excluded due to failed attention checks); for the second set of 18 bigrams (two bigrams were rerun), a further 40 native American English speakers were recruited on Prolific (of which 2 were excluded for not meeting the native speaker criteria). Both studies were implemented in Qualtrics. Participants were paid pro rata at $12/hour; the experiment took 8 minutes on average. In the first instance of the experiment, each participant saw 12 items as shown in Figure 2.5; in the second instance, each participant saw 18 items (3 or 6 intersective-biased, 3 or 6 privative-biased, 6 fillers), yielding 10 ratings/item.

2.4.2 Results

We find that across the board, writing biased contexts does indeed shift participants ratings in the intended direction, as shown in Figure 2.6, as well as reducing the variance. The one exception is *counterfeit dollar*, which refuses to be influenced by context at all. This can be explained simply by its meaning: *dollars* depend so heavily on having an authentic method of manufacture that any way in which they can be counterfeited, i.e. in which their method of manufacture is non-conventional, robs them of being a dollar. We fit an ordinal mixed effects model in R (Christensen, 2022; R Core Team, 2023) and find statistically significant effects for both the subsective-biased and privative-biased contexts compared to having no context ($p < 0.05$ for both; a subsective-biased context makes a high (subsective) rating 4x more likely while a privative biased context makes a high (subsective) rating only $^1/_6$x as likely). As in Experiment 2, we find no effect of frequency in this experiment: high-frequency bigrams do not have more fixed inferences and are not more resistant to being manipulated by context.

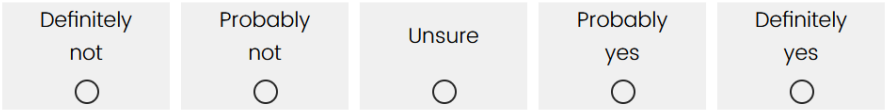
Gary is in charge of special effects for a new action film at a small movie studio. In this scene, the building that the agents are in gets set on fire and they have to make a dramatic escape through the window. Gary has rigged the building on the movie set with special pieces of oily fabric which go up the side of the building and will burn very easily, making it look like the whole building is on fire. When it's time to film, Gary lights the fabric on fire and carefully monitors it as the flames climb up the building. Gary's manager is very impressed with the fake fire.

In this setting, is the fake fire still a fire?

(**a**) *Subsective-biased context*

Julie is very excited to visit her grandparents, because she wants to make marshmallows over the fire in their fireplace. When she arrives, she discovers that it's actually a fake fire, made with flickering lights and waving pieces of fabric, which is only slightly warm.

In this setting, is the fake fire still a fire?



(**b**) *Privative-biased context*

Figure 2.5: *Screenshots of contexts for fake fire in Experiment 3.*

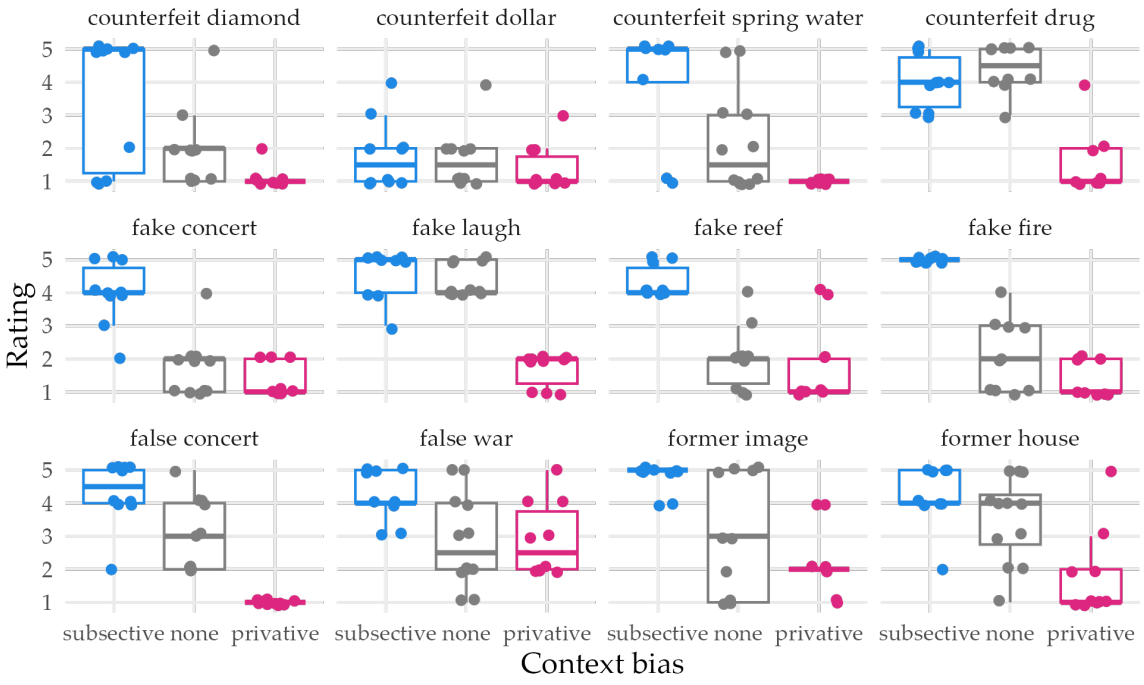


Figure 2.6: *Ratings for "In this setting, is an AN still an N?" for 12 of 28 bigrams in Experiment 3, where 1 ("Definitely not") is most privative and 5 ("Definitely yes") is most subsective. The ratings from Experiment 2 are shown in gray. The first and third columns show zero or low frequency bigrams; the second and fourth columns show corresponding high frequency bigrams.*

2.4.3 Discussion

The results from Experiment 3 show that the subsective/privative inferences drawn from adjective-noun combination are indeed context-dependent: providing different contexts can cause participants to draw quite subsective or privative inferences for the very same adjective-noun bigram. This is possible whether the bigram is high-frequency (thus potentially coming with a "default" or conventionalized context of use) or novel. Thus it is likely that the participants' imagined contexts explains some of the variance in Experiment 2, which presented the bigrams out of the blue. In other words, the meaning of adjective-noun bigrams and their privative inferences cannot be explained by memorization of a single (conventionalized) meaning or inference, since inferences must be computed productively on the fly based on the provided context (as well as world knowledge). Finally, the ability to manipulate the inferences of novel bigrams such as *false concert* supports a compositional account

for the meaning of the bigram (where context is included as part of the composition and inference-drawing process), as in Experiment 2.

2.5 Impact on existing theoretical accounts

Our experiments showcase the wide variation in privative inferences among so-called privative adjectives: first, adjectives can license either a subsective or privative inference depending on the noun (and context), and second, the same adjective-noun bigram can license either inference depending on context. These results support a compositional, context-dependent account of adjective-noun modification rather than an approach based on prior experience or convention which memorizes the meanings and/or inferences for previously encountered bigrams. Further, these results pose challenges for theories which treat privativity as a property of only the adjective, such as Partee (2010), del Pinal (2015) and Guerrini (2024a,b).

2.5.1 Widening accounts

As laid out in Chapter 1, Partee's classic account of privative adjectives (Kamp and Partee, 1995; Partee, 2007, 2009, 2010) posits that all seemingly privative adjectives in fact compose subsectively with the noun. Unlike regular subsective adjectives, however, $[\![\text{fake gun}]\!] = \emptyset$ initially, since *gun* includes only real guns and *fake* is privative (by definition). The *Non-Vacuity Principle* then coerces *gun* to expand to include both real and fake guns, so that *fake* can now act subsectively over this new expanded set. It seems reasonable to assume that Partee also intends her account to extend to other so-called privative adjectives besides *fake*, like *counterfeit* and *artificial* in English, since she cites adjectives with those meanings in Polish (Partee, 2010).

In light of the many subsective inferences with these adjectives that we have seen in these experiments, Partee's account holds up, but does not provide a lot of explanatory power. Presumably, she does not expect *stone* to always be privative, and so presumably the Non-Vacuity Principle is invoked for *stone lion* but not for *stone bowl*. Thus, presumably, *fake* may similarly yield subsective inferences, although Partee does not mention this possibility ex-

plicitly. However, Partee's account does not allow us to predict under what circumstances we yield which inference, nor why the context can shift the inference for a given bigram, such as *fake fire* in Experiment 3.

2.5.2 Dual content accounts

As discussed in Chapter 1, del Pinal (2015, 2018) provides a dual content semantics account where nouns have a conceptual structure and adjectives like *fake* draw on the C-structure of nouns like *gun* to build the new E-structure for *fake gun*. Del Pinal (2018)'s implementation of *fake* modifies *gun* to yield the semantics in (10), repeated here, for *fake gun*: firstly, in the E-structure, *fake guns* are not in the extension of guns. Secondly, *fake guns* do not have the origins of guns (the agentive property), instead, they were made to appear as if they were guns.

(10) $[\![\text{fake gun}]\!] =$

E-structure: $\lambda x.\ \neg E([\![\text{gun}]\!])(x) \wedge \neg Q_A([\![\text{gun}]\!])(x) \wedge$
$\exists e_2\ [\text{making}(e_2) \wedge \textsc{goal}(e_2, Q_P([\![\text{gun}]\!])(x))]$

C-structure: ...

While del Pinal's theory lays out in much more detail than Partee how the composition works, del Pinal (2015) stipulates that *fake guns* are not *guns* by fixing in the E-structure that a *fake gun* is not in the extension of *gun*: $\neg Q_E([\![\text{gun}]\!])(x)$. In subsequent work, del Pinal (2018) admits that this part of the E-structure is questionable for *counterfeit* and *artificial* and that it is an empirical question whether this should be included or not. Empirically, both we and Martin (2022) find that it should not be included for any "privative" adjective.

Martin (2022) adjusts del Pinal's definitions to remove privativity in the E-structure. Instead, privativity arises after composition of the adjective and noun, when set membership of the newly composed object is determined. If a targeted, negated dimension (primarily A, origin, for *fake*; T, telic, is also negated in the C-structure but not in the E-structure) is "central" to the meaning of the noun, i.e. part of the E-structure, then this results in privativity.

Which dimensions (analogous to "typical properties" of the noun) are incorporated into the E-structure is context-dependent, as it is for any use of a noun. This allows the modified account to capture the variation in our experiments by noun and by context.

2.5.3 Contextual variable-based accounts

Guerrini (2024a,b) and Reinert (2024) propose accounts of *fake* which are context-dependent: *fake* does not in fact compose with the noun directly but rather with some salient, contextually specified property, which is typically the noun but does not have to be. While Reinert (2024) does not give a more detailed lexical entry for *fake* Guerrini (2024b) further proposes a semantics for *fake* based on similarity (see Section 3.5.1), which involves negation of the noun as well as a similarity component. By virtue of this negation, privativity is explicitly part of Guerrini's definition of *fake*, following his interpretation of Partee (2010). Unlike del Pinal, Guerrini is committed to this fact. He argues that the variability in privativity in the experiments in Martin (2022), which we extend in this chapter, can be explained by the context supplying different nouns that *fake* targets. When *fake watch* has a subsective reading, *fake* (privatively) modifies another covert, contextually supplied noun, such as *Rolex*, such that *fake watch* actually means [*fake* ~~*Rolex*~~] *watch*. (*Fake Rolex* combines intersectively with *watch*.) When it has a privative reading, we have [*fake* ~~*watch*~~] *watch*, and a Partee-style mechanism widens the second instance of *watch* so that the intersection succeeds. Thanks to this context-dependence, Guerrini (2024a,b) and Reinert (2024) also account for the data presented in Experiments 2 and 3. One concern with this account is whether the contextually supplied material has to be a single noun over which *fake* acts privatively, or whether any property we can denote in our set theory suffices. Our data suggests that finding an appropriate noun would be difficult for items such as *fake fire* in Figure 2.5, or *fake concert* in (14), though there is probably a multi-word phrase or concept that can satisfy the theory.

(14) Context: A political party disguises a fundraiser as a concert so that they can hold it at a venue where political rallies aren't allowed. They even hire an up-and-coming band to sing at the event. The fake concert is a great success and the attendees enjoy the music as well as networking with the political candidates.
∴ This is a concert. (*according to experiment participants*)

2.5.4 Interim conclusion

Experiments 1-3 present experimental evidence on the variation in subsective vs. privative inferences both within adjectives and within adjective-noun bigrams, including for novel adjective-noun bigrams. We find that no adjective always yields privative inferences, lending further weight to Martin (2022) and Pavlick and Callison-Burch (2016b). We further find that for most adjective-noun combinations, privativity depends on the context as well as the adjective and the noun. Our results show that any theory of adjective-noun combination must account for the context-sensitivity of these inferences and allow generalization to novel combinations (for example, by composition)—these inferences are not so unpredictable as to need to be memorized. Theories of privativity which are not compositional (e.g. basic set complementation) or which fix privativity as a property of individual adjectives (del Pinal, 2015; Partee, 2010) with only a single method of composition do not account for the full set of our data. Compositional accounts like Martin (2022)'s modification of del Pinal (2015, 2018) and Guerrini (2024a,b) which predict that privativity is context-dependent, either by having privativity arise outside of the composition or by appealing to syntactic ambiguity, are able to account for the generalization and context-sensitivity found in our experiments. These compositional approaches aim to explain why participants are equally able to draw inferences for novel bigrams as for high-frequency ones, and what inferences we should expect given the effect of the context on available nouns or restrictions of noun denotations.

In the Section 2.7, we will dive deeper into the meaning of *fake* and explore its seemingly context-sensitive meaning and inferences. This will enable us to phrase a more comprehensive set of generalizations in Section 2.8 of how privativity and adjective-noun modification

more broadly seem to be working. In Chapter 3, we will then propose a new theoretical account of adjective-noun modification that implements these generalizations.

2.6 Even more examples of privativity

2.6.1 Material adjectives

In addition to our main dataset, Experiment 2 tested two classic examples of privativity used in the literature, *fake gun* and *stone lion*, alongside two other examples of privative inferences with material adjectives given by Partee (2010), *wooden horse* and *velvet*(*een*) *rabbit*. Figure 2.7 shows the participant ratings. On one hand, none of these examples are as clearly privative as bigrams like *fake dollar* or *fake fact*; on the other hand they clearly do not pattern with subsective adjectives behaving subsectively, like most uses of *tiny*. Closer inspection reveals a split in populations, with several participants rating e.g. *stone lion* as clearly not a lion while others rating it as clearly a lion. We attribute this to different choices in what counts as a lion: there is evidently a sense in which a *stone lion* is not a lion, but there are also other situations, such as depiction or play, where it has a sufficient quantity of lion-characteristics to be called a lion. Moreover, *stone lion* patterns very similarly to *fake lion*, suggesting again that *stone* and *fake* are not that dissimilar.[8] As Experiment 3 shows, it is important to fix the context to establish the standards for lion-hood, and also to clarify what *fake* is targeting in the case of *fake lion* (this is less important for *stone* or *counterfeit*). In the subsequent discussion, we will assume that *stone lion* and its material adjective kin are indeed privative under a sufficiently narrow reading of *lion*. We will return to the idea that context establishes the standards for lion-hood in 3.6.1.

[8]For *fake gun* and *fake lion*, we instead attribute this split distribution to a vagueness in what *fake* targets; one possible reading of *fake* is *counterfeit*, for example (more prominent for watches or handbags than for guns), which then yields a subsective reading. As we will explore more in Section 2.7, *fake* has a broader meaning than *stone* and can target various different properties.

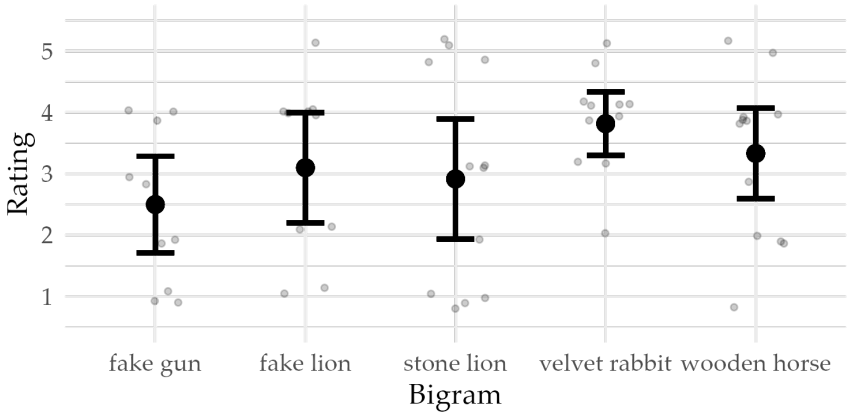


Figure 2.7: *Participant ratings for fake gun, fake lion and three bigrams with material adjectives (taken from Partee (2010)) in Experiment 2, where 1 is most privative and 5 is most subsective. • shows the mean with SE for each bigram.*

2.6.2 Color and shape adjectives

We saw in Experiment 2 that even adjectives like *illegal*, *tiny* and *multicolored*, which seem on their surface to be straightforwardly intersective or subsective, can raise the spectre of privative inferences: 6 participants rated *illegal currency* to be not or probably not currency, with a mean rating of 2.83, similar to *stone lion* at 2.91. *Tiny abundance* is rated at 3.63 on average, with one participant judging it privative and three unsure (compared to most bigrams with *tiny* which are at the ceiling of 5). Even color adjectives, often cited as the paradigmatic example of intersectivity, are not safe: *Multicolored gold* is rated privative by two participants, with an average of 3.50 (likely arising from multiple possible interpretations of gold-as-color or gold-as-paint vs. gold-as-material).

Moving beyond the items in the experiment, we can illustrate that color adjectives can be privative by considering money. Money is a useful example because its every aspect is strictly controlled, meaning that most modifications result in privativity. Unlike US dollar notes, euro notes are colored and sized by denomination. A €10 note is red and 127 × 67 mm, a €20 note is blue and 133 × 72 mm, a €50 note is orange and 140 × 77 mm, and so on. Thus, a *red €50 note* is a forgery, and not a *€50 note*.

(15) This is a red €50 note.

∴ This is not a €50 note.

The same applies to shape adjectives such as *hexagonal*, sometimes considered to be more paradigmatic examples of intersective adjectives (Guerrini, 2024c)[9] However, a *hexagonal €50 note* is also not a *€50 note* (unless it is a correctly shaped €50 note, i.e. a rectangle of 140 × 77 mm, folded into a hexagon), suggesting that even these adjectives are not immune to privative inferences.

(16) This is a hexagonal €50 note.

∴ This is not a €50 note.

It seems like in fact any adjective can result in a privative inference if the property that it supplies contradicts what the noun requires. For some adjectives, such as nationality adjectives like *Greek*, it is certainly harder to find a noun or noun phrase that would require some other nationality as well as country of origin/descent. But I think such a phrase can be imagined.

2.7 Experiment 4: The meaning of *fake*

The contexts in Experiment 3 demonstrate that *fake* can target a wide variety of properties, ranging from the fact that a fire is not burning at all to not burning the right thing (Figure 2.5), or that a watch or scarf is counterfeit. However, Martin (2022) shows that *fake* cannot just target any property: in (17), we cannot call this object a *fake gun* (though we can call it a *fake* [*metal gun*]). It seems that *fake* cannot target the fact that guns are typically made of metal.

[9]Guerrini (2024c) suggests that color adjectives are not actually examples of simple intersective adjectives, and instead take a context-sensitive parameter specifying what part of them is colored (as in *red pen* meaning *red-inked pen* or *pen red on the outside*). Guerrini suggests *Greek* and *hexagonal* as possible intersective examples.

(17) Context: Josh works at a metalworking shop and has just recently purchased a collection of guns. His only interest in purchasing the guns is to melt them down; the metals that guns are made of are valuable to him for his next project, so it's irrelevant whether the guns are broken or function. When going through the pile and melting each gun down, he finds that one doesn't melt properly, and upon closer inspection is actually made of a futuristic, high-quality plastic. It functions perfectly well, with the same internal mechanisms as a standard gun, but has none of the same materials. The manufacturer didn't want their guns to stand out, for some reason—perhaps the army didn't want to reveal their new technology—and so the plastic is designed and painted deceptively to look precisely like a standard metal gun.

(adapted from Martin, 2022)

a. #Josh says: "They sold me a fake gun!"

b. Josh says: "They sold me a fake metal gun!"

c. Josh says: "They sold me a gun."

The phrase *fake gun* suggests that something more fundamental is wrong with the gun, such as it not being designed to shoot. The key distinction here appears to be that being made of metal is merely a common property of guns, while shooting is a core property. Of course, being made of metal is very helpful for handling the heat and forces generated while shooting bullets, causing a high correlation the two properties. A survey of the contexts in Experiment 3 (as well as a good sample of LLM-generated contexts from Experiment 7; see Chapter 4) suggests that *fake* seems to target properties that are in some sense "core" to the noun or related to the noun's "essence". (17) suggests that it cannot target properties that are merely statistically correlated with such core properties. This section seeks to explore and quantify this idea. Further, we will argue that *real* exhibits the same behavior, putting *fake* in a natural class with *real* and perhaps *true* (not tested experimentally).

2.7.1 A possible definition of "core" property for *fake*

Prasada and Dillingham (2006) provide a framework and three experiments to distinguish statistically connected properties from other more "core" properties, situated with a knowledge or "theory" approach to concepts that assumes that concepts are not definitions or prototypes, but rather related to world knowledge and to each other (see e.g. Murphy, 2002 for an overview of theories of concepts accessible to linguists). Prasada and Dillingham provide three experiments with native English speakers to divide properties of concepts into two types of properties, which they call k-properties and t-properties. Intuitively, k-properties are determined by the type of thing something is, while t-properties are not. Specifically, Prasada and Dillingham argue that we represent *principled* connections between nominal concepts, viewed as types of tokens, and their k-properties. K-properties (i) license the expectation that tokens of the type will generally possess the k-property (as evaluated by statements involving kinds and generics), (ii) license explanation of the fact that a token of a type has a k-property by reference to its type, and (iii) license normative expectations. T-properties, meanwhile, are linked to concepts via statistical connections and are represented differently, even if those statistical correlations may be very strong. This perspective would allow us to class 'guns are made of metal' as a t-property of guns, and thus accord it lower status than e.g. being able to shoot, a property which *fake* can target and may well be a k-property.

Prasada and Dillingham (2006) show that k-properties (principled connections) are preferred over t-properties (statistical connections) in the following contexts:

1. Properties ascribed to bare plural kinds in generic sentences using *by virtue of* (*their Experiment 1*)
2. Explanations of properties by referring to the concept ("type") (*their Experiment 2*)
3. Normative statements about properties using *by virtue of* (*their Experiment 3*)
4. Properties ascribed to singular kinds in generic sentences
5. Properties ascribed to definite, taxonomic / well-established kinds in generic sentences
6. Properties in explanations answering "What's an N?"

As an illustration, here are all these contexts with their example k-property *dogs are four-legged* vs. their example t-property *barns are red*.

(18) a. Dogs, by virtue of being the kinds of things they are, are four-legged.

b. #Barns, by virtue of being the kinds of things they are, are red.

c. Why does that (pointing to a dog) have four legs? Because it is a dog.

d. #Why is that (pointing to a barn) red? Because it is a barn.

e. Dogs, by virtue of being dogs, should have four legs.

f. #Barns, by virtue of being barns, should be red.

g. A dog is four-legged.

h. #A barn is red.

i. The dog is four-legged.

j. #The barn is red. (*this particular infelicity may stem from* barn *not being a sufficiently good* (*taxonomic*) *kind, but the point extends to t-properties of natural kinds*)

k. What's a dog? A dog is a four-legged animal.

l. #What's a barn? A barn is a red building.

Greenberg (2003) discusses a similar contrast between bare plural generics and indefinite singular generics: both can involve a "definitional" or "in virtue of" accessibility relation between the noun and the property, but bare plurals can also involve a "descriptive" or "inductive" relation. Indefinite singular generics further require that the property is associated with the noun via known facts, norms or stereotypes and that the noun is a "reasonable cause" of the property (both implemented formally using possible worlds).

For each of their three experiments, Prasada and Dillingham find a statistically significant effect of being a k-property, as judged intuitively by the experiments in advance of the experiment, and the acceptability of the relevant sentence, demonstrating that these tests can distinguish the two kinds of property. They argue for a categorical rather than gradient

distinction between k-properties and t-properties based both on the data pattern and a need for a coherent theory. They evaluate 90 properties, evenly split between properties judged to likely be k-properties and ones predicted to be t-properties, and evenly split between natural, artifact and social kinds.

We can use Prasada and Dillingham's list of experimental items to test whether *fake* can target precisely *k*-properties, by constructing a context for each property that describes something that is an *N* in all respects except that it lacks that property. If *fake N* can be used felicitously to describe this object, then *fake* can target that property. Likewise, if *not a real N* is felicitous, then *real* is sensitive to the same. (19) and (20) show a pair of examples for *trampoline*, with Prasada and Dillingham bare plural "by virtue of" rating in parentheses.

(19) Target property: Trampolines are bouncy. (*k-property, rating 5.94/7*)
Context: Mary and her son Charlie are trying out everything in the trampoline section at a large toy store. One of the things there says it's a trampoline, but when they test it, they find it isn't bouncy.

a. That's not a real trampoline. Real trampolines are bouncy.

b. That's a fake trampoline.

c. #That's a trampoline.

(20) Target property: Trampolines are black. (*t-property, rating 1.94/7*)
Context: Mary and her son Charlie are trying out everything in the trampoline section at a large toy store. One of the things there says it's a trampoline, but when they look at it, they find that it isn't black. Charlie tests it and finds that it still works well.

a. #That's not a real trampoline. Real trampolines are black.

b. #That's a fake trampoline.

c. That's a trampoline.

In this case, *fake* and *real* indeed target the k-property *bouncy* and not the t-property *black*. Here, bounciness appears to be so core to trampolines that negating it causes the object to no longer be a trampoline. However, this is not true of all properties targetable by *fake*, as we know from the many subsective judgments in Experiments 2 and 3. While many k-properties in Prasada and Dillingham have a core or definitional quality, (21) is an example of one that does not, which *fake* can indeed target:

(21) Target property: Lifeguards can swim. (*k-property, rating 6.33/7*)
Context: Adam wants to be a lifeguard at his local pool because he wants to make friends with a girl who often swims there. Adam can't actually swim, but he thinks he can still be a good lifeguard as long as he can throw the life ring to people. He fakes a swim test certificate and is able to get the job.

a. Adam is not a real lifeguard. Real lifeguards can swim.

b. Adam is a fake lifeguard.

c. Adam is a lifeguard. (*in the sense of having the job, despite not being qualified for it*)

However, not all properties labeled as principled connections (k-properties) by Prasada and Dillingham are equally valid targets for *fake*. (22) illustrates one such case:

(22) Target property: Eggshells are fragile. (*k-property, rating 6.33/7*)
Context: Peter is taking part in his university's egg drop competition, where eggs are dropped from a height and students have to design a contraption to protect them. He learns about a genetically engineered species of chicken whose eggshells are much less fragile and can survive falls from heights of several feet, designed to reduce breakages during packing and transport. They otherwise look identical to normal eggs. He smuggles in an egg with this tough eggshell into the competition and uses it when it's time to test his invention.

a. #That's not a real eggshell. Real eggshells are fragile.

b. #That's a fake eggshell.

c. That's a fake fragile eggshell.

d. That's an eggshell.

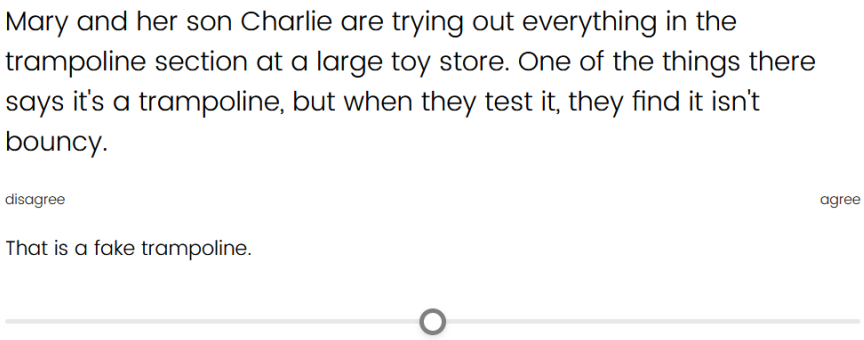
Experiment 4 explores the relationship between the properties tested in Prasada and Dillingham's experiment and the acceptability of *fake* and (*not*) *real*. Specifically, Experiment 4 tests the acceptability of *fake* and *not real* in contexts where a particular property is negated, using k-properties and t-properties from Prasada and Dillingham (2006) as well as some contexts from Experiment 3. We expect *fake* and *not real* to be more acceptable in contexts with k-properties than with t-properties, and we further expect to see some distinctions between some of the items that Prasada and Dillingham classified (before collecting their data) as k-properties, since perhaps some of these so-called k-properties are not actually connected to the noun concept in a sufficiently principled way.

2.7.2 Method

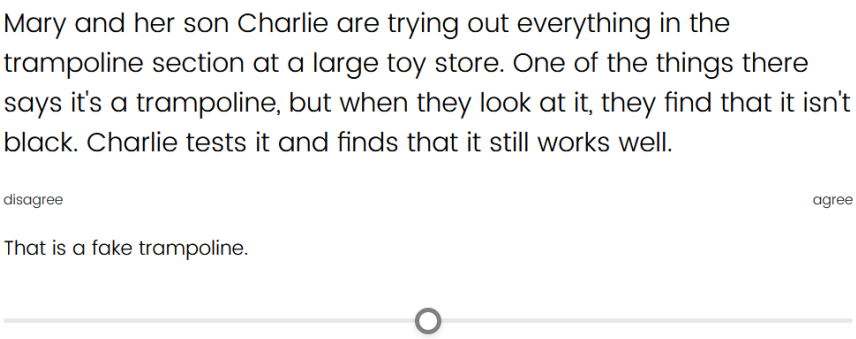
Experiment 4a asks participants to evaluate statements like *This is a fake trampoline* given a context describing a trampoline-like object with the target property negated, on a 100-point scale from "disagree" to "agree". Two example questions are shown in Figure 2.8. Experiment 4b repeats the same experiment but studies (*not*) *real* instead of *fake* and asks people to evaluate *This is not a real trampoline* for the exact same contexts.

We tested 42 noun/property/context triples covering a total of 24 nouns, balanced by expected acceptability of fake. We use 19 nouns from Prasada and Dillingham (2006), associated with 24 total k/t-properties, as well as 3 nouns from our context experiment and *gun* (since it is so common in the literature).[10] We focus on nouns in Prasada and Dillingham's data which are associated with both a k-property and a t-property, and for which it was possible to write a context where *fake* targeted that property. For 9 nouns which were only associated with a k-property, we added a second property that was either (a) a t-property (such as color, modeled after typical t-properties in their data) if the original property seemed to

[10]For *gun*, we specifically include the *fake metal gun* context (17) from Martin (2022) that we will see in Section 3.5.2 when discussing (unconstrained) context-sensitivity.



(**a**) *Negated k-property: trampolines are bouncy*



(**b**) *Negated t-property: trampolines are black*

Figure 2.8: *Screenshots of questions in Experiment 4*

be a true k-property, or (b) a more likely candidate for a k-property if not. This resulted in 20 items with k-properties where we expected *fake* to be acceptable, 10 items with "k-properties" which we suspected might not be as core and might result in lower ratings (see the discussion of finer-grained property types below), and 10 items with t-properties where we expected *fake* to be unacceptable, plus a further 2 items which tested contexts used in Chapter 3 (contexts (50) and (39); these test other aspects of the acceptability of *fake* besides what noun properties it can target).

We recruited 22 native American English speakers on Prolific for Experiment 4a and 20 native speakers for Experiment 4b, using the same criteria as in Experiment 2 and Experiment 3. We excluded 1 participant from Experiment 4a for not meeting the native speaker criteria and 1 participant for having an anomalous overall rating distribution (suggesting the participant did not do the task as intended), resulting in 20 participants per experiment. Each participant saw all 42 questions, divided into 3 blocks such that no noun was present twice in the same block. The block order was randomized and questions were ordered randomly within the blocks. We provided 3 training examples using other adjectives to demonstrate the use of the scale.

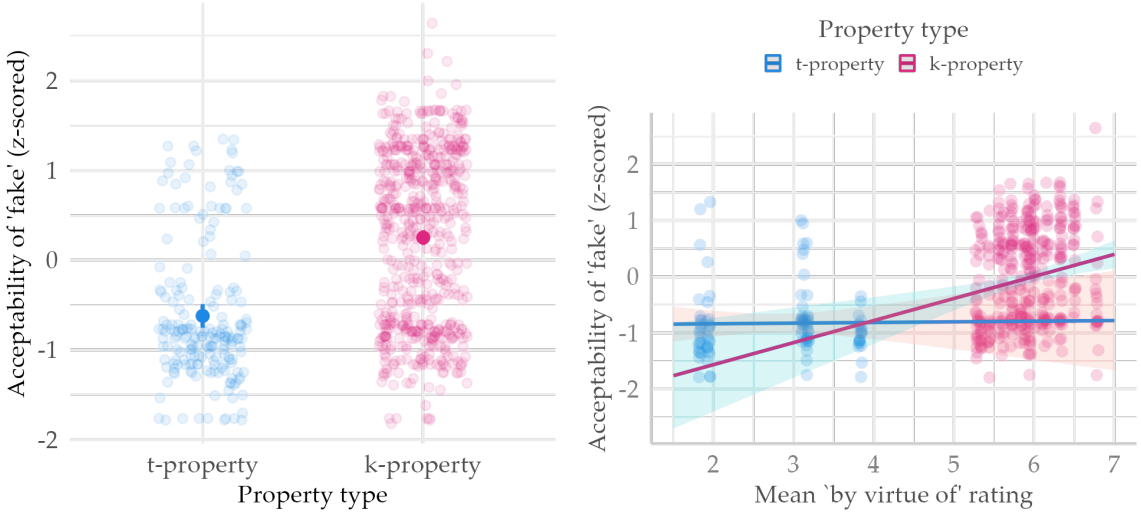
2.7.3 Results

Initial results for *fake*

We z-score all ratings to account for variation in how participants interpret the scale as well as variation in participants' thresholds to describe something as *fake* or *not real*. We exclude one item from the analysis (*scissors cut*) as the context did not actually sufficiently support *fake*,[11] as well as the 2 items which tested other aspects of *fake*. We confirm the initial judgments given for the examples in (19)–(22), with at least 75% of participants aligning with the expected judgments (z-score < 0 if marked #, z-score > 0 if marked acceptable). We then fit an initial linear mixed effects model that predicts rating for *fake* as a function of the prop-

[11]The context used an example of old and thus no longer sharp scissors; old and/or broken objects appear not to be targetable by *fake* no matter how much deception is involved. More research is needed here.

erty type as labeled by Prasada and Dillingham, with block presentation order as a random effect.[12] We find a significant effect of property type, namely that k-properties are rated 0.44 standard deviations higher than t-properties ($p < 0.001$), as shown in Figure 2.9a. We also see significant variation among k-properties, with a standard deviation (of ratings within a property) of 0.74 on average compared to 0.67 for t-properties.



(**a**) *Effects plot for rating by property type*

(**b**) *Effects plot by property type and P&D rating*

Figure 2.9: *Effects plots predicting the acceptability of fake based on the a priori property types and bare plural generic + "by virtue of" ratings in Prasada and Dillingham (2006).*

This variation is partially but not fully explained by the ratings obtained by Prasada and Dillingham. We use ratings from their Experiment 1 because it uses the most linguistic of the three tests, bare plural generics with *by virtue of*. Adding the ratings (shown in Figure 2.9b), we see a significant effect of rating ($p < 0.05$) and a significant interaction ($p < 0.05$); the effect of type is still significant ($p < 0.05$). This yields a marginal $R^2 = 0.164$, compared to $R^2 = 0.147$ with just the property type, which is an improvement but still leaves much variance unexplained.

[12]Since we are using z-scores, we do not use a random effect for participant.

Some of this variation is simply due to participants' differing standards and the quality of the context. (*Fake* also requires an element of deception, as discussed in Chapter 3, and different contexts implement this to presumably varying degrees of success.)

Results with finer-grained property types

In fact, dividing Prasada and Dillingham's "k-properties" into subclasses explains a good deal of the variation. By careful introspection about the nouns and their properties, we can divide them into the following three classes: (1) "essential" k-properties, which are privative if negated, (2) "non-essential" k-properties which hold of any prototypical instance but may be negated (yield a subsective inference), and (3) "borderline / statistical" so-called k-properties which do not actually need to hold of even prototypical instances, though they typically do, but are often highly statistically correlated with some other property which is actually a k-property. An example of an essential k-property is *trampolines are bouncy*, as discussed above. A clean example of a non-essential k-property is *snow forms in clouds*[13] (easier judged in the singular as *a snowflake forms in a cloud*): the indefinite singular generic seems acceptable, and certainly snowflakes typically form in clouds, but we are willing to call snow/snowflakes generated by a snow machine *snow* (and also *fake snow*, as we show in this experiment). *Lifeguards can swim*, as seen in (21), was also judged to be in this category. An example of a borderline/statistical 'k-property' is *eggshells are fragile*: it does not seem necessary for even prototypical eggshells to be fragile, but rather this seems statistically correlated with the fact that baby creatures have to break out of their shells; this is easier if they are fragile. (A baby dinosaur with excellent beak strength need not have a fragile egg by objective measures.) This subdivision was done by the author before seeing the results of this experiment, in order to explain examples like the non-fragile eggshell discussed in (22). A follow-up experiment which classifies these properties using participant judgments

[13]This property is not included in Prasada and Dillingham (2006), but was added to supplement their k-property *snow is white*, which was not obviously a k-property.

is a clear next step, providing a suitable design can be found which has participants give the properties a sufficient amount of careful thought.

We fit a revised linear mixed effects model which predicts the z-scored acceptability of *fake* as a function of these four property types: essential k-property, non-essential k-property, borderline/statistical k-property and t-property. We use Helmert coding to compare each property type to the combination of lower-ranked types. We find a significant effect of each type of k-property ($p < 0.001$ in all cases). This also explains a much higher proportion of the variance, with a new marginal $R^2 = 0.345$, much higher than the previous model at (marginal) $R^2 = 0.164$. There is still a fair deal of variation, but we see much clearer clusters in Figure 2.10; the remainder can likely be attributed to participant variation and degree of context effectiveness.

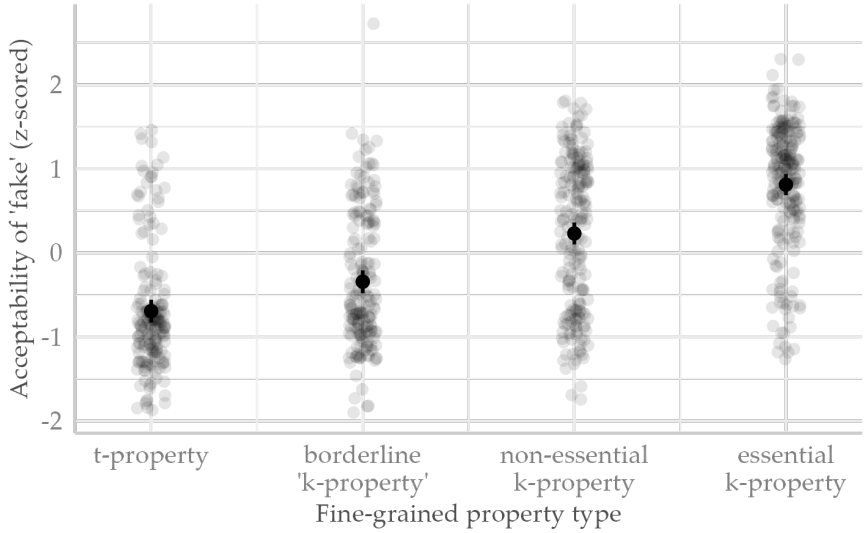


Figure 2.10: *Effects plot for rating by property type*

Results for (*not*) *real*

We fit the same linear mixed effects models for (*not*) *real* and find that all the same effects are significant. Moreover, a linear regression that predicts the mean z-scored rating of *not real* from the mean z-scored rating of *fake* for each property finds a significant, high correlation:

(adjusted) $R^2 = 0.821$ ($p < 0.001$), shown in Figure 2.11. This suggests that *fake* and *real* do indeed target the same kind of property.

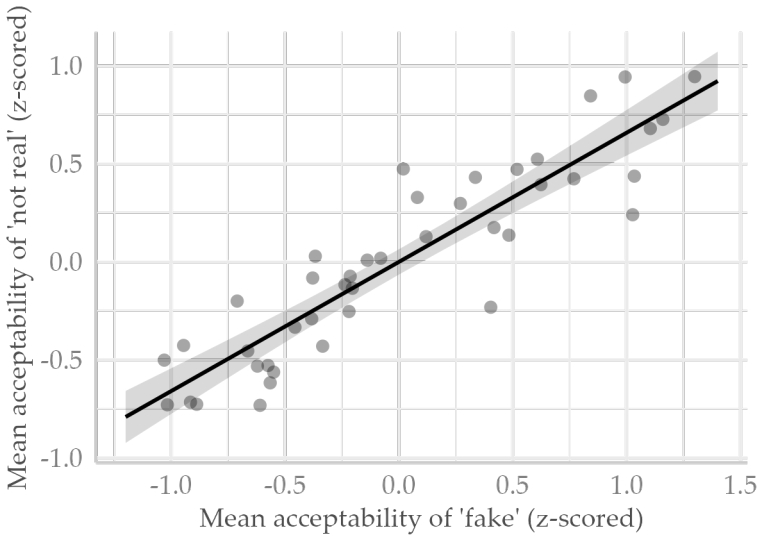


Figure 2.11: *Effects plot for the correlation of fake and (not) real*

2.7.4 Discussion

We observe a ranking of properties associated with concepts, from most core to least core (each category includes the ones above it):

1. "Essential" properties that, when they do not hold, always result in a privative inference (such as *bouncy* for *trampoline*).
2. Properties that can be targeted by *fake* and *real*.
3. Properties that support generics with singular kinds, *by virtue of* statements in generics with bare plurals and explanations (of the sort in Prasada and Dillingham (2006)).
4. Properties that support generics with bare plurals, such as *trampolines are black*.

We would like to collapse the distinction between 2. and 3. and collapse the peculiar sensitivity of *fake* and *real* to "core" properties with the notion of "core" that powers k-properties and explanation in Prasada and Dillingham (2006). However, we found a number of prop-

erties that Prasada and Dillingham labeled as k-properties / principled connections that *fake* cannot target. Recall that the original labeling in their experiment was based simply off the experimenter's native speaker judgments, in order to be able to perform a statistical analysis between the a priori label and the participant ratings. Given (1) our findings, (2) the actually somewhat noisy judgments obtained by Prasada and Dillingham, and (3) the closer examination of individual examples like *eggshells are fragile*, which suggests that that particular so-called k-property might actually be more statistical, it seems that the threshold for a k-property is actually higher than where Prasada and Dillingham originally drew it, and that *fake* does indeed target k-properties.

Let us thus assume that what *fake* targets are k-properties, and that some, but not all, k-properties lead to loss of type membership, i.e. privativity, when they are negated. One such example is *trampolines are bouncy*, while *dogs are four-legged* is not in this category: three-legged dogs are still dogs. This explains why that *fake* is sometimes privative and sometimes not, as observed in our data. If we further assume that anything that leads to loss of type membership when negated is a k-property, then this makes an interesting prediction: since *fake* can target any k-property, then *fake* should be felicitous as a substitute for any adjective that yields privativity in a certain bigram. We should be able to use *fake currency* to describe *illegal currency* and *fake €50 note* to describe an *red €50 note* (recall that €50 notes must be orange).[14] While proper experimental evidence is needed, intuitively, the acceptability of *fake* in these contexts does seem to hold, which is encouraging.

I believe mapping *fake* to k-properties to be the most intuitive way to capture the patterns we see here. More broadly, the individual items in Experiment 4 provide a catalog of 42 properties that *fake* can and cannot target, in addition to the items in Experiment 3 which *fake* can presumably target (though we do not have explicit acceptability ratings there). These examples range from definitional-seeming properties like *trampolines are bouncy* to much less obvious ones, such as *plans are carried out* or the fake concert example (14) from Experiment 3,

[14]This prediction is also made by an account that just stipulates that *fake* is privative, but since we are trying to move away from that, this provides an interesting test.

which we also included in Experiment 4 and does not even have an obvious *concerts are ...* paraphrase. These items provide a useful set of test cases for any theory that wishes to account for the behavior of *fake*.

Finally, previous experiments on *real* and *true* (Knobe, Prasada, and Newman, 2013) show the ability of these adjectives to target *dual character concepts* such as *friend* or *rock music*, which display both concrete and abstract (often normative) characteristics. We potentially see an effect of dual character in the subsective example for *fake lifeguard*, which distinguishes between having the job and being able to carry it out well/as intended. Our experiment shows that *real* can also apply to other nouns that are not regarded as having a dual character by virtue of targeting their k-properties. While we might gain some traction on *real* and dual character concepts by positing that these abstract, normative properties are all k-properties, it seems unlikely that the concrete properties can be taken as t-properties (meaning that their presence would not be sufficient for the felicity of *real*). Some concrete properties such as professors being people who are appointed by the university and carrying out research seem to be k-properties, as judged by the singular indefinite test *a professor is appointed by the university / a professor carries out research*, even if that professor is only motivated by money and not scientific advancement, and thus not a *true professor* in some sense. Instead, it seems like *real* and *fake* have dual character-associated readings when they happen to enforce or negate abstract/normative k-properties, and not otherwise. Meanwhile, *true* may be further restricted to only target abstract/normative k-properties, yielding this association with dual character concepts that possess such k-properties.

2.8 Generalizations about privative inferences

As we have seen in the experiments in this chapter, privative inferences can arise not just with traditional "privative" adjectives like *artificial*, *counterfeit* and *fake*, but also with material adjectives like *stone* and even color and shape adjectives like *red*, *hexagonal* and *tiny* which are typically classed as intersective. Further, these inferences only arise when these adjectives are combined with certain nouns (and sometimes only in certain contexts with that noun).

I propose that adjective-noun modification and the resulting inferences are governed by the following three generalizations. In Chapter 3, I will then lay out a theoretical account that implements these generalizations.

(G1) **Generalization 1:** Privative inferences occur when the adjective modifies properties that are "crucial" to noun-hood in the given context.

Adjectives modify a certain property or set of properties, such as being made of stone (or, for *counterfeit*, not having the brand that it claims to be have). When the adjective composes with the noun, it takes in some concept[15] and returns a new concept which has different properties to the original. When we then evaluate the truth conditions of "a/this A N is an N", we compare this new concept against the criteria that the predicate N supplies for membership. Privativity arises when the new concept fails these criteria.[16] Thus, *counterfeit money* is not money because being made by the relevant treasury is crucial to being money, while a *counterfeit watch* is still a watch because the brand/method of manufacture does not prevent it from functioning as a watch. Likewise, a *stone lion* is not a lion in most contexts because most contexts require a lion to be alive and breathing.

(C1) **Corollary 1:** Any adjective can in principle yield a privative inference, if there is a noun such that the property value that the adjective sets is antithetical to the meaning of the noun.

Item (C1) captures the insights from Experiment 2 and Section 2.6 that all kinds of adjectives from *fake* to *red* can yield privative inferences with the right noun.

---

[15] I will use 'concept' loosely here; formally, this will be modeled as an intension of which certain properties are true or false in w.

[16] It falls out from this two-stage process (first compose, then draw inferences) that in many cases, humans may not actually bother to draw the inference at all, leaving it undetermined if the discourse context does not require it. Our experiment setup in Chapter 2 forces people to draw the inference.

In practice, nonsubsective adjectives like *alleged* and *possible* are likely to be mostly immune from this generalization. Since the properties they contribute relate to speech acts or are modalized, it is unlikely that a language contains a noun or noun phrase that one might reasonably want to express that would contradict them.[17]

(C2) **Corollary 2:** It is not fruitful to distinguish a class of "privative adjectives" in the semantic theory, since the behavior of these adjectives (except possibly *fake*) can be explained using the same mechanisms as adjectives like *stone* which are not typically considered privative. (The only remaining use of the term is to group adjectives which have higher-than-usual likelihood of privative inferences.)

(C3) **Corollary 3:** The problems we need to solve for composition are (a) how to retain the properties that the adjective does not modify (to be *faithful* to the noun meaning during composition, as discussed in Chapter 1) and (b) how not to yield a contradiction between new properties introduced by the adjective and existing properties of the noun (which happens in cases of privative inferences).

We will discuss this last corollary more in Section 3.1, where we discuss our options for theoretical accounts. Subsective lexical entries are one, but not the only, way of solving problem (a) in Corollary (C3).

(G2) **Generalization 2:** The effect of context on privativity inferences can primarily be explained by the discourse providing extra information that is not specified by the adjective. There is no need for context-sensitivity in the adjective meaning to explain variation in privative inferences, except for *fake* (see (G3)).

[17]There may be a reading of *possible contradiction* or *possible impossibility* that is privative, but the primary reading is that we are not sure whether it is a contradiction / impossibility or not, not that it is a contradiction / impossibility that can possibly occur.

For some nouns like *spring water*, labeling something as *counterfeit spring water* (as in Experiment 3) does not determine whether a privative inference is warranted or not. Whether it is made of spring water or not is underspecified; it depends on how it was counterfeited. The fact that the additional contexts provided in Experiment 3 can shift the inference in these cases is just that: the context updates the common ground with additional information about which properties were modified—in this case, whether the counterfeiting involved tap water or just spring water from the wrong spring—filling in the underspecified information and determining the inference. The meaning of *counterfeit* remains the same in either kind of context (as it remains the same regardless of the noun). The mechanisms behind this "context-sensitivity" are fundamentally no different from the fact that something like *plastic knife* does not determine whether the knife has other properties like being *sharp*: while plastic utensils are often made of a cheap plastic that is not particularly sharp, it could be made of a more advanced plastic that holds an edge as well as metal. Whether or not a *plastic knife* is made of high enough quality plastic to be sharp is underspecified, just as it is underspecified how the spring water was counterfeited. Of course, a privative inference is more striking than how sharp a knife is, and may give us more pause about how to correctly implement adjective modification, but the mechanisms in our theory need be no different.

Of all the "privative" adjectives discussed in this dissertation, only *fake* shows any effect of context-sensitivity in its meaning. This context-sensitivity is still quite limited, as we showed in Experiment 4 that it can only target k-properties of the noun and cannot target other salient properties like being made of metal in (17).

(G3) **Generalization 3:** *Fake* and *real* target properties that are "core" to the noun – and is not sensitive to anything else, no matter how contextually salient. These properties are the same properties that indefinite singular kinds target in generic sentences. Context determines which property *fake* and *real* target.

In the next chapter, we will see how to implement these insights in a new theoretical account of adjective-noun modification.

Chapter 3

"Responsible" adjective-noun modification

3.1 Choices for adjective modification

Generalization (G1) and its corollaries suggest that we want a unified account of adjective-noun modification that does not distinguish between so-called privative adjectives and other adjectives. Privativity, which is equivalent to there being an (apparent) contradiction between some property required by the adjective and some property required by the noun, should be handled as and when it arises. Corollary (C3) spells out the specific issues our theory needs to handle: (a) how to retain the properties that the adjective does not modify (be faithful to the noun meaning) and (b) how not to yield a contradiction between new properties introduced by the adjective and existing properties of the noun (which happens in cases of privative inferences).

One way to solve (a) is of course to subsect the noun, using either predicate modification or an adjective lexical entry that involves $\lambda N.\lambda x.\ \ldots \wedge N(x)$. However, this is not the only option. The lexical entry for *counterfeit* in del Pinal (2018), which uses dual content semantics, solves (a) by describing the relationship of a *counterfeit* N to N in terms of the telic and perceptual dimensions of N (only; the origin dimension A is notably omitted). Under this view,

counterfeit is nonsubsective and does not run into contradictions, even when the inference is privative.[1] This shows that there are two broad shapes a theoretical account that implements (G1) can take:

(A1) Use an overtly intersective or subsective lexical entry (of the form $\lambda N.\lambda x.\ ... \wedge N(x)$) for many adjectives. Then use a repair mechanism to resolve contradictions with the noun when they arise. Subsectivity occurs by default for these adjectives; privativity arises contingently, in the same instances where the repair mechanism kicks in. A few adjectives, such as *alleged*, remain nonsubsective.

(A2) All adjective lexical entries appear nonsubsective. Each adjective is responsible for retaining enough properties of the noun to get the correct meaning, while not generating any contradictions. Subsectivity or privativity arise contingently for some adjectives depending on the noun; others, like *alleged*, are genuinely nonsubsective.

A well-known account of type (A1) is Partee (2010), which invokes context to widen the noun meaning in order to allow subsection to proceed. As discussed in Chapter 2, this widening approach can in principle account for all our data, as well as ability for all adjectives to yield privative inferences. However, there are two issues with it. Firstly, type (A1) accounts capture the intuition that most adjectives really are subsective most of the time and that privativity is relatively rare by treating privativity as an exception.[2] This comes at the expense of capturing Corollary (C1) well: it may seem like a contradiction to say that (almost) all adjectives can be privative and also that all such adjectives are subsective by default. The second issue is broader, and involves how the account invokes context. The burden of explaining when we get a subsective vs. privative inference is pushed almost entirely onto the context, plus whatever is contained in the opaque (atomic?) lexical entries of adjectives like *fake*. As we discussed in Chapter 2, this handles but does not explain all of our inference

[1] In del Pinal (2018), *counterfeit* is actually always privative by construction of the lexical entry, but this negation component is not required, as del Pinal mentions.

[2] Of course, frequency need not be reflective of underlying meaning: a prime case are scalar implicatures, where the strong meaning is very frequent but we take the underlying meaning to be the weak one.

variation both between bigrams (Generlization (G1)) and within a given bigram (Generalization (G2)). Further, the way in which the context can act on the noun is not constrained. It is assumed that the context widens *gun* in *fake gun* to include just those not-guns which are *fake guns*, and not also e.g. *giraffes*, or, more importantly, *toy guns*, such that *fake gun* would be able to refer to a *fake toy gun* which is actually a perfectly functional *gun* (del Pinal, 2018). We will discuss issues with context-sensitive theories of this sort in Section 3.5.1 as part of the setup for including context in our own theory.

It is possible to fix the overgeneration issues with these context-invoking theories by appropriately constraining the role of the context, especially for the noun widening mechanism, which is already tied to the noun and can thus be constrained via the noun. Section A.6 illustrates how this can be done using the constrained contextual modulation function introduced by del Pinal (2018). However, appealing to context at all to resolve privativity seems unnecessary. It is not the wider discourse context that causes *stone lion* or *counterfeit money* to be privative, but simply the fact that *stone* combines with *lion*, or *counterfeit* with *money*. Moreover, as we just discussed, handling privativity as an exception that gets resolved by contextual widening does not capture the intuitions of Generalization (G1) and Corollary (C1) that adjective modification is simply about modifying whatever properties the adjective targets, and that any adjective can in principle yield a privative inference if it targets properties particularly "core" to a noun. Per Generalization (G2), context does have an influence on inferences, but only when the adjective-noun combination underspecifies the relevant properties.

Two other accounts of type (A1) are Hogeweg (2012) and Oliver (2014), which assume that adjective-noun modification is intersective and then provide optimality theory (OT) accounts of semantics paired with rich representations of nouns. (Oliver (2014) also appeals to qualia (Pustejovsky, 1991) to structure the representation of nouns, just like del Pinal (2015, 2018).) The OT component provides a repair mechanism: it resolves contradictions by dropping contradictory parts of the noun meaning, privileging the adjective meaning. These accounts capture Generalization (G1) more directly: privativity arises when a contradiction

occurs between the noun and adjective meaning (Oliver, 2014), though adjective meanings are still intersective by default.

I will pursue an account of adjective modification that is similar in spirit in that it involves a system of contradiction resolution, but I will use a simple notion of entailment which can be built directly into the adjective lexical entry, rather than appealing to a separate constraint or meta-constraint ranking.

3.2 Dual content semantics with k-properties and genericity

The foundation for our account will be a dual content semantics, building on del Pinal (2018), since this allows us to directly model the close link between concepts and noun meanings in our semantics. In a dual content semantics, lexical entries have an E-structure which represents their extension and also a C-structure which "describes a set of beliefs about the extension [and] encodes what, according to our best psychological theories, are the basic components of the corresponding kinds of concepts" (del Pinal, 2018, p.175). In his examples, del Pinal uses the four qualia from Pustejovsky (1991)—CONSTITUTIVE, PERCEPTUAL, TELIC and AGENTIVE—to illustrate the dimensions of natural and artifact kinds (*lion* and *gun* respectively), as shown in (23) and (24):

(23) Lexical entry for *gun* in del Pinal (2018):

$[\![\text{gun}]\!]^c_M$ =

E-structure: $\lambda x.\ \textsc{gun}(x)$

C-structure:

CONSTITUTIVE: $\lambda x.\ \textsc{parts-gun}(x)$

PERCEPTUAL: $\lambda x.\ \textsc{perceptual-gun}(x)$

TELIC: $\lambda x.\ \textsc{gen}\, e\, [\textsc{shooting}(e) \wedge \textsc{instrument}(e, x)]$

AGENTIVE: $\lambda x.\ \exists e_1\, [\textsc{making}_w(e_1) \wedge \textsc{goal}(e_1, \textsc{gen}\, e[\textsc{shooting}(e) \wedge \textsc{instrument}(e, x)])]$

(24) Lexical entry for *lion* in del Pinal (2018):

$[\![\text{lion}]\!]^c_M$ =

E-structure: $\lambda x.\ \textsc{lion}_w(x)$

C-structure:

CONSTITUTIVE: $\lambda x.\ \textsc{substance-lion}_w(x)$

PERCEPTUAL: $\lambda x.\ \textsc{perceptual-lion}_w(x)$

TELIC:

AGENTIVE: $\lambda x.\ \exists e_1\ [\textsc{biological-birth-lion}_w(e_1, x)]$

These lexical entries are not intended to give a full definition of *gun* or *lion* but rather to enable his derivations; del Pinal notes that different classes of term may involve different dimensions. Our first step will be to replace qualia with a more recent psychological theory of concepts, in the spirit of del Pinal's original description of C-structure.

3.2.1 Replacing dimensions with properties

One recent "psychological theory that encodes components of concepts" is Prasada and Dillingham (2006), which we discussed in reference to the kinds of property that *fake* can target in Section 2.7 in Chapter 2. Recall that in Prasada and Dillingham (2006), each concept is associated with a collection of k-properties, which it has principled connections to, as well as a collection of t-properties, which it merely has statistical connections to.We will take the C-structure of a noun to contain k-properties and t-properties instead of a fixed set of four dimensions. This means that every noun has its own variably long list of properties[3]; it is no longer the case that every noun has a perceptual property (labeled as such), for example, though perceptible nouns will have properties relating to their perceptual aspects.[4,5]

[3]We assume this list to be finite or at worst countably infinite.

[4]I will assume that humans are able to classify a given property of a noun as being e.g. a perceptual or origin or authenticity sort of property; we will leverage this in the lexical entry for *counterfeit*. I will not assume that the properties are labeled a priori in this way.

[5]This does not commit us to a particular way of representing noun concepts, but one way to consider it, following Prasada and Dillingham (2009), is to view k-properties as a structured representation of links between concepts and other concepts.

We posit that the k-properties are identified as such and that there is an operator $\mathbb{P}_k$ that can retrieve just the k-properties from the C-structure and ignore the t-properties. We will use $\mathbb{P}$ to refer to the total list of properties.

Further, we will assume that the C-structure properties are ranked by centrality, following del Pinal (2018). In particular, all k-properties are ranked above all t-properties.[6] The motivation for a ranking within k-properties will become clear when we discuss *fake* and *real* in Section 3.4, which we will argue preferentially target higher-ranked (more central) k-properties.

The last piece we need is a new notation for the C-structure: while it is still a list (now of properties instead of dimensions), I will denote it as a sum $\sum_{P \in \mathbb{P}}$ for ease of operating over it in our lexical entries. The + operator simply adds items to the list. I will assume that $Q + \sum_{P \in \mathbb{P}}$ adds a new property to the beginning of the list, i.e. Q is ranked above the existing properties. C-structure properties of nouns will be of type $\langle s, \langle e, t \rangle \rangle$[7], but in general the type of a C-structure property of an expression will be the same as the type of the expression, however complex. Note that we are using an intensional semantics with a world variable w of type s here, so E-structure (Extensional structure) becomes something of a misnomer.

Since our list of properties is now much longer than del Pinal's original four dimensions, it is not possible to give a realistic lexical entry for *lion* anymore. However, let us assume that at minimum, the k-properties of *lion* include *lions are born from lions*, *lions have fur*, *lions have manes*,[8] *lions are lion-shaped* and *lions are mammals*. The ranking of these properties is an empirical matter which I will set aside in the case of lions.

---

[6]This feature of C-structure, as well as its general structure as a list, more closely resembles a prototype theory of concepts; the k-properties we are using are situated by Prasada and Dillingham (2006) as knowledge theories or "theory theories" of concepts (see Murphy (2002) i.a. for an overview of theories of concepts). We need our properties to be *at minimum* accessible as a list; whether they in fact have a richer structure can be set aside for our purposes.

[7]This may be an oversimplification if properties contain e.g. denotations of gradable adjectives that require a context parameter, such as to encode the k-property *basketball players are tall*; this can be adjusted if needed.

[8]Is is a well-known fact of the kind literature (Chierchia, 1995) that *a lion has a mane* is felicitous even though female lions, who form a large part of the lion population, do not have manes.

(25) New lexical entry for *lion* with incomplete C-structure

$[\![\text{lion}]\!]^w =$

E-structure: $\lambda x.\ \textsc{lion}_w(x)$

C-structure: $(\lambda x.\ \textsc{born-from-lion}_w(x)) + (\lambda x.\ \textsc{furry}_w(x)) + (\lambda x.\ \textsc{has-mane}_w(x)) +$
$(\lambda x.\ \textsc{lion-shaped}_w(x)) + (\lambda x.\ \textsc{mammal}_w(x)) + \ldots$

Lions are lion-shaped is perhaps a tautology, but consider this an abbreviation of a longer description of the shape of a lion, which surely forms part of the concept of lion even if it is not easily expressed in language. *Lions are born from lions* is more problematically circular, but this seems a necessary part of the concept, however this circularity actually gets resolved in our minds. Lastly, note that last k-property links to the concept of *mammal*, which will cause the C-structure of *lion* to entail all sorts of mammal-properties, even if those are not themselves k-properties of *lions*. (Prasada and Dillingham (2006) argue that k-properties are not directly inherited: having four legs is a k-property of a dog but not of a dachshund.)

3.2.2 Compositional C-structure

As in del Pinal (2018), we will assume that C-structure is included in our compositional semantics: composite phrases also have a C-structure. In order to achieve this, lexical items with functional types will have a C-structure which specifies a function from an existing C-structure to a new one. In particular, adjectives will be treated as modifiers—functions from properties to properties of type $\langle s, \langle\langle s, \langle e, t\rangle\rangle, \langle e, t\rangle\rangle\rangle$—and so their C-structure will consist of a function that takes in a C-structure with $\langle s, \langle e, t\rangle\rangle$ type properties and outputs a C-structure with properties of the same type that incorporates the adjective meaning.[9] (Section 3.7.3 discusses how to handle predicate modification under this account.) This replaces del Pinal's formulation which used function application dimension-by-dimension; this no longer works

[9]For typing purposes, we will set aside the fact that most of these adjectives are also at least somewhat gradable (the way that color adjectives are somewhat gradable) and so might involve a degree or comparison class in their typing (Kennedy, 2007; Kennedy and McNally, 2005; Klein, 1980). (67) in Section A.1 gives an example where *artificial* is gradable.)

since we no longer have shared, labeled dimensions between lexical items. We will show how this function over the whole C-structure is implemented for adjectives in Section 3.3.

Formally, we can adapt the dual content semantics provided by del Pinal (2018) as follows. (His semantics also includes a contextual modulation operator mod_l which is omitted here; we'll return to it in Section 3.6.)

(26) Terminal node interpretation

If α is a terminal node, then $[\![\alpha]\!]^c_{M_E}$ is the E-structure of α as specified in the lexicon, relative to c, $[\![\alpha]\!]^c_{M_C}$ is the C-structure of α as specified in the lexicon, relative to c, and $[\![\alpha]\!]^c_M = \langle [\![\alpha]\!]^c_{M_E}, [\![\alpha]\!]^c_{M_C} \rangle$.

(27) Function application

If α is a branching node and $\{\beta, \gamma\}$ is the set of α's daughters then

a. $[\![\alpha]\!]^c_{M_E} = [\![\beta]\!]^{c_1}_{M_E}([\![\gamma]\!]^{c_2}_{M_E})$

b. $[\![\alpha]\!]^c_{M_C} = [\![\beta]\!]^{c_1}_{M_C}([\![\gamma]\!]^{c_2}_{M_C})$

c. $[\![\alpha]\!]^c_M = \langle [\![\alpha]\!]^c_{M_E}, [\![\alpha]\!]^c_{M_C} \rangle$

Once we have a composite C-structure, we depart from the notion that its properties are k-properties. K-properties are used to provide the C-structure for content words such as nouns and maybe verbs (see Section 3.7.4), but Prasada and Dillingham (2006) argue that composite phrases such as *white bear* do not have k-properties (though *polar bear* does). The C-structure for composite phrases becomes just a semantic tool. We need composite phrases to have C-structure since adjectives will invoke the C-structure of the phrase that they modify (Section 3.3) and we can apply adjective modification recursively (Section 3.3.3).

However, we will not assume that all words have a contentful C-structure, in line with del Pinal (2018). In particular, there is no obvious C-structure for function words.[10] It is an open question whether function words have no C-structure and act as a point of "closure", such

[10]This point becomes even more clear once we discuss contextual modulation in Section 3.6, where the C-structure is what permits the meaning of nouns (and by extension other phrases) to be restricted or widened in context, following del Pinal (2018). Function words should not be able to undergo such contextual modulation.

that all composite phrases containing function words also have no C-structure, or whether their C-structure is something like an identity operator. More research is needed to determine what kind of data would answer this question.

3.2.3 Linking E-structure and C-structure via genericity

Replacing del Pinal's dimensions with k-properties move also allows us to tie C-structure more explicitly to kinds and genericity, and spell out the link between E-structure and C-structure. Recall from Section 2.7 in Chapter 2 that k-properties are associated not just with "in virtue of" expressions but also with the generic sentences involving the singular indefinite kind: *a dog has four legs*. Generics with bare plural kind expressions, meanwhile, permit k-properties and also t-properties: *dogs have four legs*, *barns are red*, but # *a barn is red* (Greenberg, 2003; Prasada and Dillingham, 2006).

Firstly, I will argue that C-structure is what kind readings are derived from, although I will not pursue a specific implementation of this. Martin (2022) suggests that C-structure can be mapped onto the sortal concepts underlying kinds, as shown in (28), but how exactly singular indefinite kinds pick out only k-properties while bare plurals can also target t-properties needs to be worked out.

(28) Mapping of kinds to C-structure suggested by Martin (2022, p.219)

$\mathbf{DOG} = \lambda w.\imath x.\ *\textsc{dog}_w(x)$

where **DOG** is the kind for *dog* of type $\langle s, e\rangle$, $*$ is a pluralization operator (Link, 1983) and DOG is the internal function (sortal concept) that builds the plurality. Martin suggests that DOG is the part of the kind associated with the C-structure of *dog*.

Secondly and more importantly, this close relationship between C-structure and kinds / genericity allows us to specify the relationship between E-structure and C-structure without reducing the E-structure to C-structure (i.e. while preserving extensions/intensions as atomic). By Prasada and Dillingham's criterion, k-properties and t-properties are properties that hold in bare plural generic sentences; k-properties also hold for indefinite singular

generic sentences. We can use the semantics of (bare plural) generics to relate properties in the C-structure to the predicate in the E-structure.

The semantics of generics is governed by three components: (1) modal quantification over worlds (with a modal base and ordering source that is accommodated given the context), (2) quantificational force in these worlds, typically *every* unless specified otherwise, and (3) accommodation of felicity conditions for the generic to hold (Chierchia, 1995; Greenberg, 2003; Krifka et al., 1995). For example, we interpret *lions have manes* as "for every x and every w' in the (stereotypical) modal base B_w which is most normal/stereotypical according to the ordering source, if x is a lion in w' and the felicity conditions for having a mane (namely, being male) hold of x in w', then x has a mane in w'." This can be expressed using a genericity operator, as shown in (29), which contains the modal base, ordering source and appropriate quantifier (as quantificational adverb), including the quantificational variability associated with it.[11] Following Chierchia (1995), we write the conditions as restrictions on GEN instead of as a conditional.

(29) A lion has a mane.

$\textsc{Gen}\ x, w'\ [\textsc{lion}_{w'}(x) \wedge C(x, w')]\ \textsc{has-mane}_{w'}(x)$

where C represents the contextual felicity conditions derived from *lion* and *has a mane*.

While the modal base may vary between generics, we expect the sorts of generics that are k-properties to be evaluated with respect to a normal or stereotypical modal base, which means that the actual world w_0 is also expected to be a sufficiently "normal" member of this modal base. In other words, if the felicity conditions C hold of some x in w_0, then we expect the generic to hold of it in w_0 as well. Specifically, lions that meet the felicity conditions for lions having manes (namely being male, not being sick, injured or shaved, etc.) should indeed have manes in w_0, if *lions have manes* is true.

[11] I adopt the contextual felicity semantics from Chierchia (1995) but the world (as opposed to situation) semantics from Krifka et al. (1995). See Greenberg (2003) for a more detailed semantics that also incorporates contextual relevance.

We can now use this generic semantics to constrain the atomic predicate in our E-structure in terms of the properties in the C-structure.

(30) **Genericity Constraint on E-structure:**

For any expression e of type $\langle e, t\rangle$,

$\bigwedge_{\text{P}\in\mathbb{P}(e)}$ GEN x, w' $[M_E(e)(w')(x) \wedge C_{\text{P}}(x, w')]$ $\text{P}(w')(x)$

where $M_E(e)$ returns the E-structure of e and $\mathbb{P}(e)$ its C-structure properties[12].

In other words, in all sufficiently normal worlds w' in the modal base, for each property P of the C-structure (k-properties and t-properties), if x is in the intension of e and meets the felicity constraints C_{P} for that property, then P must be true of x.

Applied to *lion*, this constraint specifies that in all sufficiently normal worlds, including the actual world w_0, lions that meet the relevant conditions are lion-shaped, have manes etc. Although shaved lions and female, maneless lions exist in the actual world and fall under the extension of *lion*, these are excluded from the Genericity Constraint by the felicity conditions in the generic semantics. Importantly, the Genericity Constraint must hold not just for nouns, but also for composite phrases such as adjective-noun phrases, which also have a (composed) C-structure.This will allow us to do detailed composition using the C-structure and then have this constrain the E-structure of the composite phrase in a transparent fashion, even though the E-structure of the composite phrase is prima facie opaque.

3.3 Conflict resolution built into adjective modification

3.3.1 Responsible adjective modification

The key idea behind accounts of type (A2) is that each adjective must combine with the noun "responsibly", retaining as many properties of the noun as it can while not generating any contradictions. Widening theories that use set intersection treat the adjective and noun as equals, and have the composition mechanism and/or the context resolve any issues. This

[12]Strictly, $\mathbb{P}(e)$ obscures an application of the $M_C(e)$ operator to first retrieve the C-structure of e.

view takes an asymmetric view where the adjective, as modifier, bears the responsibility for carrying out the modification of the noun correctly.

It is possible that the way that adjectives contribute their meaning and retain the correct parts of the noun is highly adjective-specific and noun-specific. For example, the process of applying *stone* to a noun might involve reasoning like "if the noun is bowl, then keep the form and the function and just add that it's made of stone; if it's knife, leave the function and the sharpness aspect of the form underspecified as it depends on the type of stone; if it's a lion, only keep the form…". Under such a view, adjective lexical entries are best left as black boxes. However, it is more interesting to assume that the way that adjectives contribute meaning and resolve conflicts with the noun is via a shared mechanism. I will assume that all adjective lexical entries can be decomposed into two parts: a contribution of new meaning, which may or may not depend on the noun, and a noun meaning incorporation + conflict resolution process that takes in the noun and only returns the parts of its meaning that do not conflict the adjective meaning. Conflict resolution is accomplished by iterating over the noun's C-structure properties and keeping only those that do not result in a contradiction (whose negation is not entailed by the adjective meaning). A schema is shown in (31).

(31) **Schema for adjective lexical entries**

$[\![\text{A}]\!]^{w,c} =$

E-structure: $\lambda N.\lambda x.f_E(N)(c)(w)(x)$

C-structure: $\lambda N.\,(\lambda x.f_C(N)(c)(w)(x)) +$

$$\sum_{\text{P}\in\mathbb{P}(N)} \lambda x.(\neg(f_C(N)(c)(w)(x) \rightarrow \neg\text{P}_w(x)) \rightarrow \text{P}_w(x))$$

where w is the current world, c is the context, f_E and f_C are adjective-specific functions (see below), and N represents the noun that the adjective applies to.

f_E is an opaque function from the intension of the noun to the intension of the adjective-noun phrase, which may depend on the context c for some adjectives, such as gradable adjectives (Kennedy, 2007; Kennedy and McNally, 2005; Klein, 1980) and adjectives like *good* and *skillful* (Maienborn, 2020; Martin, 2022; Reinert, 2024 i.a.). As discussed in Section 3.2.3, we will

not attempt to decompose the E-structure of composite phrases any more than we typically decompose the extension/intension of content words such as *lion*. Instead, we will rely on the Genericity Constraint in (30) to constrain the E-structure according to the C-structure, where we do specify the exact effects of the composition.[13] f_C returns a (k-)property that contributes only the new meaning added by the adjective, and may or may not depend on N and c.[14] f_C will not depend on N if the adjective's meaning contribution can be expressed as a type $\langle e, t \rangle$ property, as we will see for *stone* in the next section, but will depend on N for nonsubsective adjectives as well as for *counterfeit*, *fake* and *real* (see Section 3.3.2 and Section 3.4). It will depend on the context c whenever f_E depends on the context, such as for gradable adjectives (and also *fake* and *real*).

The conflict mechanism as stated in (31) captures the "if there is no contradiction, keep the property" intuition described above. It is logically equivalent to the disjunctive formulation in (32): "either the property applies or there is a contradiction (so the property does not apply)".

(32) **Disjunctive schema for adjective lexical entries** (equivalent to (31))

$[\![\text{A}]\!]^{w,c} =$

E-structure: $\lambda N.\lambda x.f_E(N)(c)(w)(x)$

C-structure: $\lambda N.\,(\lambda x.f_C(N)(c)(w)(x)) +$

$$\sum_{\text{P} \in \mathbb{P}(N)} \lambda x.(\text{P}_w(x) \vee (f_C(N)(c)(w)(x) \rightarrow \neg \text{P}_w(x)))$$

where w is the current world, c is the context, f_E and f_C are adjective-specific functions (see below), and N represents the noun that the adjective applies to.

---

[13] Section A.4 discusses an alternative proposal which does decompose the E-structure, but this requires some complexity to quantify over properties in the E-structure while avoiding reducing the E-structure to the C-structure.

[14] Since we add this new property at the top of the list of properties ordered by centrality, it becomes a k-property of the bigram.

I will give this conflict resolution mechanism as part of the lexical entry of every adjective we discuss.[15] However, since the mechanism is shared between adjectives, it could also be housed in a shared location. A particularly compelling choice would be the Distributed Morphology categorizing head a, which combines with roots to create adjectives (Halle and Marantz, 1993). In this case, the adjective root would just have the tuple $\langle f_E, f_C \rangle$ as its meaning, and the head a would slot f_C into the conflict resolution mechanism. This could also hold the key to handling the difference between adjectives as modifiers versus adjectives in predicate position. Perhaps only adjective roots in modifier position combine with a modification a_M head that introduces this conflict resolution mechanism. The ones in predicate position could instead combine with a different a_P head that does not contribute this extra mechanism, and instead extracts f_C in the cases where it is of type $\langle s, \langle e, t \rangle \rangle$ (or a context-dependent equivalent type). We return to this in Section 3.7.3.

This conflict resolution process is similar to the Optimality Theory constraint resolution accounts proposed by Hogeweg (2012), Kluck (2007), and Oliver (2014) for *stone lion*, but happens directly inside the adjective meaning rather than as a separate component of the semantics, and need not posit a separate (meta-)constraint ranking. Unlike Hogeweg (2012) and Oliver (2014), who both assume that adjective meanings are of type $\langle e, t \rangle$ and that adjective-noun modification is fundamentally intersection, it also allows us to handle subsective adjectives whose meaning contribution f_C depends on the noun N, such as *fake* (which targets a k-property of the noun).

Splitting the adjective denotation into two parts in the C-structure—a core adjective meaning f_C and the process which combines it with the noun meaning—also allows us to address the faithfulness constraint on compositionality raised in Chapter 1. Recall that if we take composition to be function application, then the faithfulness constraint on composition that

[15]This does not mean that every adjective in the lexicon will use this schema, though most can. This mechanism in its current form is not appropriate for modal nonsubsective adjectives like *alleged*, which want to make no commitments (even in their C-structure) about noun-hood in w_0; commitments are made only in some modal worlds. Using entailment in w_0 derives no contradictions and thus incorrectly preserves all the noun k-properties in w_0; we actually want to remove them all. Further refinement on the use of entailment vs. some other more sophisticated operator is needed to capture nonsubsective adjective meanings with a schema like this.

it should be as faithful as possible to the "adjective meaning" and the "noun meaning" falls to whichever of the two is the function that takes the other as the argument, namely the adjective. With an atomic E-structure, we cannot distinguish a notion of "adjective meaning" from how it composes with the noun. In the C-structure, however, these things are split, and we can see clearly how faithfulness to the core adjective meaning and the noun meaning are implemented.

In addition to the Genericity Constraint, we can now assume a second constraint, a meaning postulate, which constrains the atomic predicates in the E-structure via the C-structure. As discussed in Chapter 1, our notion of composition includes that composition should be faithful to the meanings of parts. In particular, adjective-noun composition should be faithful to the meaning of the noun and preserve as much of it as possible. We implement this in the C-structure by keeping all the noun's properties except those that contradict. This meaning postulate constraints the otherwise opaque, atomic E-structure. In the case when there are no contradictions in the C-structure, then the E-structure should also preserve the full noun meaning: specifically, it should be *subsective*.

(33) **Subsective Meaning Postulate**

For any adjective following the schema in (31) with E-structure function f_E and C-structure function f_C, f_E is subsective for all nouns N ($f_E(N) \subseteq N$) unless there is some k-property $\textsc{p} \in \mathbb{P}_k(N)$ such that $f_C(N)$ contradicts it.

Formally, for any world w and context c,

$$\forall N\,[\,\forall \textsc{p} \in \mathbb{P}_k(N)\,\exists x\,[\textsc{p}(w)(x) \wedge f_C(N)(c)(w)(x)]]$$
$$\rightarrow \forall x\,[f_E(N)(c)(w)(x) \rightarrow N(c)(w)(x)]$$

For example, let N be *bowl*, $\textsc{p}$ be "bowls are bowl-shaped" (BOWL-SHAPED) and let f_C be the core meaning of *stone*, MADE-OF-STONE. (See Section 3.3.2 for more details on the lexical entries for *stone* and *bowl*.) Then if there exists some x which is bowl-shaped and also made of stone, then BOWL-SHAPED and MADE-OF-STONE are not in contradiction, and so (if that is true for all the other properties $\textsc{p}'$ as well, then) *stone* should be subsective for *bowl*.

We apply this meaning postulate to all adjectives treated in this dissertation, including *stone*, *counterfeit* and *fake*. This has the effect that they all yield subsective inferences "by default" (when there are no contradictions)—of course, for adjectives like *fake*, there will frequently be contradictions. This meaning postulate is a stipulative postulate, not an explanation of *why* adjectives like *stone* are subsective. It captures our intuitions that even though we need the E-structure predicates in nouns and adjectives to be atomic to handle edges cases, there is something we know about them, namely that they should be subsective in these cases. Section A.4 discusses an alternative account that proposes a non-atomic E-structure, at the expense of additional (potentially controversial) assumptions.

In principle, we could also specify an Intersective Meaning Postulate. For many adjectives, such as *stone* as we will see below, their core C-structure meaning f_C does not depend on N and is type $\langle s, \langle e, t\rangle\rangle$ (predicative), and so we get a fundamentally intersective-looking effect in the C-structure. If desirable, the parallel effect this should have on the E-structure can be formalized in a stricter meaning postulate.

(34) **Intersective/Subsective Meaning Postulate** (optional)
For any adjective following the schema in (31) with E-structure function f_E and C-structure function f_C,

a. If f_C is of type $\langle s, \langle e, t\rangle\rangle$ and does not depend on N, then f_E is **intersective** for all nouns N ($f_E(N) \subseteq N$) unless there is some k-property $\textsc{p} \in \mathbb{P}_k(N)$ such that $f_C(N)$ contradicts it. Formally, for any world w and context c,
$\forall N\,[\,\forall \textsc{p} \in \mathbb{P}_k(N)\,\exists x\,[\textsc{p}(w)(x) \wedge f_C(c)(w)(x)]]$
$\rightarrow \forall x\,[f_E(N)(c)(w)(x) \leftrightarrow f_C(c)(w)(x) \wedge N(c)(w)(x)]$

b. Else, if f_C depends on N, then f_E is **subsective** for all nouns N ($f_E(N) \subseteq N$) unless there is some k-property $\textsc{p} \in \mathbb{P}_k(N)$ such that $f_C(N)$ contradicts it. Formally, for any world w and context c,
$\forall N\,[\,\forall \textsc{p} \in \mathbb{P}_k(N)\,\exists x\,[\textsc{p}(w)(x) \wedge f_C(N)(c)(w)(x)]]$
$\rightarrow \forall x\,[f_E(N)(c)(w)(x) \rightarrow N(c)(w)(x)]$

3.3.2 Lexical entries and derivations

In this section, we will give example lexical entries for *stone* and *counterfeit* and provide example derivations of privative and subsective inferences when these inferences arise from the choice of noun.

Stone bowls and *stone lions*

While our adjectives are formally all of type $\langle s, \langle\langle s, \langle e, t\rangle\rangle, \langle e, t\rangle\rangle\rangle$, we will assume that *stone* corresponds to a MADE-OF-STONE property of type $\langle s, \langle e, t\rangle\rangle$: that is, for a given entity x, we can tell whether x is made of stone independently of what noun x is described as. Whether such a property exists is a decision that has to be made for each adjective on a case-by-case basis, and will determine whether the adjective yields intersective inferences in cases where it does not conflict with the noun (i.e. when the second half of its meaning preserves all the properties of N). I will assume that *stone* is indeed a bona fide adjective in English, as judged by the acceptability of sentences like *This statue is stone* (*but that statue is wooden/plastic*), but other material adjectives like *wooden* can be used interchangeably in its place for the purposes of these worked examples.

(35) Lexical entry for *stone*

$[\![\text{stone}]\!]^w =$

E-structure: $\lambda N.\lambda x.\ \textsc{stone}_w(N)(x)$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{made-of-stone}_w(x)) +$

$$\sum_{\textsc{p} \in \mathbb{P}(N)} \lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg \textsc{p}_w(x)) \rightarrow \textsc{p}_w(x))$$

The Subsective Meaning Postulate applies to *stone*.

I will assume a very toy lexical entry for *bowl* that just includes BOWL-SHAPED and HOLDS-SHAPE (when poked/used) as its sole k-properties; the point is that *stone* does not contradict any k-properties of bowl-hood, and so we get a subsective (actually, intersective) reading in the C-structure. (If it contradicts any t-properties, such as bowls typically being made of some

other material, these will get kicked out, but t-properties are not used for determining noun-hood.) The Genericity Constraint and the Subsective Meaning Postulate then ensure that (in all normal worlds), the extension yields a subsective inference too.

(36) Derivation of a subsective inference for *stone bowl*

a. $[\![\text{bowl}]\!]^w$ =

E-structure: $\lambda x.\ \textsc{bowl}_w(x)$

C-structure: $(\lambda x.\ \textsc{bowl-shaped}_w(x)) + (\lambda x.\ \textsc{holds-shape}_w(x))$

b. $[\![\text{stone bowl}]\!]^w = [\![\text{stone}]\!]([\![\text{bowl}]\!])$ =

E-structure: $\lambda x.\ \textsc{stone}_w(\textsc{bowl}_w)(x)$

C-structure:

$(\lambda x.\ \textsc{made-of-stone}_w(x)) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{bowl-shaped}_w(x)) \rightarrow \textsc{bowl-shaped}_w(x))) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg(\textsc{holds-shape}_w(x))) \rightarrow \textsc{holds-shape}_w(x)))$

$=(\lambda x.\ \textsc{made-of-stone}_w(x)) + (\lambda x.\ \textsc{bowl-shaped}_w(x)) + (\lambda x.\ \textsc{holds-shape}_w(x))$

c. Application of Genericity Constraint to *stone bowl*:

$$\bigwedge_{\textsc{p}\in\mathbb{P}([\![\text{stone bowl}]\!])} \textsc{Gen}\ x, w'\ [M_E([\![\text{stone bowl}]\!])(w')(x) \wedge C_{\textsc{p}}(x, w')]\ \textsc{p}(w')(x)$$

$\Leftrightarrow$(GEN x, w' [$\textsc{stone}_w(\textsc{bowl}_w)(x) \wedge C_1(x, w')$] $\textsc{made-of-stone}_w(x)$) $\wedge$

(GEN x, w' [$\textsc{stone}_w(\textsc{bowl}_w)(x) \wedge C_2(x, w')$] $\textsc{bowl-shaped}_w(x)$) $\wedge$

(GEN x, w' [$\textsc{stone}_w(\textsc{bowl}_w)(x) \wedge C_3(x, w')$] $\textsc{holds-shape}_w(x)$)

where C_1 represents the felicity conditions for *stone bowls are made of stone*, C_2 for *stone bowls are bowl-shaped* and C_3 for *stone bowls hold their shape*.

d. Application of Subsective Meaning Postulate to *stone*:

$\forall N\,[\ \forall \textsc{p} \in \mathbb{P}_k(N)\ \exists x\,[\textsc{p}(w)(x) \wedge \textsc{made-of-stone}_w(x)]]$

$\rightarrow \forall x\,[\textsc{stone}_w(N)(x) \rightarrow N(c)(w)(x)]$

Let N be *bowl*:

$[\ \forall P \in \mathbb{P}_k([[\text{bowl}]])\ \exists x\ [P(w)(x) \wedge \textsc{made-of-stone}_w(x)]]$

$\rightarrow \forall x\ [\textsc{stone}_w(\textsc{bowl}_w)(x) \rightarrow \textsc{bowl}_w(x)]$

e. Subsective inference: By assumption in (36b), there is no contradiction between any of the k-properties of *bowl* and MADE-OF-STONE in w: there exists some x in w that is made of stone, is bowl-shaped and holds its shape. So the antecedent of (36d) is true in w. Thus, $\forall x\ [\textsc{stone}_w(\textsc{bowl}_w)(x) \rightarrow \textsc{bowl}_w(x)]$, i.e. a *stone bowl* is a *bowl*.

The privative inference for *stone lion* is derived in a similar way, using the lexical entry for *lion* from (25), except that instead of invoking the subsective meaning postulate for *stone* we will invoke the Genericity Constraint on both *stone lion* and *lion* to yield the privative inference. Again, we omit the t-properties for simplicity, since they have no bearing on the inference.

(37) Derivation of a privative inference for *stone lion*

a. $[[\text{lion}]]^w =$

E-structure: $\lambda x.\ \textsc{lion}_w(x)$

C-structure: $(\lambda x.\ \textsc{born-from-lion}_w(x)) + (\lambda x.\ \textsc{furry}_w(x)) +$

$(\lambda x.\ \textsc{has-mane}_w(x)) + (\lambda x.\ \textsc{lion-shaped}_w(x)) +$

$(\lambda x.\ \textsc{mammal}_w(x))$

b. We assume that $\textsc{made-of-stone}_w(x)$ entails $\neg\textsc{born-from-lion}_w(x)$, $\neg\textsc{furry}_w(x)$ and $\neg\textsc{mammal}_w(x)$ in all sufficiently normal worlds (which we take to include w), by virtue of the meanings we assume these functions have.

Formally, $\textsc{Gen}\ x, w'\ (\textsc{made-of-stone}_{w'}(x) \rightarrow \neg\textsc{born-from-lion}_{w'}(x))$ etc.

We further assume that all the properties in the C-structure of *lion* as well as *made-of-stone* are interpreted generically with respect to the same modal base (in the context in which *stone lion* is uttered).

c. $[\![\text{stone lion}]\!]^w = [\![\text{stone}]\!]([\![\text{lion}]\!]) =$

E-structure: $\lambda x.\ \textsc{stone}_w(\textsc{lion}_w)(x)$

C-structure: $(\lambda x.\ \textsc{made-of-stone}_w(x)) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{born-from-lion}_w(x))$

$\rightarrow \textsc{born-from-lion}_w(x)])) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{furry}_w(x))$

$\rightarrow \textsc{furry}_w(x))) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{has-mane}_w(x))$

$\rightarrow \textsc{has-mane}_w(x))) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{lion-shaped}_w(x))$

$\rightarrow \textsc{lion-shaped}_w(x))) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{mammal}_w(x))$

$\rightarrow \textsc{mammal}_w(x)))$

$= (\lambda x.\ \textsc{made-of-stone}_w(x)) + (\lambda x.\ \textsc{lion-shaped}_w(x))$

All the k-properties except LION-SHAPED are removed by virtue of the entailments in (37b) which we assume to hold in w.

d. Application of Genericity Constraint to *stone lion*:

$$\bigwedge_{P \in \mathbb{P}([\![\text{stone lion}]\!])} \textsc{Gen}\ x, w'\ [M_E([\![\text{stone lion}]\!])(w')(x) \wedge C_P(x, w')]\ P(w')(x)$$

$\Leftrightarrow (\textsc{Gen}\ x, w'\ [\textsc{stone}_{w'}(\textsc{lion}_{w'})(x) \wedge C_1(x, w')]\ \textsc{made-of-stone}_{w'}(x)) \wedge$

$(\textsc{Gen}\ x, w'\ [\textsc{stone}_{w'}(\textsc{lion}_{w'})(x) \wedge C_2(x, w')]\ \textsc{lion-shaped}_{w'}(x))$

where C_1 represents the felicity conditions for *stone lions are made of stone* and C_2 for *stone lions are lion-shaped*.

e. By (37b) and (37d),

$\textsc{Gen}\ x, w'\ [\textsc{stone}_{w'}(\textsc{lion}_{w'})(x) \wedge C_1(x, w')]\ \neg\textsc{born-from-lion}_{w'}(x)$

f. Application of Genericity Constraint to *lion*:

$$\bigwedge_{\text{P}\in\mathbb{P}([\![\text{lion}]\!])} \textsc{Gen}\ x, w'\ [M_E([\![\text{lion}]\!])(w')(x) \wedge C_{\text{P}}(x, w')]\ \textsc{p}(w')(x)$$

$$\Leftrightarrow (\textsc{Gen}\ x, w'\ [\textsc{lion}_{w'}(x) \wedge C_3(x, w')]\ \textsc{born-from-lion}_{w'}(x)) \wedge \ldots$$

where C_3 represents the felicity conditions for *lions are born from lions*, C_4 the felicity conditions for *lions are furry*, etc.

g. Apply modus tollens[16] to the implicit conditional inside Gen (Krifka et al., 1995) in (37f): if x is a lion in w' and satisfies C_3 then it must be born from a lion in w'. So if x is not born from a lion then either it is not a lion or it does not satisfy C_3.

h. Almost-privative inference: By (37e) and (37g),

$$\textsc{Gen}\ x, w'\ [\textsc{stone}_{w'}(\textsc{lion}_{w'})(x) \wedge C_1(x, w')]\ (\neg\textsc{lion}_{w'}(x) \vee \neg C_3(x, w'))$$

We derive a privative inference in sufficiently normal worlds, which we take to include the actual world w in the sort of modal base we are dealing with, with the exception of these two felicity conditions C_1 and C_3. Stone lions which do not meet the felicity conditions for being made of stone (C_1) may in fact be lions. To address this, we might apply an additional meaning postulate to adjectives that their E-structure always respects the felicity conditions of their C-structure predicate—that is, stone lions always meet the felicity conditions for (lions) being made of stone. This meaning postulate can be applied to all adjectives except nonsubsective ones like *alleged*, which need further study outside of this dissertation.

(38) **Generic Felicity Meaning Postulate**

For any adjective following the schema in (31) with E-structure function f_E and C-structure function f_C, for all nouns N, let C_{N,f_C} be the felicity conditions associated with the generic "Ns are f_C". Then f_E encodes (entails) those felicity conditions. Formally, for any world w and context c,

$$\forall N\ \forall x\ [f_E(N)(c)(w)(x) \rightarrow C_{N,f_C}(w)(x)]$$

[16] If $(p \rightarrow q)$ then $(\neg q \rightarrow \neg p)$.

With this meaning postulate, we derive the following improved inference, since the restrictor $\textsc{stone}_{w'}(\textsc{lion}_{w'})(x)$ now entails $C_1(x, w')$:

(37) i. $\textsc{Gen}\ x, w'\ [\textsc{stone}_{w'}(\textsc{lion}_{w'})(x)]\ (\neg\textsc{lion}_{w'}(x) \vee \neg C_3(x, w'))$

The only remaining difference between this and a proper privative inference is the presence of C_3, which is the felicity condition associated with the k-property that *stone* negated that we used in the inference, in our case *lions are born from lions*. (We could have used any of the k-properties that *stone* negated, including *lions are mammals*.) This is only relevant if it is possible for $\neg C_3(x, w') \wedge \textsc{lion}_{w'}(x)$ to hold; otherwise, $\forall x\,[\neg C_3(x, w') \rightarrow \neg\textsc{lion}_{w'}(x)]$ and we get a privative inference. In our case, this seems likely: *lions are born from lions* and *lions are mammals* are such straightforward statements that they may not have any relevant felicity conditions, and so we would expect *stone lion* to be truly privative. But other k-property-negating adjective-noun bigrams such as *maneless lion* do make use of this distinction: female maneless lions can still be lions because they do not meet the felicity conditions for having manes in the first place. In general, this is an acceptable price to pay to capture the potential differences between concepts and intensions/extensions.

All of this is occurring under a strict, literal definition of *lion*.[17] We will see in Section 3.6 how the meaning of nouns such as *lion* can be loosened in context such that *lion* can still be used to refer to stone lions, as is in fact common in language use.

Counterfeit as an example "privative" adjective

In this section I will illustrate that we can use exactly the same mechanisms as for *stone* to derive the subsective inference for *counterfeit watch* and the privative inference for *counterfeit money*. All we need is an appropriate lexical entry for *counterfeit* and entries for *watch* and *money* that do or do not contradict what *counterfeit* entails.

[17] Using a strict definition of HAS-MANE, we also dropped *has-mane* from the C-structure of *stone lion*. In fact, a *stone lion* has a *stone mane*, just not a literal mane. This may be included in our LION-SHAPED k-property. Alternatively, this may hint at a bigger issue of the way that k-properties seem to refer to other concepts, such as *lion* to *mane*. Perhaps *stone* actually applies recursively to some of these properties, potentially following a mereological structure. I will leave this exciting but difficult to (semantically) implement possibility to future research.

The key insight of this section is that it is possible to give a lexical entry for *counterfeit* that is not inherently privative and that follows the schema in (31), and thus that there is no need for any invocation of context or other special treatment of these adjectives. Section A.1 further demonstrates this point by providing lexical entries for our remaining so-called privative adjectives *artificial*, *false* and *knock-off*. We will leave *former* to other work, since it displays significant additional complexity, such as scoping over possessives, and may just yield privative inferences as an implicature (Bigley, 2025; Morzycki, 2016; Reinert, 2024).

The meaning of *counterfeit* involves two core components: the fact that the object in question appears to have some authentic origin or brand but does not, and an element of deception. That deception is necessary for both *fake* and *counterfeit* is shown in (39):[18]

(39) Context: An art student paints a copy of a famous painting by Michelangelo as a way to learn and better appreciate the artist's techniques. The art student keeps the painting for himself and has no intention of claiming it as anything other than his own work.

(adapted from Guerrini, 2024b)

a. #This is a fake/counterfeit painting.

b. #This is a fake/counterfeit Michelangelo painting.

c. This is a painting.

d. #This is a Michelangelo painting.

We will reduce the core of *counterfeit* to the hopefully more atomic predicate of properties AUTH-ORIGIN(Q) that tests whether Q is a property that captures whether something has the origin or brand that it purports to have, and then existentially close over such properties.[19] This appeals to the idea discussed in Section 3.2 that while k-properties are not labeled like del Pinal's dimensions, humans do have a way of classifying whether a given property of

[18]Judgment confirmed in Exp. 4: 16 of 20 participants found *that is a fake painting* unacceptable in this context.

[19]Del Pinal (2018) appeals to the origin dimension AGENTIVE of his C-structure to define *counterfeit*. While this potentially works, it is unclear whether the origin of a typical *watch* necessarily includes authenticity; most watches do not have recognizable brands and authenticity is not at stake. More to the point, we no longer have an origin dimension since we are using k-properties and t-properties instead of dimensions.

a noun is of a certain type (perceptual, authentic origin, etc.) or not when "asked" (by the adjective) to do so.

It is a somewhat thorny question whether *counterfeitness* can be evaluated without the presence of the noun; in other words, whether it corresponds to a property of type $\langle s, \langle e, t \rangle \rangle$ or not and by extension whether AUTH-ORIGIN needs to depend on the noun N. *Counterfeit* can be used in predicate position. Further, I can point at some unknown artefact that has what looks like a brand logo emblazoned across it and say *this blicket / this is counterfeit*. You do not appear to need to know what a *blicket* is to determine that what is wrong is that it was not made in the right kind of blicket factory or sold by an authorized blicket seller. On the other hand, the fact that this refers to a *blicket factory* or a *blicket seller* suggests that the meaning of *counterfeit* perhaps does depend on the noun; perhaps when *counterfeit* is used in predicate position a suitable noun is contextually supplied. (See discussion of predicate position in Section 3.7.3.) We will conservatively say that *counterfeit* does depend on the noun and evaluate AUTH-ORIGIN(Q) on a dimension Q of N; thus *counterfeit* has a selectional restriction that it can only apply to nouns which have such an origin dimension (k-property or t-property). While this will make the derivation easier, it is not critical to our general point that *counterfeit* is neither intrinsically privative nor context-dependent.

I will spell out this first component of the meaning of *counterfeit*—appearing to have some authentic origin or brand but not actually having it—but leave the deception component opaque as DECEPTION(x, Q) where Q is the origin property defined by the first component. As Guerrini (2024b) argues, deception can be implemented as "intent to cause a belief that the object has property Q". This can be cashed out straightforwardly using the well-studied possible world semantics of intent, belief and causality; for brevity I will not do this in our lex-

ical entry. For the first component, I will use the implementation of SEEM-LIKE from Guerrini (2022) that Guerrini (2024b) uses in his lexical entry for *fake*.[20,21]

(40) Lexical entry for *counterfeit*

$[\![\text{counterfeit}]\!]^w =$

E-structure: $\lambda N.\lambda x.\, \textsc{counterfeit}_w(n)(x)$

C-structure: $\lambda N.\, (\lambda x.\, \textsc{cf}_w(N)(x)) + \sum_{\textsc{p} \in \mathbb{P}(N)} \lambda x.(\neg(\textsc{cf}_w(N)(x) \rightarrow \neg \textsc{p}_w(x)) \rightarrow \textsc{p}_w(x))$

where $\textsc{cf}_w(N)(x) = \exists Q \in \mathbb{P}(N)[\textsc{auth-origin}_w(Q) \wedge \textsc{seem-like}_w(x, Q) \wedge \neg Q(w)(x) \wedge \textsc{deception}_w(x, Q)]$

We can now derive the privative inference for *counterfeit money* by making the assumptions that *money* has a k-property MADE-BY-FEDERAL-BANK (or similar) that is the right type of authentic origin property, i.e. AUTH-ORIGIN(MADE-BY-FEDERAL-BANK) holds. I will lay out the remainder of derivations informally; the formalism will be the same as in (36) and (37).

(41) Derivation of a privative inference for *counterfeit money*

a. The lexical entry for *money* includes a k-property MADE-BY-FEDERAL-BANK (or similar) that satisfies AUTH-ORIGIN(MADE-BY-FEDERAL-BANK); this is the only dimension of *money* that does so. It also includes k-properties such as being exchangeable for goods and some t-properties.

b. When evaluating *counterfeit money*, the only choice for Q in the "core meaning" component $\textsc{cf}_w([\![\text{money}]\!])(x)$ is MADE-BY-FEDERAL-BANK; thus this gets negated in the C-structure for *counterfeit money*.

[20]Guerrini's implementation of *seem-like* depends on the context, strictly speaking: x seems like a Q if there is a contextually specified set of properties that they share. In that respect, *counterfeit* is in fact context-sensitive. However, this context-sensitivity should have no bearing on the privative/subsective inferences we draw, especially since we fix $\neg Q(x)$.

[21]Instead of existentially closing Q, we could also have it be supplied by the context. However, it is unclear what this gains us; at best this helps us in situations where I am looking at a counterfeit Rolex watch and calling it just a *counterfeit watch*, in which case Q is presumably *Rolex*. If "context" is just a way of getting at the noun, we might as well have AUTH-ORIGIN depend on N and use a contextually supplied silent noun in predicate position.

c. The Genericity Constraint ensures that entities in the extension of *counterfeit money* negate PRINTED-BY-FEDERAL-BANK in sufficiently normal worlds, including the actual world, provided the felicity conditions for being *counterfeit money* (according to the C-structure, i.e., the conditions for $\text{CF}_w([\![\text{money}]\!])(x)$) are met. By the Generic Felicity Meaning Postulate, we assume that being in the extension of *counterfeit money* in these worlds entails these felicity conditions.

d. The Genericity Constraint applied to money requires that PRINTED-BY-FEDERAL-BANK holds of entities in the extension of *money* in sufficiently normal worlds, provided the felicity conditions for being printed by a federal bank hold.

e. By modus tollens applied to the conditional inside GEN, in all sufficiently normal worlds, including the actual world, *counterfeit money* is either not *money* because PRINTED-BY-FEDERAL-BANK does not hold of it, or some particular instance of *counterfeit money* may be *money* if it did not meet the felicity conditions for being printed by the federal bank in the first place.

f. As for *stone lion*, it is hard to imagine something in the actual world that does not meet the felicity conditions for being printed by the federal bank while also being money, so we expect a fully privative inference in the actual world.

For *counterfeit watch*, we want to derive a subsective inference. Observe that *watches have brands* sounds felicitous. This means that *having a brand* (or strictly, having that brand authentically) is either a k-property or a t-property of watches. One way to derive this inference is to assume that having a brand is a t-property of *watch*: while most watches have brands / authentic origins, this is not necessary in order to be a watch. Then the derivation goes as follows:

(42) Derivation of a subsective inference for *counterfeit watch*

a. The lexical entry for *watch* includes a t-property *has-brand* that satisfies AUTH-ORIGIN and is the only dimension of *watch* that does so.

b. When evaluating *counterfeit watch*, this t-property *has-brand* is the only choice for Q, thus *has-brand* is negated in the C-structure of *counterfeit watch*. The k-properties of *watch* are not affected by this and are all preserved in the C-structure of *counterfeit watch*.

c. The Genericity Constraint ensures that entities in the extension of *counterfeit watch* have all the k-properties of *watch*, since *counterfeit watch* preserves them all.

d. The Subsective Meaning Postulate applied to *counterfeit* with N as *watch* then yields a subsective inference since no properties were negated in the C-structure.

An alternative approach is to assume that *counterfeit* actually requires that the dimension that it targets is a k-property of the noun, and that this selectional restriction thus coerces nouns to have such authentic origin properties as k-properties. This brings it more in line with *fake*, which we concluded in Section 2.7 in Chapter 2 only targets k-properties, and can mean *counterfeit*. However, this does not allow us to derive subsective inferences using the Genericity Constraint, since the Genericity Constraint enforces k-properties but *counterfeit* is now negating one). We will discuss this in more detail in Section 3.4.2 when we discuss how to derive subsective inferences with *fake*.

3.3.3 Recursive adjective modification

We need to assume C-structure is compositional in this theory for two reasons: (1) to constrain the E-structure of the composed phrase and (2) because privative inferences can occur in recursive adjective modification, so our conflict resolution process must be able to operate on adjective-noun phrases, not just on nouns.

In (43), *edible flowers* refers to types of flower like rose or hibiscus, which can safely be eaten by humans and taste pleasant. We can then depict these types of flower as plastic models, resulting in *plastic edible flowers*—a particular kind of *plastic flower* that depicts these types of flower. Notably, the final result is no longer edible, since it is made of plastic.

(43) Context: A store sells many kinds of plastic and glass flowers. A chef is looking to decorate his restaurant only with (imitations of) edible flowers, such as roses and hibiscus, which he also uses in his cooking. He prefers plastic ones since they are less likely to break. He asks the store owner where to find them.
"You can find some plastic edible flowers over there."

We derive a privative inference that *plastic edible flowers* are not *edible flowers* (though they are *plastic flowers*) following exactly the same schema as for *stone bowl* for *edible flower* and the same schema as *stone lion* for *plastic* [*edible flower*], because *plastic* causes the edibility property contributed by *edible* to be dropped. The full derivation is given in Section A.2. This works smoothly in our theory but is actually harder to account for with a widening theory, since the whole phrase *edible flower* must be widened, not just the head, requiring a notion of (appropriately constrained) widening that applies to phrases, not just nouns.

3.4 Accounting for *fake* and *real*

Section 3.3 sets up a new framework for adjective modification that handles privative inferences directly within the modification semantics. We are now in a position to slot *fake* and *real* into this theory. This will also be our first encounter with context, which we will discuss in greater detail in Section 3.6. Unlike other "privative" adjectives like *counterfeit* and *artificial*, we will argue that *fake* is context-sensitive and can target different k-properties depending on the context.

3.4.1 *Fake* and *real* as selectors over C-structure properties

In Section 2.7 of Chapter 2, we established that *fake* and *real* target k-properties and require the noun to either have (for *real*) or not have (for *fake*) that property. In Section 3.2 of this chapter, we set up our C-structure to contain k-properties and t-properties, and specified that the k-properties can be singled out by $\mathbb{P}_k$. All we need now is to have *fake* and *real* select among the properties in $\mathbb{P}_k$.

We can implement this in two ways: either *fake* and *real* can select a k-property to apply to conditioned on the context, or we can existentially close and just say that there exists some k-property that they target. In the interest of more transparent inference derivations, we will implement this as a context-sensitive CHoice operator over Properties ChP that takes in the context and the list of k-properties $\mathbb{P}_k$ and returns one. Just like *counterfeit*, we also need a deception component for *fake* (see the same example (39)). This yields the lexical entry in (44). The lexical entry for *real* will also use ChP to assert instead of negate a k-property; we'll state it formally in Section 3.4.3.

(44) Lexical entry for *fake*

$[\![\text{fake}]\!]^{w,c} =$

E-structure: $\lambda N.\lambda x.\ \textsc{fake}_w(N)(c)(x)$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{fk}_w(N)(c)(x)) +$

$$\sum_{\textsc{p}\in\mathbb{P}(N)} \lambda x.(\neg(\textsc{fk}_w(N)(c)(x) \rightarrow \neg\textsc{p}_w(x)) \rightarrow \textsc{p}_w(x))$$

where $\textsc{fk}_w(N)(x) = \neg(\textsc{ChD}(\mathbb{P}_k(N))(c))(x) \wedge \textsc{deception}_w(x, (\textsc{ChD}(\mathbb{P}_k(N))(c)))$

In other words, x is *fake* N in a context c (in sufficiently normal worlds / according to the C-structure) if the k-property of N returned by the choice operator in this context, $Q = \textsc{ChP}(\mathbb{P}_k(N))(c)$, does not hold of x and there is an intent to deceive people that Q holds of x.

We will further assume that ChP respects the centrality ranking of k-properties and preferentially chooses higher-ranked properties. This accounts for the intuition that the most salient reading of *fake Rolex* is *counterfeit Rolex*, and not that it is not even a watch at all. If we assume the authenticity k-property for Rolexes (something like *Rolex watches are made by Rolex, the company*) is very important for Rolex watches and ranked high, while *Rolexes are watches* is also a k-property of *Rolex* but is ranked much lower, then this difference in reading availability can be explained nicely by our account.

We are now in a position to derive privative inferences for *fake*. Since we are now using a context c, I will assume that our modal base for GEN is restricted to worlds in which c also holds and where c has the same effect on CHP as in the actual world.

(45) Derivation of a privative inference for *fake watch* in a context c where what is salient is that watches should tell the current time and we are discussing a watch which does not, such as a theater prop.

a. Assume that *watch* has a k-property TELLS-TIME that it shows the current time (or is at least capable of doing so).

b. Assume that $\text{CHP}(\mathbb{P}_k([\![\text{watch}]\!]))(c) = \text{TELLS-TIME}$.

c. Evaluating *fake watch* adds $\neg(\text{CHP}(\mathbb{P}_k([\![\text{watch}]\!]))(c))(x) = \neg\text{TELLS-TIME}(x)$ to the dimensions of *fake watch*. This contradicts the k-property $\lambda x.\text{TELLS-TIME}(x)$, so that property is removed by the conflict resolution mechanism in the latter half of *fake*'s lexical entry.

d. The Genericity Constraint applied to *watch* requires that all k-properties of *watch* hold in sufficiently normal worlds, provided their felicity conditions are met. The Genericity Constraint applied to *fake watch* asserts that $\neg\text{TELLS-TIME}(x)$ holds of fake watches in sufficiently normal worlds in context c.

e. By modus tollens applied to the conditional inside GEN, in context c *fake watches* are either not *watches* in sufficiently normal worlds, including the actual world, or the particular *fake watch* may be a *watch* but did not meet the felicity conditions for telling the time in the first place. While there are cases such as *broken watches* which fail to meet these conditions and are still *watches*, none of these seem relevant for our theater prop case in c.

3.4.2 Subsective inferences with *fake*

A second immediate observation that follows from (45) is that the privativity inference in (45) does not actually depend on *c*. It doesn't matter what k-property ChD picks: the sheer fact that it picks a k-property causes the *fake N* to violate the requirements of the Genericity Constraint on *Ns* in sufficiently normal worlds such as the actual world (unless it fails to meet the felicity conditions for the k-property). This outcome is not specific to our lexical entry of *fake*. It is guaranteed as soon as we combine the following:

- Assumption 1: *Fake* targets (negates) k-properties (Section 2.7 in Chapter 2)
- Assumption 2: K-properties hold of their nouns in generic sentences (with bare plurals and indefinite singulars) (Prasada and Dillingham, 2006)
- Assumption 3: Properties that hold in generic sentences are true in sufficiently normal worlds defined by Gen's modal base, which usually includes the actual world for characterizing generic sentences like ours (Krifka et al., 1995)
- Inference 1: Therefore, in these worlds, k-properties hold of their nouns.
- Inference 2: Therefore, negating a k-property causes privativity in these worlds, except in cases where the felicity conditions for that k-property do not hold and it is still possible to be an instance of the noun.
- Inference 3: Therefore, *fake* is always privative in these worlds for all stereotypical examples of the noun which do satisfy all k-properties (or even just this k-property) and their felicity conditions.

Our Genericity Constraint simply formalizes Assumptions 2 and 3. It is unclear if this is a desirable outcome. On one hand, this appeals to many people's intuitions in prior literature (del Pinal, 2018; Guerrini, 2024b; Partee, 2010) that *fake* really is inherently privative on some level. Many of the subsective inferences we observe with *fake* are in explicit contexts in Experiment 3 and Experiment 4, meaning the inference holds in some context *c* about some specific object *x*, though there were some k-properties that seemed to generally yield subsective inferences, such as *lifeguards can swim*.

There are four possible paths we can take, with different paths likely applicable for different subsective inferences:

(P1) The context c can only occur in worlds w' which are not sufficiently normal in our (contextually supplied) modal base and ordering source for the Genericity Constraint to apply. This will apply in some cases, but probably not to any of our examples.

(P2) The particular described object x does not meet the felicity conditions for the k-property being targeted. This may apply to examples such as a blunt tapestry needle being used as a *fake* (*sharp*) *needle* theater prop (standing in for a sharp sewing needle; an example in Experiment 4); this tapestry needle is an atypical needle and likely fail to meet the felicity criteria for *needles are sharp*.

(P3) The noun refers to a dual character concept (Knobe, Prasada, and Newman, 2013), of which *lifeguard* may be an example, such that "x is a lifeguard some sense but x is not a lifeguard in some deeper sense". In that case, *lifeguards can swim* is a k-property associated only with the deeper sense, and not with the superficial sense, and so *fake* is privative with regards to the deeper sense (to which it is applied) but subsective with regards to the superficial sense, which may be used as the predicate in *Adam* (*the fake lifeguard*) *is a lifeguard*.

(P4) *Fake* applies to a narrower sense of the noun given by the context, not to the standard use of the noun. This seems intuitive for *fake watch*, which is taken to mean something like *fake designer watch* when it is subsective. Section 3.6 gives an implementation for this as part of our discussion of context-sensitivity. This is in the spirit of Reinert (2024) and Guerrini (2024b) (see Section 3.5), but Section 3.6's implementation is much more constrained than theirs.

Further, we also observe subsective inferences in Experiment 2, which asks questions involving indefinite singulars that seem to be generics: *Is a fake watch still a watch?* While 2 of 12 participants do respond with a privative rating, 7 assign a subsective rating (the remaining

3 choose "Unsure"). In fact, however, we seem to get similar behavior with properties that hold in generics: if you turn the classic generic statement *A lion has a mane* into a question *Does a lion have a mane?*, it is easy to imagine some people answering yes while other people might answer "well, it depends on whether it's a female lion...". This response invokes the felicity conditions for *a lion has a mane* rather than tacitly accommodating them. It would seem that when the generic is asked as a question, it's possible to quantify not over all lions that meet the felicity conditions, but rather over all lions, yielding a "probably" or "sometimes" response based on the number of lions that do have manes. This thus supports the felicity condition path, Path (P2).

More study of the individual contexts is needed to determine whether Paths 1-3 can cover the data we see (and how exactly Path (P3) would work for dual character concepts), or whether we need to resort to Path (P4) to derive subsective readings with *fake*. If we choose Path (P4), we must adequately constrain what kinds of noun can be supplied—see Section 3.5 for discussion of the overgeneration that occurs otherwise. Section 3.6 gives a constrained account of noun narrowing based on del Pinal (2018), with an explicit example of a subsective inference for *artificial reef*, by only allowing noun narrowing based on existing k-properties and t-properties.[22]

A final outcome of this discussion is that we do not get a perfect parallel between *fake* and *counterfeit*, even though *fake* sometimes seems to take on the meaning of *counterfeit*. We derived subsective inferences for *counterfeit* for nouns like *watch* by assuming that the property that *counterfeit* targeted was a t-property of *watch*. *Fake*, by assumption in Section 2.7 in Chapter 2 and by virtue of our lexical entry, only targets k-properties. Path (P4) allows us to coerce the noun into a narrower reading which may promote certain t-properties to be k-properties, such as authentic origin being a k-property of *designer watch* but only a t-property of *watch*.

[22]Allowing contextual coercion of the noun, and by extension what is a k-property might help explain some of the intermediate ratings we saw in Experiment 4 where some properties labeled k-properties by Prasada and Dillingham were deemed only somewhat felicitous for *fake* to target in the provided contexts by some participants. However, this must be adequately constrained in order to avoid opening a Pandora's box.

3.4.3 What is the meaning of *real*?

The mirror observation of this discussion for *fake* is that *real* is entirely vacuous in the C-structure and has only a subtle effect on the E-structure in sufficiently normal worlds, namely by targeting felicity conditions. Its lexical entry, shown in (46), is the mirror of *fake*: instead of negating a k-property, it asserts it. Since this property was already in the C-structure, this does nothing in the C-structure. What we want *real* to be doing is enforcing this property in the E-structure, where it may not always hold if its felicity conditions are not met. This lets us say things like *Real lions have manes!* (del Pinal, 2018) and place a constraint on lions which previously only held if the felicity conditions were met (in sufficiently normal worlds).

(46) Lexical entry for *real*

$[\![\text{real}]\!]^{w,c} =$

E-structure: $\lambda N.\lambda x.\, \textsc{real}_w(N)(c)(x)$

C-structure: $\lambda N.\, (\lambda x.\, \textsc{rl}_w(N)(c)(x)) + \sum_{P \in \mathbb{P}(N)} \lambda x.(\neg(\textsc{rl}_w(N)(c)(x) \rightarrow \neg P_w(x)) \rightarrow P_w(x))$

where $\textsc{rl}_w(N)(x) = (\textsc{ChP}(\mathbb{P}_k(N))(c))(x)$

In other words, x is *real* N in a context c (in sufficiently normal worlds / according to the C-structure) if the k-property of N returned by the choice operator in this context, $\textsc{ChP}(\mathbb{P}_k(N))(c)$, holds of x.

The fact that *real* might be vacuous is already put forward by Partee (2010) and observed as early as Lakoff (1987). However, if we apply the Generic Felicity Meaning Postulate from (38) to *real*, we see that it is not vacuous.[23] The Generic Felicity Meaning Postulate requires that $\textsc{real}_w$ (f_E for *real*) entails the felicity conditions of $\textsc{rl}_w$ (f_C for *real*). The felicity conditions of $\textsc{rl}_w$ are precisely the felicity conditions of whatever k-property ChP selected. In other words, applying *real* to a noun's k-property removes the requirement that the felicity condition for that k-property must hold in order for the k-property to also hold of the extension in

[23]Section A.4.5 gives an alternative lexical entry for *real* which implements the desired k-property enforcing effect directly on the E-structure, without need for a meaning postulate.

sufficiently normal worlds, such as the actual world. Thus, using *real lions* in a context where manes are salient means that only male lions with manes qualify as *real lions* in that context.

3.5 Context and previous accounts of privativity

The last generalization that we need to implement in our new theory is Generalization (G2), namely that privativity inferences can vary for a given adjective-noun bigram when different contexts are supplied. Before we do that, it is important to look at existing context-dependent accounts of privativity. As hinted at earlier, these accounts do not constrain (sufficiently, or at all) how context acts in these cases and end up overgenerating meanings of *fake* that are not attested. This will motivate us to use a constrained version of contextual modulation based on del Pinal (2018) that does not encounter these issues.

3.5.1 Context-sensitive accounts of *fake*

Noun widening As laid out in Chapter 1 and Chapter 2, one option put forward by Partee (2010) is to assume that adjectives like *stone* and *fake* are subsective by default and that the Non-Vacuity Principle widens the noun in cases of privativity. This type of widening account can be viewed as a contextual modulation account, since context modulates the noun meaning (del Pinal, 2018; Recanati, 2010).

***Fake* takes a contextual variable** Reinert (2024) and Guerrini (2024a,b) propose accounts of *fake* where *fake* does not in fact operate on the noun, but in fact on a contextually supplied variable C, which is often but need not be identical to the noun. $[\![\text{fake}]\!](C)$ then intersects with the noun via predicate modification. Both of these accounts also use noun widening in order to accomplish this intersection in cases of privativity. The addition of a contextual variable to a widening account allows them to explain subsective inferences with *fake*: in these cases, *fake* applies (privatively) to some other contextually specified noun or property, such as [*fake* ~~*Rolex*~~] *watch*. (47) shows Reinert's use of a contextual variable for *fake*.

(47) $[\![\text{fake}]\!] = \lambda x_e.\, \text{fake}_w(\mathcal{D}_{\text{CONT}})(x)$

where 'fake' is a constant of type $\langle s, \langle s, \langle e, t\rangle, \langle e, t\rangle\rangle\rangle$ and $\mathcal{D}_{\text{CONT}}$ represents the context-dependent aspect in which x is fake.[24]

Reinert's *Preference principle for the pragmatic specification of free variables* ensures that the context assigns the adjacent noun's meaning to $\mathcal{D}_{\text{CONT}}$ by default. The motivation for this move is that adjectives like *skillful* and *good* can take a contextual variable (or other silent contextually supplied content) (Maienborn, 2020; Martin, 2022; Reinert, 2024): contexts like (48) show that *skillful* need not apply to the noun adjacent to it.

(48) Context: At a music school, there is an annual chess contest in which the different music classes compete against each other. Usually, the pianists are not doing very well, but things are different this year. As it turns out, some of the new piano students have been playing chess for years, and have won almost all of their matches. One of the students of the violin class says to one of his classmates: "Wow, we have a lot of skillful pianists here this year!" (Reinert, 2024)

Guerrini's implementation takes a similar shape: he classes *fake* and *skillful* as "type 1" adjectives, which compose following the schema in (49) using a reduced relative clause analysis of adjectives.[25] (49) shows the formulation given in Guerrini (2024a); Guerrini (2024b) uses a syntactically realized covert N (following Martin (2022) on *good*/*skillful*) instead of a contextual variable. The contextual variable C plays a similar role to Reinert (2024), albeit minus the principle to constrain C to preferentially choose the noun.

[24]This wording is quoted directly from Reinert (2024)—Reinert does not further explain the idea of "context-dependent aspect". Her examples suggest that it is another noun phrase like *Rolex* or *Gucci product* and do not provide a mechanism for explicitly linking it to x as an "aspect of x".

[25]In fact, Guerrini provides two versions of adjective modification for prenominal and postnominal modification (needed for Italian). We use the Italian postnominal adjective modification variant for English prenominal adjectives since they share the same pattern of inferences; Guerrini claims that prenominal Italian adjectives only yield privative inferences.

(49) $[\![\text{Det Adj N}_{\text{type\_1}}]\!] = [\![\text{Det}]\!]([\![\text{Adj}_{\text{type\_1}}]\!](C) \cap [\![\text{N}]\!])$

with C an implicit variable.

If Adj has a privative lexical meaning and $[\![N]\!] \subseteq C$, the above meaning is vacuous. In this case, the noun is broadened via a Partee-style broadening mechanism. This broadened extension will encompass at least the following set: $[\![\text{N}]\!] \cup [\![\text{Adj}_{\text{type\_1\_privative}}]\!](C)$

Guerrini (2024b) also gives a non-atomic lexical entry for *fake* that demonstrates how point (a) of Corollary (C3) (inheriting enough properties of the noun) is solved despite *fake* being privative.

Both of these accounts thus build on noun widening to explain the subsective inference of *fake watch* as $[\![\text{fake}]\!]([\![\text{Rolex}]\!]) \cap [\![\text{watch}]\!]$ and the privative inference of *fake gun* as applying to the noun directly $([\![\text{fake}]\!]([\![\text{gun}]\!]) \cap [\![\text{gun}_{\text{widened}}]\!])$.

3.5.2 Issues with unconstrained context-sensitivity

These accounts of privativity (as exemplified by *fake*) invoke context in two places: first, to widen the noun when *fake* is privative, and second, in the contextual variable accounts, to supply an appropriate contextual variable when the overall inference is not privative. However, none of these theories place any restrictions on what set the context widens the noun to or what values the contextual variable can take. This causes each of these steps to generate unattested meanings of *fake N*.

Del Pinal (2018) observes that it should be possible to widen *gun* to mean *toy gun* in contexts where this is salient, and where some kind of widening is required in order to resolve *fake*. However, *fake gun* does not mean *fake ~~toy~~ gun* even in a context like (50):[26]

[26]Judgment confirmed in Exp. 4: 16 of 20 participants found *that is a fake gun* unacceptable in this context.

(50) Context (adapted from del Pinal, 2018): Some terrorists planned a mass shooting at a Wild West-themed costume party. They expected a lot of people to dress up as cowboys and carry toy guns as part of their costume. The terrorists also dressed up as cowboys, but they disguised the gun they wanted to use for the shooting to look like a toy gun, so that they could sneak it into the party.

a. #I heard some disturbing news. Some terrorists constructed fake guns and planned to use them to attack a costume party.

b. I heard some disturbing news. Some terrorists constructed fake toy guns and planned to use them to attack a costume party.

Further unattested meanings are generated when we allow a contextual variable. The context in (17), repeated from Chapter 2, makes *metal gun* salient, but *fake gun* cannot mean [*fake ~~metal gun~~*] $\cap$ *gun* in this context.

(17) Context: Josh works at a metalworking shop and has just recently purchased a collection of guns. His only interest in purchasing the guns is to melt them down; the metals that guns are always made of is valuable to him for his next project, so it's irrelevant whether the guns are broken or function. When going through the pile and melting each gun down, he finds that one doesn't melt properly, and upon closer inspection is actually made of a futuristic, high-quality plastic. It functions perfectly well, with the same internal mechanisms as a standard gun, but has none of the same materials. The manufacturer didn't want their guns to stand out, for some reason—perhaps the army didn't want to reveal their new technology—and so the plastic is designed and painted deceptively to look precisely like a standard metal gun.

(adapted from Martin, 2022)

a. #Josh says: "They sold me a fake gun!"

b. Josh says: "They sold me a fake metal gun!"

c. Josh says: "They sold me a gun."

In this example, the proposed contextual variable C, *metal gun*, is a subset of *gun*. The overgeneration of such accounts is even more striking when we consider examples such as (51) where the supposed C is not a subset of N. This is possible with *skillful*, as in (48), but is quite infelicitous for *fake*. We cannot use *fake rock formation* to mean [*fake ~~reef~~*] $\cap$ *rock formation*, as would be predicted by contextual variable accounts.

(51) Context: A resort in Caribbean wants to encourage more scuba diving in their area, but unfortunately there are no coral reefs nearby. Their divers find an extremely unusually shaped rock formation just off their coast which has the perfect shape to look like a reef. They cover it in realistic-looking plastic corals and start feeding the local fish to attract them in large numbers. Visiting tourists are happy to scuba dive with all the fish, even though they know it's just a manmade attraction.

a. #This is a fake rock formation.

b. This is a fake reef.

c. This is (among other things) a rock formation.

Similar examples can be constructed for other "privative" adjectives such as *artificial* and *counterfeit* that these theories would presumably handle in the same way as *fake*—see Section A.5.[27]

It is possible to overcome these issues, especially in the case of noun widening, which is already tied to the noun. Section A.6 illustrates how the constrained contextual modulation function introduced by del Pinal (2018) can be used to implement noun widening and avoid the examples above. This constrained version of noun widening accounts for all our data, although it does not implement the generalizations in Section 2.8 as well as our new theory. Our theory concludes that contextual modulation is not necessary to account for privative inferences when the adjective modifies a property that is crucial to the noun. However, we

[27]Note that these theories do not require a contextual variable for material adjectives like *stone*, since these are not assumed to be inherently privative. Subsective inferences with material adjectives presumably arise by default under these accounts, while widening is invoked for privative cases like *stone lion*.

will invoke the context to explain the fact that inferences vary for many other adjective-noun bigrams depending on the discourse context provided.

3.6 Explaining context-sensitive inferences

Generalization (G1) states that privativity occurs whenever an adjective modifies properties that are core to a noun's meaning. In principle, subsective inferences occur whenever that is not the case. However, in many cases, information that might be relevant to a privative inference is left underspecified—for example, *counterfeit* says that the origin is not authentic but does not specify exactly how it was tampered with. Thus, *counterfeit spring water* of some particular brand may or may not be *spring water* depending on what happened during manufacture. The discourse context can fill in this information and shift the inference one way or the other. A second effect of context that we observed Experiment 3 is that the context sometimes clarifies the standards for noun-hood, essentially modulating the noun meaning before the adjective applies to it and before the same noun is used in predicate position of the inference. (*Is a fake reef still a reef?* It depends how picky we're being about reefs.)

These effects occur entirely independently of the contradiction resolution done by the adjective in Section 3.3, something which our theory allows us to nicely separate. Instead, we will show how standard discourse updates and the constrained contextual modulation mechanism proposed by del Pinal (2018) account for this variation.

3.6.1 Context-specific standards for noun-hood

Contexts such the ones used in Experiment 3 in Chapter 2 play two roles: firstly, they fill in information to specify the properties of the adjective-noun bigram that are underspecified by the bigram itself, as in the examples in Section 3.6.2. Secondly, they also provide the standards for what counts as a noun in this context. This is particularly evident in the minimal pair in (52a) and (52b), adapted from the *fake reef* example used in Experiment 3, but can be seen to a lesser degree in many other of the contexts in Experiment 3 and throughout this chapter.

(52) a. Context: Off the coast of Australia sits an artificial reef, which is a manmade structure covered in thousands of zoo-grown corals. A few years into the project, it supports a whole ecosystem of fish and corals. Martha, a marine biologist, is compiling a list of reefs that might be affected by coral bleaching. Since the corals on the artificial reef are in danger too, she includes it on her list of reefs.
Martha says: "The artificial reef is / counts as a reef."

b. Context: Off the coast of Australia sits an artificial reef, which is a manmade structure covered in thousands of zoo-grown corals. A few years into the project, it supports a whole ecosystem of fish and corals. Martha, a marine biologist, is compiling a list of reefs to list as natural heritage sites. She does not include the artificial reef despite its biodiversity, since it is manmade.
Martha says: "The artificial reef is not / does not count as a reef."

What we can call something is known to be context-dependent and flexible. As concrete evidence, Martin (2022) found that participants were perfectly happy to call toy bears *bears* even when (pictures of) real bears were also available to choose (though the presence of real bears did slightly reduce toy bear choice). This is reflected in the difficulty in determining what exactly is in the extension or intension of any given concept, which we typically choose to denote atomically as $\lambda x.\textsc{bear}(x)$ rather than specifying in more detail.

You may be content to say that the extension of *bear* or *reef* in a world w is always contextually determined and that this is not a job for semantics. However, del Pinal (2018) provides a constrained contextual modulation function mod_l that can account for the type of noun meaning variation we see in relation to privativity inferences, while also constraining it, by virtue of referring to the noun's C-structure. This explains why we can use *bear* to refer to things that have properties of bears (such as looking like bears) but not to, for example, *giraffes*.

While I will refer the reader to del Pinal (2018) for the technical implementation, the idea is that the contextual modulation function mod_l can apply to the E-structure of any noun (or indeed any composite expression that has a C-structure). It can then adjust the E-structure of the noun in one of two ways:

1. **Contextual widening**: mod_l can return a *k-property operator* ("dimension operator" in del Pinal, 2018) that selects one or more k-properties from the C-structure. This operator replaces the E-structure of the noun with the conjunction of these properties, which will typically result in a wider meaning than before.

2. **Contextual narrowing:** mod_l can return an *enrichment operator* that selects one or more k-property or t-property from the C-structure. This operator conjoins these properties with existing the E-structure of the noun and sets this conjunction as the new E-structure. This results in a narrower meaning whenever the enrichment operator selects non-crucial properties.

Del Pinal's original formulation allows mod_l to target any dimension; since we now include t-properties in our C-structure, I will restrict dimension/property operators to k-properties only to avoid unwanted widenings (such as widening *barn* to *things that are red*).[28] This restriction is not necessary for enrichment operators.[29]

A typical case of widening with mod_l is that it chooses a k-property operator that only returns perceptual dimensions, yielding a meaning of roughly "looks like an N".[30] Widening of this sort is applied whenever you point at a stone lion or toy lion and say "That's a lion", since $\lambda x.\, \textsc{lion}_{w_0}(x)$ ordinarily requires other lion-properties beyond just appearance.

A case of narrowing with mod_l would be a child pointing at a female lion and saying "That's not a lion, it doesn't have a mane!" The enrichment operator picks out the k-property *lions have manes* and conjoins it to the intension in the E-structure to get the new intension $\lambda x.\, \textsc{lion}_w(x) \wedge \textsc{has-mane}_w(x)$, enforcing that all lions have manes in the intended meaning of *lion* in this sentence.

[28]The original formulation is also somewhat ambiguous about how many dimensions it can target, but later examples in del Pinal (2018) with *real* make it clear that it can target multiple, not just one.

[29]In fact, if we assume that authenticity is a t-property of *watch*, we can get the desired reading of *watch* as *designer watch* which we need for subsective *fake watch* by enriching *watch* with this authenticity t-property, in the same way as the enrichment derivation for *artificial reef* below.

[30]This is accomplished trivially with del Pinal's four qualia dimensions by selecting just the perceptual dimension P, but requires a collection of appropriate perceptual properties for our property-based C-structure.

Unlike previous accounts of free contextual modulation (Lasersohn, 2012; Pagin and Pelletier, 2007; Recanati, 2010; Szabó, 2012), mod_l is constrained to only widen or narrow along properties provided by the C-structure. There is no way to widen *lion* to mean *giraffe*, since *lion* does not include enough giraffe-dimensions beyond e.g. being a mammal.[31] Exactly how constrained this modulation operator is depends on what properties are in our C-structure. But by virtue of them having to be k-properties of the noun, which must hold in all sufficiently normal worlds, we already rule out a lot of unattested meanings.

We can now explain our two *artificial reef* examples in (52) using these tools. In both cases, we get a privative inference in general, just as discussed for *fake*, since we assume that *artificial* is targeting the property of the reef forming naturally and we assume that that is a k-property of *reefs*. However, we can take either Path (P2) (not meeting felicity conditions) or Path (P4) (contextual coercion of the noun) to get a subsective reading in our particular case. Which path we take depends on the meaning we assign to *reef*.

One option is to take (52a) to involve the default definition, in which case (52b) involves a narrowed definition that insists on the reef forming naturally. In this case, forming naturally is a k-property of reefs, but our particular reef does not meet the felicity constraints for it. mod_l returns an enrichment operator for (52b) that conjoins this property with the original E-structure and forces it to hold, such that *reef* no longer refers to the artificial reef.

The other, mirror option is to view (52b) as using a default, narrow definition of *reef* which has forming naturally as a k-property and (52a) as involving a widened definition that drops this k-property. "Dropping" this k-property to have *reef* refer to the artificial reef anyway is achieved by having mod_l return a k-property operator in (52a) that replaces the intension in the E-structure (which requires forming naturally) with some conjunction of reef k-properties that does not include the natural formation k-property.

[31] Technically, mod_l does allow the widening of *lion* to mean *mammal*, if *lions are mammals* is a k-property of lions. However, this is only possible in a context where this makes sense; it is hard to think of a context where just saying *mammal* would not be a better choice than widening *lion* to mean *mammal*.

3.6.2 Discourse context resolves underspecification of inferences

Adjectives modify or add particular properties to the noun. In most cases, these changes do not impact the properties crucial to the noun, and so we get a subsective inference, absent any other information. However, a lot of properties, including properties that might have an impact on noun-hood, are left underspecified in this process. For example, *counterfeit* merely states that the method of manufacture is non-authentic, but not exactly how. The examples (53a) and (53b) from Experiment 3 show how the context can fill in what happened in the counterfeiting process to swing the inference one way or another.

(53) a. Context: A factory that produced both Poland Spring and Crystal Geyser water bottles realized at the end of the work day that they'd bottled both types of water using the spring water used for Crystal Geyser. Instead of making everyone stay late, they shipped it anyway, thinking no one would notice the difference. However, some dedicated Poland Spring fans quickly realized the taste was different, and were very angry that they had been sold counterfeit spring water.
∴ This is spring water.

b. Context: A factory producing bottled spring water found out that the spring they were using suddenly dried up. The manager panicked because they had promised to ship a lot of water that day, and if they didn't, everyone would likely lose their jobs. So the manager decided to just replace the spring water with tap water, and continue selling the water until they found a new spring. Before they were able to find one, the factory got in huge trouble for selling counterfeit spring water.
∴ This is not spring water.

Since correct branding is not essential to (bottled) spring water (i.e. not a k-property), we derive a subsective inference for *counterfeit spring water* in sufficiently normal worlds such as the actual world. How do we get the privative inference? The privative inference arises simply because the context asserts that the stuff sold was tap water, not spring water. In fact, the question is why we can describe this stuff as *counterfeit spring water* at all, if *counterfeit*

spring water is generally *spring water*. We can appeal to the notion of noun modulation introduced in 3.6.1 and say that the meaning of *spring water* in the bigram *counterfeit spring water* is contextually modulated in (53b) to a widened definition that only preserves properties to the effect of "has the appearance of spring water". This essentially brings back a widening account of privativity, but only in cases where the "default" inference is subsective (that is, not for *counterfeit money* or *stone lion*). Since it is context causing the inference to be privative, asking context to widen the noun is more motivated than for *stone lion*. Context widens the noun because it would be a contradiction to describe this discourse referent x_0, described to be tap water and not spring water, as (*counterfeit*) *spring water* (that being subsective). Some looser definition of spring water must be meant. In this case, *counterfeit* still yields a subsective inference over the loosened meaning. The privative inference derives from the fact that x_0 is asserted in Sentence 3 to be tap water and that tap water is not spring water.

One might think that a second way out of this issue would be to appeal to the fact that *counterfeit* involves negation in its meaning, and that this negation somehow causes the underspecification. In fact, (54a) and (54b) show that we can get exactly the same pattern for the material adjective *plastic*, which is not typically taken to involve negation.

(54) a. Context: A company produces a watch that is made entirely out of plastic, right down to the hands and internal mechanisms. The watch functions like traditional watches but is much cheaper to make.
∴ This is a watch.

b. Context: A children's toy company makes a plastic watch to use during dress-up. It's molded from a single piece of plastic and doesn't move, with the hands always pointing to the same fixed time. John buys one for his daughter as a toy. When John goes to work, he mistakenly grabs the plastic watch instead of his own watch. John is frustrated when he realizes he can't tell the time and nearly misses his bus.
∴ This is not a watch.

This concludes our implementation of Generalization (G2).

3.7 Conclusions and contact with other phenomena

3.7.1 A new theory of adjective-noun modification

Section 3.3 proposes a new theory of adjective modification that incorporates privativity as a matter of course in the guise of potential meaning contradictions, and puts the burden on the modifier (the adjective) to resolve this. We thus directly capture Generalization (G1), that privativity occurs simply when adjectives modify crucial properties, and its Corollary (C1), that any adjective can in principle be privative. Unlike previous theories of privativity, we do not need a repair mechanism such as noun widening to explain privativity in expressions like *stone lion* and *counterfeit money*, and we do not need to invoke context at all for these inferences. We do invoke context for cases where context affects what inference is derived for a given adjective-noun bigram. However, this is constrained: the context can modulate the noun meaning only along properties of the C-structure and does this only in cases where a non-standard inference is derived by virtue of a non-standard context, such as a privative inference for *counterfeit spring water* even though *counterfeit* and *spring water* are not inherently contradictory. This mechanism of noun modulation is entirely separate from adjective-noun modification and its mechanism to derives privativity, capturing Generalization (G2).

We also invoke context for *fake*, but again in a constrained fashion: the context only specifies which of a noun's k-properties *fake* picks to negate. The fact that C-structure contains k-properties and that *fake* targets k-properties by picking a property from the C-structure implements Generalization (G3). Moreover, the Genericity Constraint makes clear the tight relationship between k-properties and generic sentences. By restricting *fake* to the C-structure, we do not hit the overgeneration issues for more context-dependent accounts of *fake* (Guerrini, 2024a; Partee, 2010; Reinert, 2024) that we will discuss next in Section 3.5.1. Our theory also neatly explains predictions for uses of *literally* with *stone* and *fake*, discussed in Section A.3.

We improve upon del Pinal (2018), whose dual content semantics we build on, by removing privativity from the adjective lexical entry and adding a mechanism of conflict resolution

instead. This allows us to account for *stone lion* and *counterfeit money* in the same straightforward fashion; del Pinal's lexical entry for *steel* is intersective and would result in an unresolved contradiction for *steel lion*. By replacing his qualia dimensions with k-properties, we are also able to explain *fake*'s particular sensitivity to k-properties; del Pinal's $\neg Q_A$ component of *fake* (negating the origin of typical instances of the noun) might in principle entail negating some k-property, but this is unclear at best. Further, although del Pinal (2018) proposes his account of privative inferences and his constrained contextual modulation operator mod_l in the same paper, he does not combine the two. Our account shows how mod_l can naturally be extended to explain why inferences can be subsective or privative in different contexts, by virtue of the noun meaning being modulated by the context.

3.7.2 Conceptual combination

This dissertation approaches adjective-noun modification primarily from a linguistic perspective which is concerned with inferences and uses a referential view of semantics that assumes extensions and possible worlds. In doing so, we were obliged to also connect to a notion of concepts, since both adjectives and nouns are content words associated with rich concepts, some kind of reference to which is necessary to explain the behavior that we saw. There is a great deal of literature about conceptual combination in the theory of concepts that does not take an extensional perspective (see Murphy, 2002, Chapter 12 for an overview). In no small part, this is because extensions are a poor way of modeling concepts, as reflected in our difficulty here in bridging the gap between E-structure and C-structure and our reliance on the complex semantics of generics. The account in this chapter implements conceptual combination and the resulting inheritance of noun properties, including conflict resolution between adjective and noun properties, using the familiar (to semanticists) tool of entailment, which allows us to derive further entailments that we are interested in for our specific data. Ultimately, this dissertation still takes del Pinal's view that the C-structure—and, by extension, how it composes—should be a bridge to the theory of concepts and should represent "our best psychological theory". The theory proposed here is certainly not intended to

provide a conclusive theory of how adjective concepts modify noun concepts at every level of detail which would explain all the typicality effects and other effects studied by previous work on conceptual combination (Murphy, 2002; Smith, Osherson, et al., 1988 i.a.). The actual process of adjective-noun modification at the conceptual level is undoubtedly richer and more complex than the (somewhat) toy examples given here. The contribution of this chapter is to specify some behaviors that need to happen at the interface between language and concepts in order to explain phenomena around privativity that have previously primarily drawn the attention of linguists, and to provide at least an outline of how they should occur in terms that make sense to a semanticist.

As an illustration, we compare our account to the influential model of adjective-noun conceptual combination proposed by Smith, Osherson, et al. (1988). Their model assumes that adjectives target fixed dimensions of nouns, such as a color dimension for color adjectives, similar to del Pinal' assumption that nouns have a fixed set of labeled dimensions which adjectives modify. This assumption is criticized by subsequent work since it is difficult to find a dimension for many adjectives, such as *corporate* (Murphy, 2002). It also does not account for non-intersective adjectives or adjectives like *corporate* whose meaning is argued to depend on the noun (*corporate stationery* has a different contribution to *corporate account*, etc.). Our account allows for a varying and unlabeled set of noun C-structure properties (which may be structured in some way beyond an ordered list, but need not be for our account) and views the adjective as just adding another such property; any properties of that sort that the noun previously had will be dropped (i.e. overwritten) if they are contradictory. Further, the meaning contribution of adjectives in our account may depend on the noun (depend on N), making it easy to account for the variability in the meaning of adjectives like *corporate* within the lexical entry if so desired.

A notable point of this dissertation is that there is a significant distinction between conceptual combination in the form of adjective-noun modification and conceptual combination in the form of noun-noun compounds. Adjective-noun modification is a significantly more constrained process: we expect the composition to be as faithful to the adjective and noun

meanings as possible, and many meanings are unattested (see Section 3.5). We also expect adjective-noun composition to be deterministic (given a context), and thus generalization to novel adjective-noun bigrams should be possible (and is possible, as shown in Experiment 2). For noun-noun compounds, there is far more variability and dependence on world knowledge for how the two nouns should combine, as in the much-cited example of *olive oil* vs. *baby oil* (ten Hacken, 2016).

3.7.3 Predicate position

This theoretical proposal focuses solely on adjective-noun modification as an instance of how two content words compose when one modifies the other. Our lexical entry for adjectives is built around the case of adjectives being noun modifiers. Thus, we have given adjective lexical entries of type $\langle s, \langle \langle s, \langle e, t \rangle \rangle, \langle e, t \rangle \rangle \rangle$. Of course, adjectives may also occur in predicate position in many (though not all) languages, including English, in which they should be type $\langle s, \langle e, t \rangle \rangle$.

There remains significant debate over whether adjective types should be homogenous ("generalizing to the worst case", Montague, 1970) or heterogenous, and how to derive one reading from the other to yield a more unified theory of adjectives (Demonte, 2011; Hamann, 1991; Morzycki, 2016). One path is to start with the predicative type $\langle s, \langle e, t \rangle \rangle$ and derive attributive (modifier) readings; the other is to start with the function of properties modifier type and derive predicative readings. The typical way that function/attributive theories of adjectives handle predicate position is by assuming that predicate position involves a silent, contextually specified noun (Montague, 1970; Parsons, 1972), or a related context-dependent strategy that "sharpens" the meaning of the adjective to no longer rely on a noun (Kamp, 1975). Meanwhile, predicative theories of adjectives use a special composition rule of predicate modification to allow predicative adjectives to combine with the noun.

Siegel (1980) shows that Russian long- and short-form adjectives can help clarify this debate. Russian long-form adjectives have an additional morpheme compared to their corresponding short form, and may occur in predicative or attributive (modifier) position, while

short-form adjectives may only occur in predicative position. (Examples from Matushansky, 2008 and Siegel, 1980.)

(55) a. xorošaja teorija
good-LONG theory

b. *xoroša teorija
good-SHORT theory

(56) Zimnie noči budut dolgimi / dolgi
winter nights will.be long-LONG / long-SHORT
'The winter nights will be long.'

Morzycki (2016) summarizes Siegel's conclusions as follows: (1) Short-form adjectives seem to be property-denoting, while long-form adjectives appear to be modifiers. (2) Since these property-denoting short-form adjectives cannot occur in modifier position, we might not want a rule of predicate modification, as this would allow them to occur there. Requiring modifiers to be a functional type would explain the ungrammaticality of short-form adjectives in this position. This supports accounts like ours, which do not use predicate modification. (3) The fact that long-form adjectives can appear in predicate position despite being modifier-denoting supports the idea of having silent nominals in certain predicate position, which the adjective then combines with.

This third point is further supported by a difference in readings between the long and short form in predicate position observed by Siegel:

(57) a. studentka umna
student intelligent-SHORT
'The student is generally intelligent.'

b. studentka umnaja
student intelligent-LONG
'The student is intelligent as a student.'

The long form *umnaja* appears to be combining with (a silent copy of?) *student*. Moreover, when no such noun is contextually salient, the long form is marked:

(58) a. Oleg umen
Oleg intelligent-SHORT
'Oleg is intelligent in general.'

b. Oleg umnyj
Oleg intelligent-LONG
'Oleg is intelligent as a ...what?'

As we discussed in Section 3.3.1, our new theory can be extended to account for predicate position by saying that the conflict resolution component of adjectives that makes adjectives into modifiers of nouns is not part of the lexical entry of the adjective root, and is instead provided at the locus of modification, for example by a Distributed Morphology categorizing head a_M that is specific to modification (Halle and Marantz, 1993). In this case, roots whose meaning contribution (f_C in schema (31)) is of type $\langle s, \langle e, t\rangle\rangle$ and does not depend on the noun could combine with a different categorizing head a_P in predicate position that just returns f_C as-is.[32] This is similar to the suggestion in Siegel (1980) that there is a transformation in the lexicon that licenses predicative adjectives in modifier position. Adjective roots that do depend on a noun can also be available in predicate position if they can combine with a_M plus a silent noun provided by the context, as we discussed for *counterfeit*.

While it may seem awkward to posit silent material, doing so allows us to explain context-sensitive behavior such as the context-sensitivity of *skillful* (Martin, 2022; Reinert, 2024) as well as why adjectives like *counterfeit* and *intelligent* in Russian are only sometimes felicitous in predicate position. Since a bleached noun like *entity* cannot combine with (or does not resolve the ambiguity of) *counterfeit* or *intelligent*, we predict these adjectives to be felicitous precisely when context (such as the subject of the sentence) provides an appropriate noun.

[32]Roots that do not have an appropriate meaning contribution should fail to combine with a_P by virtue of lacking an exponent for it, this meaning being impossible to expone.

Kennedy (2012) suggests that we have further evidence for this approach since some languages like Vata (Niger-Congo) genuinely seem to lack predicative adjectives (Baker, 2003) and use sentences with overt, semantically bleached nouns to express sentences like (57a). More broadly, there is significant cross-linguistic and intra-linguistic variation in this space: other languages like Slave (Athabaskan) appear to lack attributive adjectives (Baker, 2003), and even within English there are adjectives like *alive* and *remiss* which can only be used predicatively, while adjectives like *alleged* can only be used as attributively (Morzycki, 2016). In other words, it may be necessary to have a significant level of type variation and stipulation in the lexicon, and it may not be possible to have unified single-type cross-linguistic theory. This dissertation proposes a unified theory of adjective-noun modification that cuts across adjective classes[33] and can be transparently extended to handle predicate position and the associated item-by-item variation involved, including the Russian data.

This theory is intended to be compatible with existing accounts of (morpho-)syntactic variation in adjective-noun composition across languages (Cinque, 2010; Martin, 2022; Tat, 2011 i.a.): the theory of this chapter applies whenever these accounts propose that an adjective combines with a noun in a modifier relation, and explains *how* this modification occurs, regardless of how the syntax achieves that (overt noun, covert noun, reduced relative clause or otherwise). We expect *stone lion* and its kin to be privative in any language that has appropriate material adjectives, by virtue of how these concepts modify each other, though a language need not lexicalize any particular example we use here.[34] It is also intended to be compatible with existing accounts that study the specific meaning and/or context-dependence of adjectives, such as accounts of gradable adjectives (Kennedy, 2007; Kennedy and McNally, 2005; Klein, 1980) or the context-dependence of *skillful*-class adjectives (Maienborn, 2020;

[33]While we do not discuss nonsubsective adjectives like *alleged*, it should be possible to extend the theory to them once their core meaning is properly established; they will simply not use the conflict resolution mechanism in the actual world.

[34]English is a convenient language to study because it lexicalizes a range of material adjectives as well as both *real* and *fake*. By contrast, German lacks an analogue for English *fake* to such an extent that English *fake* is now being borrowed into German (p.c. Carla Spellerberg), although it can express a similar concept using *nicht echt* ('not real').

Reinert, 2024) or color adjectives (Guerrini, 2024c; Kennedy and McNally, 2010). The meanings arrived at can be slotted into the "core" meaning in our adjective schema (with a slight adjustment in typing for gradable adjectives).

Lastly, drawing this distinction that some adjectives may occur in predicate position if their core is type $\langle s, \langle e, t\rangle\rangle$, while others possess a more complex core and may not, essentially recovers the intersective/subsective distinction that many semanticists use to classify adjectives, as discussed in Section 3.3.1 when we discussed subsective and intersective meaning postulates. We can read the class of the adjective off the type of its core meaning and/or its ability to combine with the predicative categorizing head a_P, and further signal it by attaching the appropriate meaning postulate. The point of Generalization (G1) is simply that "privative" is not among the classes we can distinguish: intersective, subsective and nonsubsective are still distinct and useful classes.

3.7.4 Adverb-verb modification

A second area outside the noun phrase in which this chapter might have applications is adverb-verb modification. Under an event semantics view, verbs are also content words that describe a primitive in our semantics—events instead of entities—and also have rich conceptual structures (Casati and Varzi, 2008). It should be possible to extend the same notion of modification and conflict resolution via C-structure to adverb-verb modification, simply by replacing type e with v in all our schemas and constraints.

In fact, cases of privative verb modification do exist, especially with *falsely*. While we have mainly left discussion of *false* to Section A.1, one of the senses of *false* is that the propositional content that the noun refers to is false, as for *false rumor*, *false statement*, *false proof* etc. Intuitively both (59a) and (59b) have the same truth conditions and privativity inferences, and should be captured in the same way.

(59) a. Sam had a false proof of the theorem.
∴ Sam did not have a proof of the theorem.

b. Sam falsely proved the theorem.

∴ Sam did not prove the theorem.

We even see some productive use of *fake* as an adverb, at least in some colloquial English, as these two examples from online forums attest:[35]

(60) Context: Caption of a screenshot of a post on social media platform Reddit about someone complaining after being fired.
I'm sure these things happen but it's written in such a fakely Reddit way.

(61) Context: Post about FACEIT, a gaming platform.
Someone fakely requested to deactivate my account and FACEIT MIA just went ahead and did it.

The one issue with extending this account to adverb-verb modification is that not all parts of the sentence need have a C-structure. If the adverb is adjoined to the VP or vP and not directly to the V head, then the verb first combines with its complement DP before combining with the adverb. DPs usually contain a function word as their head. Function words, as discussed in Section 3.2.2, most likely have no C-structure of their own. If this has the effect of ending the composition of C-structure (as opposed to passing the C-structure through unmodified), then the C-structure of the VP is not accessible to the adverb for it to modify. Depending on the locus of adverb adjunction, this would to be resolved. I am not aware of any previous work in privativity in adverb-verb modification, and there is much exciting research to be done to extend this account (or any other suitably constrained account of privativity) to this domain.

[35] Sources:
https://www.reddit.com/r/AmITheAngel/comments/1iq4wye/im_sure_these_things_happen_but_its_written_in/
https://www.reddit.com/r/FACEITcom/comments/xt2smu/someone_fakely_requested_to_deactivate_my_account/

This ends our discussion of the theory of adjective-noun modification and what we can learn from privativity about how content words like adjectives and nouns compose. In the remainder of this dissertation, we will turn to investigating generalization to novel adjective-noun bigrams, first by extending Experiment 2 that we conducted in Chapter 2 on humans to LLMs in Chapter 4. Chapter 5 then investigates an analogical model of adjective-noun inferences that could also explain generalization as an alternative to composition.

Chapter 4

Is artificial intelligence still intelligence? [1]

Preface

We will begin this chapter by asking whether modern large language models (LLMs) are able to derive human-like subsective and privative inferences for adjective-noun combinations and generalize to novel combinations. As we motivated in Chapter 1, this is not just an interesting question about LLMs, but is also useful for linguistic purposes for two reasons, which we will lay out in detail in Section 4.8.1 and Section 4.8.2 at the end of this chapter.

Firstly, attempting to establish evidence of compositionality in other language-using systems besides humans can give us clearer standards of how to evaluate humans. Much of the design in Chapter 2, such as controlling for frequency / memorization and using such a large set of nouns, was motivated in part by knowing that confounds such as frequency and item-by-item variability would need to be ruled out for LLMs. Further, as we will discuss in Section 4.8.1, people find themselves hesitant to ascribe things like compositionality to LLMs even though they pass the same behavioral tests that humans do (at least to a reasonable stan-

[1]The main body of this chapter was previously published as Ross, Davidson, and Kim (2024) ("Is artificial intelligence still intelligence? LLMs generalize to novel adjective-noun pairs, but don't mimic the full human distribution"), co-authored with my advisors Kathryn Davidson and Najoung Kim.

dard). This at minimum motivates us design better experiments to test compositionality, and motivates some researchers to redefine what they mean by compositionality or redefine what we mean by evidence for it (McCurdy et al., 2024). Such considerations only emerge once we start asking these questions about sufficiently capable non-human systems; otherwise, it is easy to build definitions of behavior that are either too broad or are only applicable to humans by design (Pavlick, 2025).

Secondly, establishing that LLMs show reasonably human-like behavior on these tasks allows us to use them as models of (parts of) human behavior to test hypotheses that are hard to model in formal semantics alone. While we will introduce Experiment 7 as a way of achieving human-like behavior in LLMs, Section 4.8.2 explains how we can view the same experiment as testing a hypothesis about how context affects human behavior that is hard to break down in our semantic theory.

4.1 Introduction

As the pretraining datasets of LLMs grow, it becomes increasingly difficult to test whether LLMs can generalize to unseen instances of linguistic phenomena, since it is hard to systematically exclude them from the pretraining data (Kim, Linzen, and Smolensky, 2022). Adjective-noun combinations, in particular with so-called privative adjectives like *fake*, provide a good test bed for testing granular language understanding and semantic composition in LLMs, since we can find many adjective-noun bigrams which are easy for humans to understand despite being novel to them, and can further confirm that they are novel to the LLM by a string search over its pretraining corpus. Adjective-noun bigrams are naturally associated with membership inferences which allow us to test whether LLMs have successfully "composed" their meaning: is an {adjective} {noun} still a {noun}? As we laid out in Chapter 2, this inference is not as straightforward as it may seem: while a *yellow flower* is clearly still a *flower* (a *subsective* inference), a *fake* or *counterfeit dollar bill* is typically not a *dollar bill* (a *privative* inference), even though a *fake watch* is typically a *watch*. In order to draw the correct inference, humans and LLMs need to grasp not only the effect of the adjective on the noun's

properties, but also which properties are required vs. merely typical for membership in that noun/category. Moreover, like many cases of meaning, this inference depends on context. For example, a *fake crowd* might qualify as a *crowd* if it is made up of paid actors, but less so if it is just painted dummies on a movie set. Nonetheless, humans derive convergent inferences for many novel bigrams both with and without context, giving a ground truth to which we can compare LLMs.

Building on the dataset of English human judgments developed in Chapter 2, we compare LLMs of several sizes with humans for cases in which the context determines the inference. We further explore three methods to evaluate LLMs against the full distribution of human ratings when there is no provided context. We find that when the context determines the inference, recent larger LLMs closely match human behavior, while smaller LLMs only sometimes do so. Almost all LLMs, even smaller ones, are able to handle novel adjective-noun combinations as well as they handle known ones, suggesting that they evaluate these combinations on-the-fly and can generalize accordingly like humans do. This alignment in performance on observed and novel combinations continues to hold in a setting where no context is provided. However, we find that even 70B parameter LLMs are only able to approximate the distribution of ratings of a population of human raters for 75% of our dataset or less, struggling on combinations with high human variance and inferences which are rare for a particular adjective, such as a *homemade cat* not necessarily being a *cat*. Most LLMs also assign a more positive rating to the question "Is artificial intelligence still intelligence?" than most humans (see Figure 4.5). In sum, while our generalization results are highly promising, there is room for improvement on the task of matching human inferences in this category overall. We share our code and data on GitHub.[2]

[2] `https://github.com/rossh2/artificial-intelligence/`

4.2 Related work

Most previous computational work on adjective-noun composition focuses on distributional semantics using word embeddings (Boleda, Baroni, et al., 2013; Boleda, Vecchi, et al., 2012; Cappelle, Pascal, and Keller, 2018; Guevara, 2010; Hartung et al., 2017; Vecchi et al., 2017). Cappelle, Pascal, and Keller (2018) specifically analyze privative adjective embeddings, finding no difference between their embeddings and embeddings of other adjectives. Results for early LLMs are largely negative: BERT (Devlin et al., 2019) shows no evidence of compositionality for adjective-noun and noun-noun phrases, relying instead on word overlap heuristics for similarity judgments (Yu and Ettinger, 2020). Bertolini, Weeds, and Weir (2022) study the inferences of adjective-noun combinations for BERT and RoBERTa (Liu et al., 2019). They divide adjectives into three inference classes—intersective, subsective and privative—based on previous computational work (Lalisse and Asudeh, 2015; Nayak et al., 2014), and test whether LLMs can draw the correct pattern of inferences for adjectives in each class. However, recent work in linguistics (Martin, 2022; Pavlick and Callison-Burch, 2016b) as well as our results from Chapter 2 suggests that Bertolini, Weeds, and Weir's task may be ill-defined, since adjectives in these "classes" can license either a subsective or a privative inference depending on the noun and context. Indeed, Goodale and Mascarenhas (2023) find that BERT can distinguish between intersective and subsective adjectives, which are more clearly distinct.

For the "Is an X a Y" task more broadly, LLMs from BERT to Llama 2 (Touvron et al., 2023) show some, but limited abilities to judge "Is an X a Y" for hypernyms, especially with negation (Ettinger, 2020; Hanna and Mareček, 2021; Moskvoretskii, Panchenko, and Nikishina, 2024; Nikishina et al., 2023; Ravichander et al., 2020). Results from property learning also show that earlier models struggle to learn what properties are typical of nouns (Apidianaki and Garí Soler, 2021; Do and Pavlick, 2021; Pavlick and Callison-Burch, 2016a)—a key part of our task, since LLMs must know what constitutes a *watch* in order to judge whether a *fake watch* counts as one. Meanwhile, Lyu et al. (2022) find that BERT and GPT-3 (Brown et al., 2020) cannot handle inferences on recursive adjectives, such as "Is my favorite new movie my

favorite movie?", while Sathe, Fedorenko, and Zaslavsky (2024) find that GPT-2 (Radford et al., 2019), MPT 30B (Team, 2023) and other smaller models struggle to predict whether rare adjective-noun combinations are acceptable to humans. However, less is known about the capabilities of newer, larger models in these areas. Recent work with newer models including Llama 2 and GPT-3 on noun-noun compounds—whose meaning arises less straightforwardly from their parts than adjective-noun combination (ten Hacken, 2016)—shows that the LLMs do not generalize well in this case (Coil and Shwartz, 2023; Ormerod, del Rincón, and Devereux, 2024; Rambelli, Chersoni, Collacciani, et al., 2024).

4.3 Human judgment dataset

We can view Chapter 2 as creating two datasets of human judgments on adjective-noun inferences. The first, which we refer to as No-Context, collects human ratings on "Is an {adjective} {noun} a {noun}?" on a 5-point Likert scale ("Definitely not", "Probably not", "Unsure", "Probably yes" and "Definitely yes") for 798 bigrams, covering 102 nouns crossed with 6 typically-privative and 6 typically-subsective adjectives. (In this chapter, we use "(typically) privative / subsective adjective" to refer to adjectives historically classed as such, which often but not always result in the respective inference.) 180 of the 798 bigrams are zero frequency in the C4 pretraining corpus (Raffel et al., 2020), which we take as a proxy for the undisclosed pretraining corpora of the models we study. We take these bigrams to be novel to both humans and LLMs.[3] We call a bigram high-frequency if it is in the top quartile of bigrams studied in Chapter 2.

In Chapter 2, we showed that this inference depends on the adjective and noun, with bigrams with "subsective" adjectives usually (but not always; e.g., *homemade cat*) being rated subsective, while bigrams with "privative" adjectives such as *fake crowd* elicit a wide distribution of ratings from privative to subsective, with high variance per bigram. Moreover, humans converge on inference judgments for many zero-frequency bigrams, and show similar

[3]The dataset was not published when these models were trained, so there is no danger of it being included in model pretraining.

variance overall between zero-frequency and high-frequency bigrams, demonstrating that they can generalize these inferences.

The second dataset from Chapter 2, which we refer to as CONTEXT, shows that providing a context with appropriate detail is sufficient to determine the inference for typically-privative adjectives. Participants first read a short 50-word context and then answer "In this setting, is an {adjective} {noun} a {noun}?" The dataset contains 56 expert-written contexts for 28 bigrams, with one privative-biased and one subsective-biased context for each bigram. 6 bigrams are zero-frequency in C4 and a further 7 are low-frequency (below median frequency in the total set). This dataset is much smaller due to the need for trained annotators to create the contexts.

4.4 Experiment 5: In-context meaning generalization

Typical LLM evaluations assume a single correct answer for each question. Thus, we begin with the smaller CONTEXT dataset from Chapter 2 where the contexts provided are sufficient to determine the inference. Our evaluation focuses on the following two aspects: (1) whether the LLM is sensitive to the provided context, and (2) whether the LLM is able to choose the intended inference both for high-frequency bigrams it has presumably seen during pretraining and also for zero-frequency bigrams which we presume it has not.

4.4.1 Method

Experiment 5a evaluates the 28 bigrams in CONTEXT with no context provided, using the same setup as Experiment 6. While there is no single correct rating in this setting, we can use this as a baseline to see if providing a context changes the rating. Experiment 5b evaluates the same bigrams but provides the two contexts for each bigram which bias the rating for humans. For both, we adapt the method used in Chapter 2 as closely as possible for LLMs. We use the same question wording, asking "(In this context [5b],) is an {adjective} {noun} still a {noun}?" with the same 5-point Likert scale, the only difference being that the Likert scale is described in words rather than pictured. We provide 5 few-shot examples illustrating each

Table 4.1: *Few-shot prompts for Experiment 1a and Experiment 2, which give one bigram for each value on the Likert scale to demonstrate use of the scale for "is-a" judgments.*

Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", is a green pepper still a pepper? **Answer:** Definitely yes
Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", is a wooden pear still edible? **Answer:** Definitely not
Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", is a small ladder still useful? **Answer:** Unsure
Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", is melted ice still ice? **Answer:** Probably not
Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", is a short basketball player still tall? **Answer:** Probably yes

rating on the scale, shown in Table 4.1 for Experiment 5a and in Table 4.2 for Experiment 5b (same bigrams, but with contexts). These only demonstrate the use of the Likert scale and do not include any typically-privative adjectives.[4] We believe this to be a comparable setup since humans also see examples during the training phase of the experiment. Humans see only the first three examples along with a short explanation of the suggested reasoning, and are encouraged but not required to pick the suggested rating.

To get responses on the Likert scale, we calculate the surprisal of the 5 answers. The model's response is whichever of these 5 answers has the lowest surprisal. This limits us to assessing open-source models where log-probabilities of the input are available, excluding popular closed-source models like GPT-4 (OpenAI et al., 2024). We study the Llama 2 (Touvron et al., 2023) and Llama 3 series[5] (Dubey et al., 2024) in detail, as well as Mixtral 7x8B

[4]Experiment 5c in Section B.2.5 performs an ablation study on Experiment 5b with 0-shot prompting.

[5]In this chapter, Llama 3 refers to the original Llama 3 models, not the newly released Llama 3.1 models.

Table 4.2: *Few-shot prompts for Experiment 1b and 3b, which give one bigram/context pair for each value on the Likert scale to demonstrate use of the scale for "is-a" judgments. For the chat/instruction-tuned models, we format the few-shot prompt as a conversation between the assistant and the user, where each context and question is provided by the user and the assistant provides each answer (without the "Answer" prefix). For the base models (discussed in Chapter B), we concatenate the few-shot examples and use the "Answer" prefix to indicate the answer.*

Context: Sarah asks Leo to go to the store to buy a bell pepper. When he gets there, he realizes she didn't say which color pepper he should buy. He buys a green pepper. When he gets home, Sarah is disappointed, because she prefers the red ones.
Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", in this context, is the green pepper still a pepper?
Answer: Definitely yes

Context: Mark is an expert carver and carves a highly realistic pear out of dark colored wood. He hides the wooden pear in his fruit bowl among the fruit he bought from the supermarket.
Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", in this context, is the wooden pear still edible?
Answer: Definitely not

Context: Bob has climbing roses growing all up the side of his house, and wants to trim them for the first time. He needs to find a way to reach the roses higher up. He looks in his shed to see what he has and finds that he has a small ladder, which he can use to reach the roses halfway up the house, though not the ones at the very top.
Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", in this context, is the small ladder still useful?
Answer: Unsure

Context: Sam asks Carla to go to the store to buy ice for drinks for their party. Unfortunately, she leaves it in her car all day and comes back in the evening to find that it has all melted. Carla doesn't know what to say to Sam about the melted ice, which he was planning to use in their cocktails.
Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", in this context, is the melted ice still ice?
Answer: Probably not

Context: Jordan's friend is on the high school basketball team, and is the tallest among her friends. At the match, Jordan notices that her friend is actually a short basketball player, as most of the other players are taller than her.
Question: On a scale of "Definitely not", "Probably not", "Unsure", "Probably yes" or "Definitely yes", in this context, is the short basketball player still tall?
Answer: Probably yes

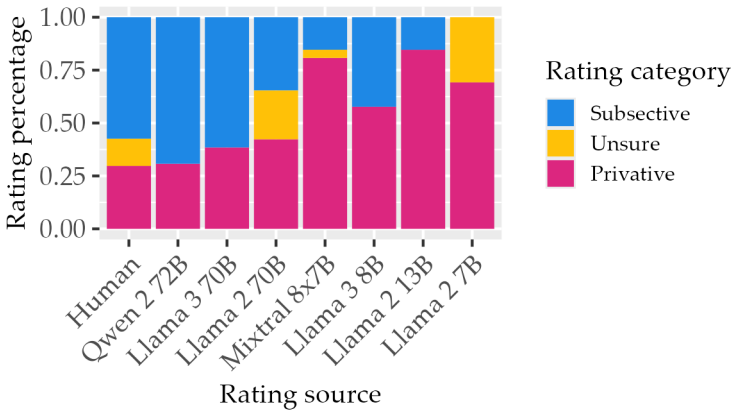


Figure 4.1: *Percentage of privative vs. subsective inferences for bigrams in Experiment 5a (no context) for each model ("Instruct/Chat" omitted for brevity).*

(Jiang et al., 2024) and Qwen 2 72B (Yang et al., 2024). We test all sizes of Llama 2 and Llama 3 to investigate whether generalization ability improves with model size. We primarily focus on instruction-tuned models; results for the base Llama models are shown in Section B.2.

4.4.2 Results

Firstly, to measure whether the provided biased contexts have a significant effect on inference ratings, we compare the results from Experiment 5a and 5b. Figure 4.1 shows the proportions of bigram ratings which are privative ("Definitely not" or "Probably not"), subsective ("Definitely yes" or "Probably yes") or neither ("Unsure") when the LLM (or human; 12 ratings/bigram for humans) rates these 28 bigrams without context in Experiment 5a. The exact proportions are shown in Table B.1. Each LLM brings its own bias: Llama 3 70B is evenly balanced, while Qwen 2 72B favors subsective ratings and Mixtral 8x7B favors privative ratings. For humans, all of these bigrams have high variance when rated in this out-of-context setting, which these percentages do not reflect.

We fit the same ordinal regression as we did in Chapter 2, `Rating ~ ContextBias`, in R (Christensen, 2022; R Core Team, 2023). While in Chapter 2 we found that for humans, both the privative and subsective contexts have a significant effect on ratings compared to rating the bigram with no context, we find that this is not the case for all LLMs. For all LLMs, we

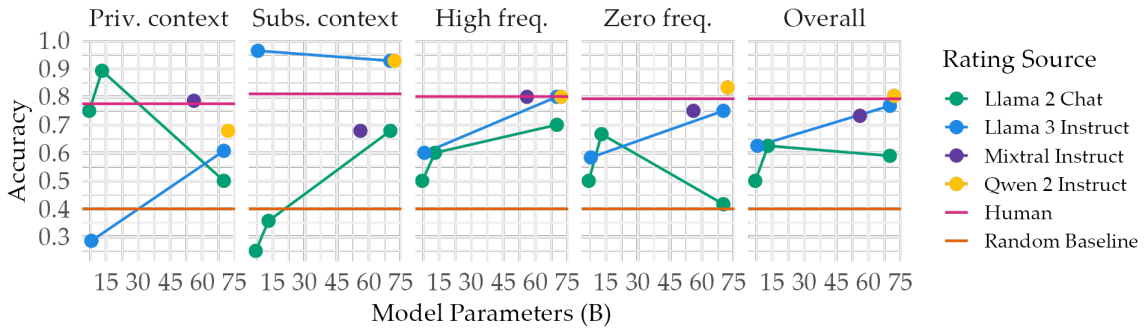


Figure 4.2: *Accuracy on the context-based inference task (Experiment 5b) overall, in privative vs. subsective contexts, and for high frequency vs. zero frequency bigrams. Accuracy on the context-based inference task increases with model parameters for all models except Llama 2 Chat, and all models except Llama 2 70B Chat can generalize to (perform similarly or better on) zero frequency (novel) bigrams.*

find that subsective contexts have a significant effect ($p < 0.05$) compared to providing no context in Experiment 5a. Privative contexts only have a significant effect ($p < 0.05$) for those models which rate many bigrams with typically-privative adjectives as subsective without context (see Figure 4.1), namely Llama 3 70B Instruct, Llama 3 8B Instruct and Qwen 2 72B Instruct.

Next, we judge the inference as correct if the rating is "Definitely/Probably not" in privative-biased contexts and "Probably/Definitely yes" in subsective-biased contexts. Figure 4.2 shows the accuracy for all language models under this metric, plus a random guessing baseline. The human results in Figure 4.2 should be viewed as a ceiling measuring the effectiveness of the context at fixing the inference, not human competence at the task.

We see that Llama 3 70B Instruct, Qwen 2 72B Instruct and Mixtral 7x8B Instruct perform similarly to humans on this task, suggesting a good ability to (a) understand the effect the context has on the thing described, (b) understand what is necessary to count as an instance of each noun, and (c) draw the correct inference based on the previous two steps. For all models except Llama 2 70B Chat, we see no difference between their accuracy on high-frequency bigrams and zero-frequency ones.

Likewise, performance scales with model size for all models except Llama 2 70B Chat. This is because this metric penalizes use of the "Unsure" rating, which Llama 2 70B Chat

often uses (see Figure 4.1), and uses more often for the 6 zero-frequency bigrams. Llama 2 13B Chat scores higher because it is more confident, even though it is sometimes confidently wrong. (Under the softer metric of "accuracy within 1 SD of the human mean", which we will introduce in Section 4.5.1, performance does indeed scale with model parameters – see Figure B.1 in the Appendix.)

4.4.3 Discussion

While model performance on judging inferences of adjective-noun combinations given a context improves with scale, we see that almost all models behave similarly for high-frequency and zero-frequency bigrams, despite presumably never having seen the zero-frequency bigrams before (or at least despite the substantial frequency gap). Thus, we conclude that models do not handle this task by memorizing inferences of noun membership during pretraining and instead assess it dynamically, thus being able to generalize (whether this process is a case of genuine adjective-noun meaning composition, as humans are presumed to be able to do in linguistic theory, or some other heuristics). While this is an exciting result, this is a rather small dataset. In the next section, we study the larger No-Context dataset from Chapter 2.

4.5 Experiments 6 and 7: Evaluating inferences without context

The No-Context dataset in Chapter 2 asks the same inference question "Is an {adjective} {noun} a {noun}?", but without providing any additional context to help determine the inference. This results in a wide distribution of human ratings for some (but not all) bigrams involving typically privative adjectives. For example, a *counterfeit dollar bill* is never judged to be a *dollar bill*. The distribution also widens for some bigrams with typically subsective adjectives, such as *homemade cat*. In this section, we propose three methods to investigate whether LLMs can match the distribution of ratings provided by humans in this context, since there is no longer a single correct answer.

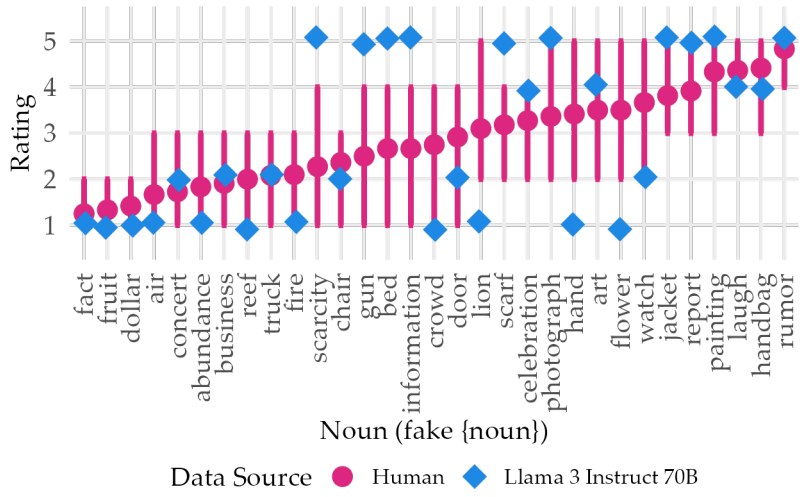


Figure 4.3: *Ratings for select bigrams involving* fake *for Llama 3 Instruct 70B, compared to the (rounded) 1 SD interval around the human mean.*

4.5.1 Method 1: Accuracy within 1 SD

The first method evaluates a single judgment derived from an LLM and asks: Is the LLM sampling from the same distribution as the humans for each bigram? A quick but coarse metric to assess this is whether the LLM's rating falls within one standard deviation of the human mean, rounded to the nearest integer rating (1 corresponds to "Definitely not", 5 to "Definitely yes").[6] This method is intuitive and easy to compute, but is also a relatively low bar. To illustrate, Figure 4.3 shows the human means with intervals 1 SD wide for a selection of bigrams of the form *fake* {noun}, with the no-context ratings from Llama 3 Instruct 70B superimposed. For *fake* in particular, these intervals are relatively wide and easy for the LLM to land in, although for subsective adjectives like *useful*, these intervals are much smaller. Conversely, subsective inferences are by far the most common inference, so we still expect high performance for subsective adjectives.

Experiment 6 obtains single ratings for each of the 798 bigrams in No-Context using the same setup as Experiment 5a (Section 4.4). Figure 4.4 shows the scores on this metric,

[6]One alternative would be to compute the inter-annotator agreement (IAA) between the LLM and the other, human annotators, but existing metrics for IAA either calculate the agreement between two annotators or the agreement across the whole group.

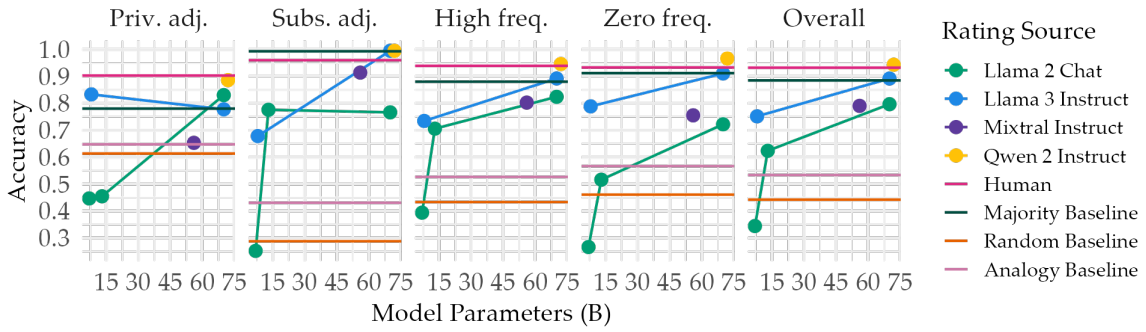


Figure 4.4: *Accuracy within 1 SD of the human mean on the no-context inference task (Experiment 6) overall, for typically privative vs. subsective adjectives, and for high vs. zero frequency bigrams. While accuracy is high, a simple "majority" baseline nearly saturates this metric.*

split by adjective type (typically privative or typically subsective) and bigram frequency. We also compare to three baselines: random, "majority", and analogy. For the random baseline, we sample 100 ratings from 1–5 for each bigram , calculate whether it is within 1 SD of the human mean for that bigram, and average the results. The "majority" baseline guesses a fixed rating depending on the adjective's underlying category as typically subsective or typically privative. Bigrams with subsective adjectives are rated "Definitely yes" (5), while bigrams with privative adjectives are rated "Unsure" (3), taking advantage of the fact that privative adjectives tend to have wide 1-SD intervals that often overlap with 3.

Finally, the analogy baseline attempts to calculate the inference by analogy, imitating reasoning such as "a fake watch is a watch, and a handbag is an expensive accessory like a watch, so a fake handbag must also be a handbag". Specifically, it uses the distance between GloVe embeddings (Pennington, Socher, and Manning, 2014) to find nearby adjectives and nearby nouns among all the adjectives and nouns used in Chapter 2, assembling those into "nearby bigrams". It then averages the human inference ratings among those nearby bigrams which are high-frequency (assumed "known") to predict the rating of the new bigram. We explore analogy in more detail in Chapter 5.

We find that most models, with the exception of Llama 2 7B Chat, perform well under this metric, but the high performance of the majority baseline highlights the leniency of this metric. In fact, only Llama 3 70B Instruct manages to outperform the majority baseline. That

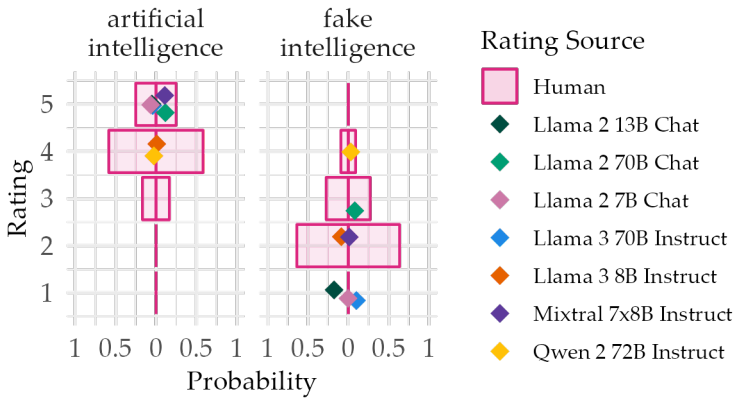


Figure 4.5: *Ratings for "Is artificial/fake intelligence still intelligence?", showing the distribution for humans and the single rating (with no context provided) for LLMs. Most instruction-tuned LLMs give a more confident (higher) rating than humans for* artificial intelligence.

said, the gap between the analogy baseline and the larger models suggests that models are doing something more sophisticated than simple analogical reasoning to high-frequency inferences that they have seen before. This is further borne out by the models' high performance on zero-frequency bigrams, as in Experiment 5b, showing again that models seem to handle this task on-the-fly rather than relying on having seen the combination during pretraining. As in Experiment 5b, performance scales with size.

For *artificial intelligence* specifically, we find that most LLMs answer "Is artificial intelligence still intelligence?" with "Definitely yes", while humans tend to prefer a more conservative "Probably yes", as shown in Figure 4.5.

While this method is appealingly simple, performance on this metric is close to saturated by the majority baseline, making it difficult to evaluate whether LLMs are performing in a "human-like" way. We next explore two methods of generating a distribution of ratings from an LLM, to see if LLMs can capture the whole human distribution rather than merely capturing a point within it.

4.5.2 Method 2: Log-probability distribution

Method Our second method obtains a distribution of ratings from the LLM by calculating the log-probabilities of all 5 answers for each of the 798 bigrams in No-Context in Experiment 6 and converting this into a probability distribution for each bigram. For each bigram, we calculate the Jensen-Shannon divergence between the distribution of ratings obtained from the LLM and the distribution given by the (normalized) human ratings.[7] 0 indicates perfect overlap, while 1 indicates maximal divergence.

Table 4.3: *Jensen-Shannon divergence between per-bigram rating distributions for humans and LLM log-probabilities, for privative vs. subsective adjectives.*

	JS Divergence		
Model	Priv.	Subs.	Total
Human	0	0	0
Llama 3 70B Instruct	0.26	0.08	0.17
Qwen 2 72B Instruct	0.33	0.08	0.19
Llama 2 70B Chat	0.18	0.25	0.22
Mixtral 7x8B Instruct	0.32	0.13	0.22
Llama 3 8B Instruct	0.18	0.34	0.26
Llama 2 13B Chat	0.25	0.35	0.30
Uniform baseline	0.20	0.46	0.34
Llama 2 7B Chat	0.29	0.46	0.38
"Majority" baseline	0.71	0.12	0.40

Results Table 4.3 shows the average Jensen-Shannon divergences, including a uniform distribution baseline and the "majority" baseline reported in Section 4.5.1. Llama 3 70B Instruct shows the lowest average Jensen-Shannon divergence across all bigrams using this method, with an excellent divergence of just 0.08 on (typically) subsective adjectives, matched by Qwen 2 72B. However, not all models are able to concentrate enough of their log-probability mass on "Definitely yes" for bigrams with subsective adjectives, such as Llama 2 70B Chat, even though all models were able to rate these bigrams as subsective when giving a single

[7] While calculating the Kullback-Leibler divergence would also let us treat the human ratings as ground truth, we prefer the Jensen-Shannon divergence because it is bounded between 0 and 1 and thus easier to interpret.

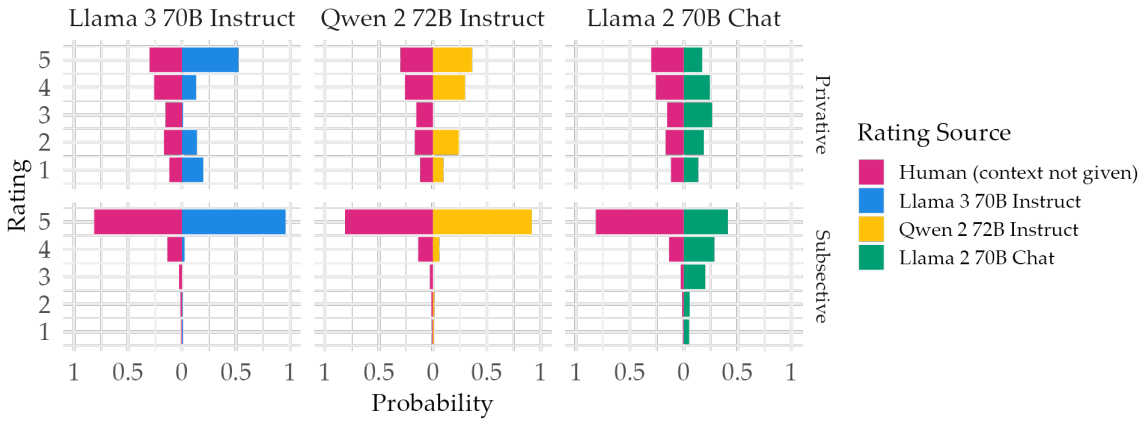


Figure 4.6: *Average log-probability distribution for (typically) subsective vs. privative adjectives for selected LLMs, compared to the average human distribution.*

rating (Table B.4). Further, all models, especially the larger ones, struggle with subsective-adjective bigrams that humans rate as somewhat privative, such as *homemade cat* or *illegal currency*. All models except Qwen 72B Instruct rate *homemade cat* as mostly subsective (5), whereas humans' ratings are distributed evenly from 1-4, and Qwen does not assign enough enough probability mass to these intermediate ratings. Figure 4.7 shows the distribution for *homemade cat* for Llama 3 70B Instruct.

For bigrams with privative adjectives, where the inference is much harder to predict, results are still promising, but there is room for improvement. Figure 4.6 shows that although Llama 3 70B Instruct has the lowest overall JS divergence for privative adjectives, it rates them as subsective (5, "Definitely yes") too often, and neither it nor Qwen 2 72B make human-like use of the "Unsure" rating (perhaps a side-effect of their helpfulness training). Section 4.5.4 discusses the item-by-item variation in JS divergence in more detail.

More broadly, it is not clear that model log-probabilities *should* map onto a distribution of how frequent different answers are among humans. Method 3 investigates a more sophisticated way of getting a distribution of ratings from LLMs which may map more closely onto the human distribution.

4.5.3 Method 3: Context generation

One source of variation in human ratings is that different humans are likely imagining different instantiations of the bigram, loosely corresponding to the different contexts in Experiment 5 (though likely much less well specified), depending on the priors they each have. They then rate "Is an AN an N?" given that imagined context. While some humans may consider multiple instantiations and form a small distribution which informs their single answer, the distribution of ratings primarily arises from consulting a larger population of humans. A language model, when prompted with no context, instead has a single set of priors and has to estimate the log-probability of each answer given that single set of priors. The next method investigates whether we can improve the fit of the model's distribution by having it generate a set of contexts and rate "Is an AN an N?" given each context, imitating this aspect of the human variation and thought process.

Method Method 3 asks the model to generate 12 different "stories" of 50-100 words involving the target bigram which "describe the {bigram} in detail", giving three of the contexts used as few-shot prompts for Experiment 5b as examples. This is somewhat similar to recent efforts to mimic human survey results by prompting or having the LLM generate personas, then generating data with those personas as context (Argyle et al., 2023; Bisbee et al., 2024; Chan et al., 2024 i.a.). Experiment 7a generates all 12 stories in one chat using a temperature of 0.6. The prompt is shown in Table 4.4. We use three contexts already used in the few-shot prompts for Experiment 5 as example stories to help control the style and level of complexity of the language. We found that providing example stories was much more effective than trying to control the output with detailed instructions. {bigram} and {a/an} are substituted in at runtime. Then, Experiment 7b uses the "In this setting, is an {adjective} {noun} still a {noun}?" design from Experiment 5b to have the model rate the bigram inference in each of these 12 contexts, yielding 12 ratings per bigram. We can then calculate the Jensen-Shannon divergence between this LLM-generated distribution and the human distribution.

Table 4.4: *Prompts for Experiment 7a to generate contexts featuring a given bigram in which the inference "Is an {adjective} {noun} a {noun}?" will later be judged.*

In this task, you will write short, simple stories of 50-100 words about specific objects or things. The story should use simple language and describe the object in detail. **Example:** Write a short, simple story about a wooden pear. **Story:** Mark is an expert carver and carves a highly realistic pear out of dark colored wood. He hides the wooden pear in his fruit bowl among the fruit he bought from the supermarket. When his friends come to visit, only one of them noticed the wooden pear hiding among the fruit. **Example:** Write a short, simple story about melted ice. **Story:** Sam asks Carla to go to the store to buy ice for drinks for their party. Unfortunately, she leaves it in her car all day and comes back in the evening to find that it has all melted. Carla doesn't know what to say to Sam about the melted ice, which he was planning to use in their cocktails. **Example:** Write a short, simple story about a short basketball player. **Story:** Jordan's friend is on the high school basketball team, and is the tallest among her friends. At the match, Jordan notices that her friend is actually a short basketball player, as most of the other players are taller than her. Even so, her friend does very well in the game and scores six points. Now, write a short, simple story about {a/an} {bigram}, which includes the phrase "{bigram}" and describes the bigram in detail. Start your story with "Story:".
Write another different story about {a/an} {bigram}. Start your story with "Story:". (*repeated 2x*)
Great! Write three more stories about {a/an} {bigram}. Number your stories with "Story 1:", "Story 2:" and "Story 3:". (*repeated 3x*)

However, this method is computationally expensive: it took us ca. 400 GPU-hours with A100s to generate the 12 contexts for our 798 bigrams with Llama 3 70B Instruct. Thus, we conduct this experiment with one model (Llama 3 70B Instruct) and demonstrate this method as a proof of concept.

Table 4.5: *Two intuitive stories generated by Llama 3 70B Instruct about a* fake crowd*, which yield privative vs. subsective inferences. We then ask Llama 3 to rate "Is a fake crowd still a crowd?" given these contexts.*

Context A	**Context B**
The new video game, "Epic Quest," was about to be launched, and the developers wanted to create a buzz around it. They decided to stage a fake crowd of fans waiting in line outside the game store on launch night. They **hired a team of people to dress up in costumes** and hold signs that read "I've been waiting for 10 hours!" [...]	The small town of Oakdale was hosting its annual Christmas market, but the organizers were worried that not enough people would show up. To create the illusion of a bigger crowd, they **set up a fake crowd of mannequins** dressed in winter coats and hats, and placed them around the market stalls. [...]
Rating: Definitely yes (subsective)	**Rating**: Probably not (privative)

Results Table 4.5 shows two LLM-written stories illustrating a privative vs. subsective inference for *fake crowd* which successfully capture human intuitions about two kinds of *fake crowd*, showing promise for this method. Section B.1 contains additional examples of generated contexts. Overall, however, the Jensen-Shannon divergences in Table 4.6 show that generating contexts actually fits the human distribution *worse* than just taking the log-probability distribution directly for Llama 3 70B Instruct.

Table 4.6: *Jensen-Shannon divergence between per-bigram rating distributions for humans and Llama 3 70B Instruct using the log-probability and context generation methods, for privative vs. subsective adjectives.*

	JS Divergence		
Method	Priv.	Subs.	Total
Log-probability	0.26	0.08	0.17
Context generation	0.38	0.11	0.24

Using a cut-off of 0.25 for JS divergence, we find that the distributions generated using context generation are reasonably human-like for only 61.4% of the 798 bigrams, compared

to 75.3% for the log-probability distributions. One possible explanation is that we are not generating the right kind, or a sufficient diversity, of contexts: we place hardly any constraints on the story generation, but perhaps e.g. explicitly asking for stories that disambiguate the target inference might match human behavior better, since humans see the "Is an AN an N?" question when imagining their "contexts". A final point of divergence from humans is Llama 3 70B Instruct's unwillingness to ever use the "Unsure" rating, but this may be an issue with this particular model rather than the method itself. Nevertheless, this method yields well-distributed (often bimodal) rating distributions, as shown in Figure 4.7, and, while not as close as the log-probability distribution overall, still approximates the human distribution well for many bigrams.

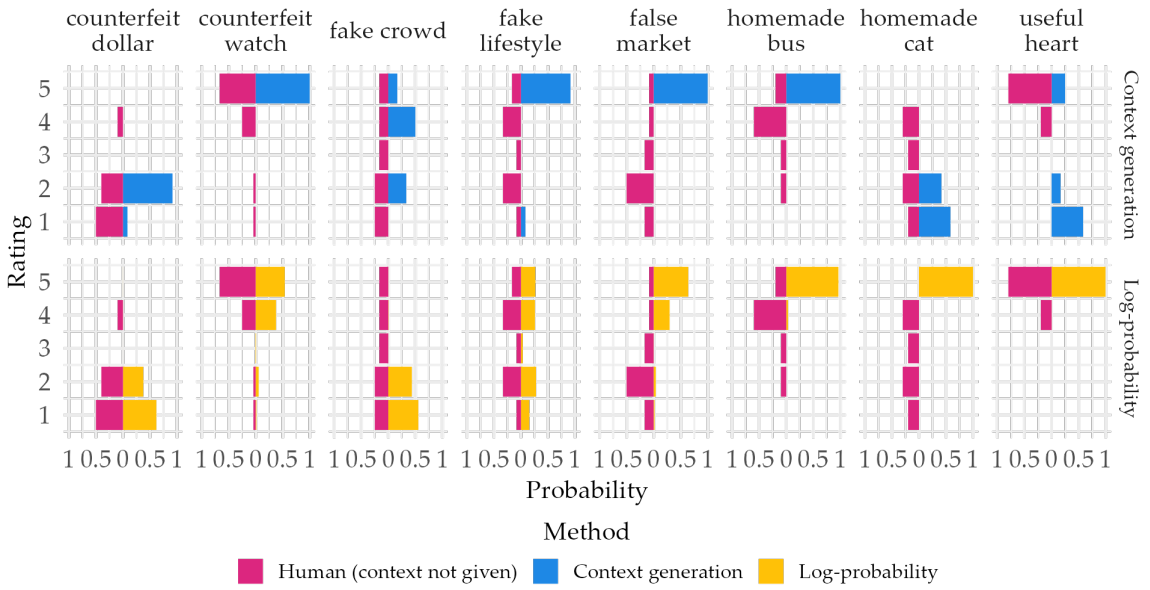


Figure 4.7: *Rating distributions over selected bigrams using log-probabilities vs. context generation, compared to the human distributions. Both methods successfully capture bigrams like* counterfeit dollar, counterfeit watch, *and both fail for* false market *and* homemade bus. *While the log-probability method fits most bigrams better, such as* fake lifestyle *and* useful heart, *the context generation method is better for* fake crowd *and* homemade cat.

4.5.4 Distribution method comparison

Overview Both methods of generating a distribution are good at capturing the narrow subsective distributions of most subsective adjectives. However, both methods struggle for items like *homemade bus*, which humans rate as more privative than subsective despite the typically-subsective adjective, resulting in a thin but long tail of high JS divergences (see Figure 4.8, which compares the distribution of JS divergences per bigram between the two methods). Interestingly, the context generation method is better able to capture this partially privative behavior for certain bigrams, such as *homemade cat*, which shows promise: it writes stories focusing around knitted or cardboard cats, matching human intuitions. (This yields a JS divergence of 0.33 compared to 1.00 for the log-prob distribution; see Figure 4.7 and Section B.1 for an example context.)

For the typically-privative adjectives which are the primary focus of this chapter, the log-probability distributions provide a better fit overall, but not for all bigrams, as shown in Table 4.7. For 45 of them, such as *fake leg*, context generation provides a better fit: many of the generated stories are about prosthetics, deemed to be *legs* (see Section B.1 for an example). More broadly, Figure 4.8 shows that there is room for improvement for both methods: both have a thick tail of bigrams whose human distributions they do not fit well.

Table 4.7: *Number of bigrams with privative vs. subsective adjectives where Jensen-Shannon divergence is low (<0.25) for both Method 2 (log-probability) and Method 3 (context generation), high (>0.5) for both, or where one Method is (<0.5 and) better than the other.*

JS divergence	Privative	Subsective
Both low (< 0.25)	103	339
Both high (> 0.5)	31	3
Log-prob > context	187	39
Context > log-prob	45	7

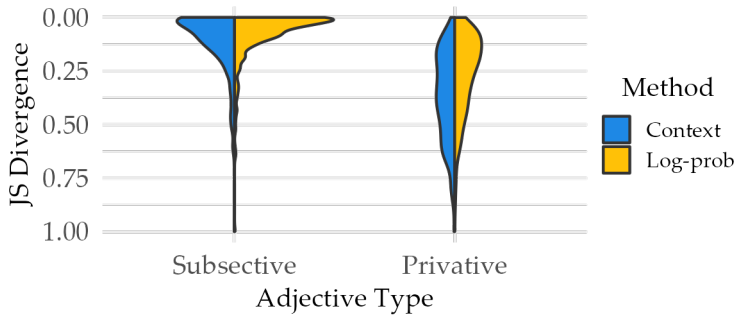
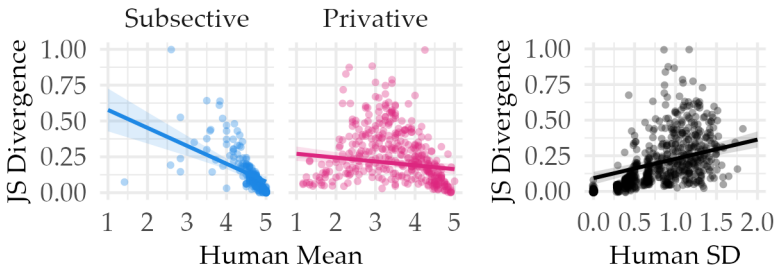


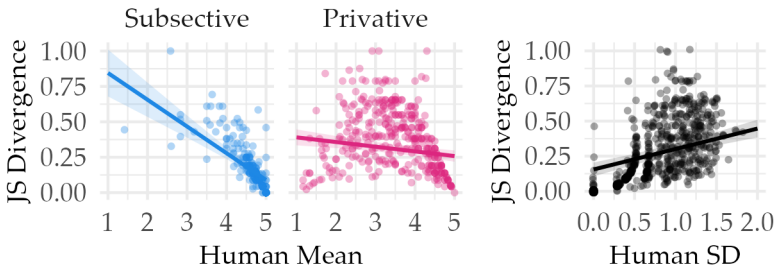
Figure 4.8: *Distribution of per-bigram Jensen-Shannon divergences between the rating distributions for humans and Llama 3 70B when obtained from log-probabilities vs. by generating contexts. 0 indicates perfect overlap, while 1 indicates maximal divergence.*

Significant predictors We fit linear regressions for JS divergence as a function of adjective type, human rating mean, human SD and bigram frequency[8] in R for each method for Llama 3 70B Instruct and find, for each method, a significant negative effect of privative adjective type and human mean on the JS divergence, as well as a significant positive effect of human SD and a significant, positive interaction between adjective type and human mean. We do not find an effect of bigram frequency. Effects plots are shown in Figure 4.9; full coefficients in Table B.7. This is an exciting result, because it shows that Llama 3 70B Instruct is similarly adept at modelling the human distribution of ratings for novel (zero-frequency) bigrams as it is for high-frequency bigrams, suggesting that it can generalize beyond its training data. Specifically, the negative effect of human mean on typically subsective adjectives supports the qualitative finding that both methods struggle with bigrams involving typically subsective adjectives with low human ratings, i.e. which have a more privative interpretation. Interestingly, we see the same effect for privative adjectives, where JS divergence increases as the human mean decreases (i.e. as the bigram is rated more privative). Finally, the significant positive effect of human SD shows that both methods struggle to predict the human distribution as human variation increases. These linear regressions achieve an R^2 of 0.44 (log-probability) and 0.55 (context generation), meaning there is still variation left unexplained by these factors. However, we find no further qualitatively interpretable patterns.

[8] `JSDivergence ~ AdjectiveType * HumanMean + HumanSD + BigramFrequency`



(**a**) *Predicted JS divergence for log-probabilities*



(**b**) *Predicted JS divergence for context generation*

Figure 4.9: *Effects plots for the regressions* `JSDivergence ~ AdjectiveType * HumanMean + HumanSD + BigramFrequency` *for each method in Section 4.5.4. There is no significant effect of bigram frequency.*

Qualitative analysis We further conduct a qualitative error analysis of the bigrams where the context generation method's JS divergence from the human distribution is particularly high. Manual inspection suggests that some issues likely stem from Llama 3 70B Instruct not picking same the (literal or abstract) noun meaning as humans: for example, most contexts for *useful heart* involve a metaphorical *heart* (e.g. the heart of a community), while *false market* contexts all involve physical markets rather than economic markets. In some cases, the model also appears to interpret the adjective metaphorically or loosely: judging by the generated contexts, Llama 3 seems to think that a *fake idea* is simply a *wrong idea* or *false idea*. While *fake* can sometimes mean *false* and vice versa, this is not the intuition the authors had about *fake idea*. Finally, some bigrams also get misunderstood as noun-noun compounds such as *counterfeit attack* in the generated context in Section B.1, or incorporated into longer noun-noun compounds, such as *useful attack dog* or *former attack helicopter*.

More broadly, while the LLM generates pleasingly diverse contexts for examples like *fake crowd*, other examples such as *fake concert* are lacking in diversity, with essentially the same blueprint being repeated in all 12 stories.

4.6 Conclusion

In this chapter, we study whether LLMs can combine adjectives and nouns to yield noun membership inferences both with and without context, and for both high-frequency and zero-frequency (presumed novel) bigrams. We find that when the inference is determined by context, large, recent LLMs make the expected inferences, while smaller and/or older LLMs only sometimes do so. All LLMs, even smaller ones, behave similarly for zero-frequency bigrams as they do for high-frequency bigrams, suggesting that they do not rely on having seen them and their associated inference during pretraining. Instead, they evaluate these combinations on-the-fly and can generalize accordingly, as humans do.

We investigate three methods to evaluate whether LLMs can extend this behavior to a setting where no context is provided, where they either need to fall within the human distribution of ratings or match the whole human distribution. While recent LLMs are able to fall within 1 SD of the human mean for up to 95% of our dataset, this is a very lenient metric. On our stricter metric of matching the human distribution (either using log-probabilities or via context generation), we find that these LLMs are good at capturing the simple distribution of bigrams like *multicolored flower* and *counterfeit watch*, but are only able to capture up to 75% of our total dataset. LLMs struggle particularly with bigrams with unusual inferences for their adjective, such as *homemade cat*, and with bigrams with high human variance. Interestingly, however, LLMs are still equally able to capture distributions of novel bigrams in this setting. In sum, our generalization results are exciting because they show LLMs are generalizing beyond their training data even in this delicate, context-sensitive task, but there is still room for improvement on the task of capturing these inferences overall. Further, we hope that the methods presented in this chapter will inspire future work which moves beyond targeting

single correct answers and begins to target human population distributions, from meaning and inferences to wider issues such as opinions and political positions.

4.7 Limitations

This chapter has a number of limitations, being the first exploration in this area. Firstly, we only study English adjective-noun inferences. The set of typically-privative adjectives and which inference they trigger with which nouns may vary between languages and cultures. Secondly, the dataset for Experiment 5, which determines the inference by providing an appropriate context, is not very large (56 bigram/context pairs), resulting in lower than ideal statistical power.

Third, for fair comparison to human participants, and since the human data demonstrates that many judgments are non-binary for this task, we use a 5-point Likert scale rather than a binary yes/no question. This assumes that the model can use the Likert scale. While there is some previous work which also uses Likert scales with similar models (Abeysinghe and Circi, 2024; Argyle et al., 2023; Chuang et al., 2024), there is no work explicitly showing that models understand Likert scales in general. Thus, this design does not let us disentangle whether the model is struggling with the task itself or simply with the use of the Likert scale, despite the 5-shot setting and the use of log-probabilities to enforce use of the scale. This especially applies to smaller models that are known to show weaker instruction-following skills, and to non-instruction-tuned models. Alternatives such as calculating the surprisal of e.g. *a fake watch is/is not a watch*, however, have their own issues, since LLMs have been shown to be sometimes insensitive to negation (García-Ferrero et al., 2023; Truong et al., 2023).

Fourth, for comparison reasons, we use the same prompts used in the human experiment to evaluate the LLMs, including few-shot examples; however, it is possible that this prompt phrasing and exact choice of few-shot examples introduces artefacts which adversely affect the LLMs' performance and "masks" their underlying ability at the task.

Finally, due to time and compute limitations, we were only able to run the context generation method for Llama 3 70B Instruct. In future work, we would like to extend this method to

other models and investigate which of the differences we see between the context generation and the log-probability sampling method for Llama 3 70B Instruct should be attributed to the model vs. to the differences between the methods.

4.8 What can we learn about humans from studying LLMs?

4.8.1 Standards of evidence for composition

Studying compositional generalization in language-using systems other than humans helps us both design these experiments more rigorously and also set standards for finding evidence of composition in humans.

Firstly, when designing experiments for LLMs, we know that LLMs are prone to "shortcuts" such as using frequency, answer length or surface-level similarity to select the answer rather than solving the task the way we intended. Controlling for these factors also leads to a more rigorous design for humans, who could in principle also be using such factors. In particular, LLMs might be solving our task by similarity and/or analogy rather than by composition; LLMs are known to implement word and phrase similarity (Pavlick, 2022) and can compute analogies (Webb, Holyoak, and Lu, 2023). Properly considering this possibility resulted in an entire chapter (Chapter 5) that applies to humans as much as to LLMs and that would not have been written otherwise.

Previous discussion of composition and symbolic reasoning has often been defined by behavior on certain tasks (Fodor, 1998; Pavlick, 2025; Quilty-Dunn, Porot, and Mandelbaum, 2023). When people find that LLMs can do these same tasks or exhibit these same behaviors, they often find themselves hesitant to ascribe compositionality or symbolic reasoning to them. It seems that behavior on such tasks may not feel satisfactory as evidence of composition after all, especially given what we know (and what we don't know) about the internals of LLMs. There are several responses to this. The first response is to say that we must dismiss our human-centric biases and conclude from these results that LLMs are indeed performing composition per our definition.

A second response is to say that behavior on a task is sufficient, but that our task was not sufficiently well designed to rule out other ways of solving it which we suspect to be at play in LLMs (but had not considered for humans). Perhaps LLMs are able to derive the same privativity inferences as humans without really composing the meaning of the phrase as an intermediate step, somehow combining cues about the inference from the adjective and cues from the noun directly without composing the meaning. In this case, the behavior is not compositional in the way that we want: while the inference about the whole may arise from information contained in the parts, our definition of compositionality (C_0 in Chapter 1) requires a grasp of the entire meaning of the whole and have it arise from the entire meaning of the parts. Under this view, we need more research to find and implement better tasks to measure behavioral reflexes of compositionality.

A third response is to tackle the intuition behind this: perhaps what we want "composition" to mean is inherently at odds with what we believe LLMs to be doing, namely combining vectorized distributional representations of words and/or word meaning into some new vectorized distributional representation by some kind of matrix multiplication we don't understand. If so, we need to redefine our definition of compositionality and/or our standards of evidence (McCurdy et al., 2024; Pavlick, 2025): under this view composition requires symbols at the appropriate level of abstraction, and we need evidence of symbolic processing and/or symbolic representations of the right sort as well as evidence that they are combined in the expected way. Nefdt (2020) defines three levels of compositionality: process, state and output; this response says that process compositionality is what we mean, and not state or output. Behavioral experiments can only measure output-level compositionality. Even with this split, there is a serious question of what level of processing we should expect compositionality at: it is not obvious whether the human implementation of composition at the lowest neural level looks that different to the description just rejected, even if we take for granted that humans use symbolic thinking at a higher level of abstraction. Baggio (2021) defines a principle of compositionality that can be applied to human processing, but it differs significantly from the linguistic principle we discussed in Chapter 1. Under this view,

we need more research in both mechanistic interpretability for LLMs (McCurdy et al., 2024) and in neuroscience and psycholinguistics for humans, so that we can perform closer and better comparisons between the two. Initial processing work on adjective-noun composition, including with privative adjectives, has not yielded many conclusive results across studies so far (Călinescu, Ramchand, and Baggio, 2025; Fritz and Baggio, 2020; Honari-Jahromi et al., 2021; Pylkkänen, 2019; Schumacher, Brandt, and Weiland-Breckle, 2018 i.a.).

Different researchers may find themselves sympathetic to any one or even multiple of these responses. What this debate shows is that just asking these questions about LLMs uncovers multiple new angles for research in humans, both in defining the phenomenon and setting the standards of evidence (and designing according experiments) and in exploring how composition works at a processing level as well as in the semantic theory.

4.8.2 Experiment 7 tests a hypothesis about humans

LLMs, by design, capture large amounts of statistical information both about language (such as frequencies of words and collocations) and also information that is contained in language about how the world generally works. Semantic theories generally capture neither of these things. The composition of content words is one of the rare phenomena where these things have an effect on a semantic phenomenon, namely (membership) inferences. In Chapter 3, we made strong and perhaps circular-seeming assumptions about the context (such as what property *fake* was targeting) in order to be able to derive any inferences at all. This explains some but not all of the human behavior found in Chapter 2. Modeling with LLMs can help us test hypotheses about other aspects of that human behavior, complementary to the contributions of the theory.

As we discussed when setting up Experiment 7, a key observation from Chapter 2 is that the inference drawn for a given adjective-noun bigram depends on the context. Fixing the context results in narrow distributions, while the out-of-the-blue formulation in Experiment 2 results in significant variability among participant judgments for some bigrams. We hypothesized that in fact, such contexts explain the variation in Experiment 2: different participants

imagine different "contexts", loosely speaking, when they answer the out-of-the-blue question. (The imagined "contexts" are likely more underspecified than the contexts provided in Experiment 3.) These may be sampled by chance or be derived from participants' different priors and experiences of the world. This variation is hard to model in a standard semantic theory and is typically left to some opaque context variable c or to underspecification. In fact, we can use LLMs to model this aspect and test this hypothesis, by having them generate a selection of explicit contexts that represent these imagined "contexts", as we did in Experiment 7. Then we can have humans or LLMs rate the inference in these contexts. If the rating distribution we get over these generated contexts matches the variation of the out-of-the-blue human rating distributions, then this is plausible evidence that the variation is indeed (partially) derived in this way. Since we established in Experiment 5 that LLMs are quite good at getting human-like inferences when given a context, we can use LLMs to substitute for human ratings here, allowing us to rate a large volume of bigrams and contexts. Indeed, Experiment 7 shows a relatively good fit to human out-of-the-blue distributions using this method, although there is still some deviation. As we discussed above, this may be an issue with the implementation details rather than the general method. Notably, when humans are coming up with "contexts" to answer the out-of-the-blue question, they are doing this with the goal of resolving the privativity inference in order to answer the question. We gave our LLM no such instructions when it generated the contexts, meaning that the contexts may not have been as helpful as the ones humans came up with for resolving the inference. It is also unclear what temperature setting on the LLM best represents the amount of variation between human participants. These difficulties can likely be overcome in future work; the fact that we get a very reasonable JS divergence between humans and LLMs even in this initial setup is highly promising and suggests that context generation may well explain a good amount of the out-of-the-blue variation between human participants.

Chapter 5

Is analogy enough to draw novel adjective-noun inferences? [1]

5.1 Introduction

How are humans able to generalize to complex linguistic expressions they have not encountered before? One view on how this can be achieved is through a mechanism of composition, determining the meaning of the phrase and any resulting inferences from the meanings of its parts (Partee, 2009; Szabó, 2012, i.a.). Others, however, believe that composition is not required: mechanisms such as analogy are sufficient to explain humans' ability to generalize to novel phrases (Ambridge, 2020; Bybee, 2010 i.a.). The same question arises when we study LLMs' ability to generalize. If they can generalize to novel phrases, is this evidence that they must be composing these phrases from their subparts, or is there another way to achieve the same results?

Chapter 2 argues that humans must be using composition, since they converge on the inferences of at least some combinations that they are assumed never to have seen before (e.g., for *fake reef* or *counterfeit scarf*, which never appear in a large corpus). Chapter 4 suggests a

[1] This chapter will also be published in the proceedings of SCiL 2025, co-authored with my advisors Kathryn Davidson and Najoung Kim, under the same title.

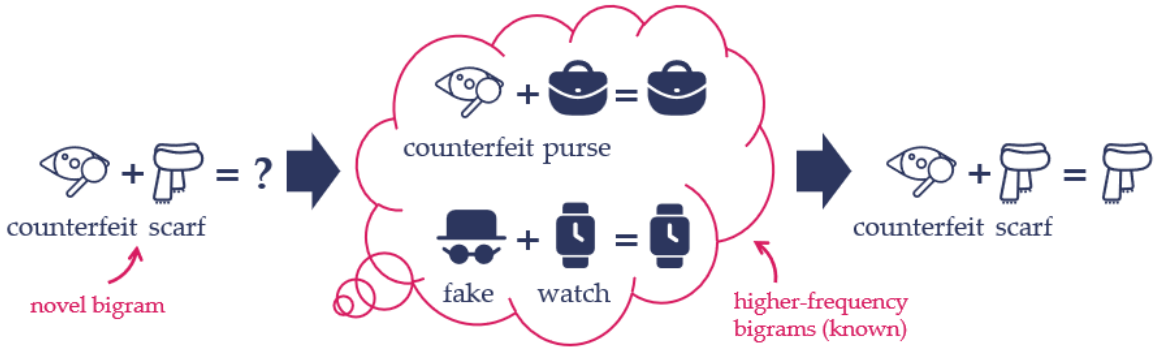


Figure 5.1: *Possible analogical reasoning to infer that counterfeit scarf is a scarf, since a counterfeit purse is a purse and a fake (or counterfeit) watch is a watch.*

similar conclusion for LLMs based on the same dataset, since LLMs show reasonably human-like behavior on at least some bigrams that are assumed not to be in the LLMs' training datasets. These combinations are interesting because the membership inferences targeted (e.g., "Is a counterfeit scarf still a scarf?") depend not just on the adjective but also on the noun, involving significant detail about how exactly the adjective affects the noun and what properties are important for membership in that noun category in typical situations.

This chapter questions these conclusions, and investigates whether this task can in fact be solved by analogical reasoning, without composition. For example, for *counterfeit scarf*, one might reason (as in Figure 5.1): "Is a counterfeit scarf still a scarf? A scarf is an accessory like a watch or a purse, and a counterfeit watch is still a watch, and a counterfeit purse is still a purse, so a counterfeit scarf is most likely still a scarf". This skips the compositional step of combining the meanings of the words to derive the meaning of the bigram, and further violates the principle of compositionality as stated by Szabó (2012) by referring to information beyond the meaning of the bigram's parts, namely the inferences associated with other adjective-noun bigrams.

We investigate analogical reasoning through two complementary approaches. First, we build a computational model of analogical reasoning which attempts to derive ratings for the low-frequency and zero-frequency (assumed novel) bigrams in the dataset of Chapter 2, by analogy to the high-frequency ones. A computational model allows us to precisely define

what we mean by analogy, and explore the consequences of different implementation decisions. Second, we ask human participants to reason analogically, guided by examples and their own intuition of what analogy means. We then evaluate how often they can produce an analogy, and whether the resulting rating distributions derived analogically is the same as the distributions from Chapter 2, where no instructions on how to reason were given. We find that the ratings derived by analogy significantly differ for several bigrams, suggesting that the original participants did not derive (all) their ratings by analogy.

Between the two methods, we find convincing evidence that while analogical reasoning produces similar results in many cases, it is not sufficient to derive the full set of inference data. Thus, we find support for the view that humans must have access to a compositional mechanism. Further, our analogy model performs worse on novel bigrams than the best LLM in Chapter 4, and our analogy model's successes and failures correlate poorly with those of the best LLM. This suggests that the LLM is not (just) using analogy in the cases where it can generalize, and supports the claim in Chapter 4 that such LLMs are performing some kind of composition (productively combining the meaning of adjective and noun) in these cases.

5.2 Related work

So-called "privative" adjectives such as *fake* pose a challenge for compositional accounts of semantics, since they cannot be simply intersected with the noun (Kamp and Partee, 1995). Multiple accounts have been proposed for how composition with privative adjectives should work (del Pinal, 2015; Guerrini, 2024b; Martin, 2022; Partee, 2010 i.a.).

Most previous computational work on adjective-noun composition using distributional semantics does not (explicitly) cover privative adjectives (Baroni and Zamparelli, 2010; Hartung et al., 2017; Vecchi et al., 2017). Boleda, Baroni, et al. (2013) cover 16 "non-intensional" adjectives, including two which are commonly taken to be privative (*former*, *mock*; see Nayak et al. (2014) for a classification). Boleda, Baroni, et al. build distributional semantic models of adjective-noun composition that use vector addition and matrix multiplication to model adjective-noun composition, but they do not cover analogy. Cappelle, Pascal, and Keller

(2018) study the distributional semantics of *fake* and bigrams in which it occurs, but do not implement any method of composition.

Ross, Kim, and Davidson (2025) gather a large quantity of offline human judgments on (privative) adjectives and their membership inferences, discussed further in Section 5.3, and Ross, Davidson, and Kim (2024) extend this dataset to assess LLMs. While Ross, Davidson, and Kim (2024) do propose a simple analogy baseline to compare to their LLMs, we propose an improved, more powerful and configurable analogy model and present a detailed analysis of its performance.

Analogy has been much studied as a core component of human reasoning (see Hofstadter, 2001 for an overview), and approaches such as construction grammar propose that analogy to known exemplars can be used to understand any novel phrase (Ambridge, 2020; Bybee, 2010). Rambelli, Chersoni, Blache, et al. (2024) propose a computational model of this process based on distributional semantics. While we also build our computational model around analogy between phrases, we only attempt to derive membership inferences from the analogy, and avoid commitment to whether the full meaning of the phrase can be accessed by analogy.

5.3 Human judgment dataset

Chapter 2 presents a dataset of human judgments on adjective-noun inferences of the form "Is an {adjective} {noun} still a {noun}?" on a 5-point Likert scale. The dataset covers 798 bigrams (102 nouns crossed with 6 typically privative and 6 typically subsective adjectives, filtered to only include combinations that make sense).[2] In this dataset, the question is presented out of the blue as a generic, rather than in a discourse context. The additional information in a discourse can sometimes determine the inference on its own (without needing to interpret the bigram at all), whereas the out of the blue setting requires some kind of reasoning strategy (composition, analogy or otherwise) to determine the inference. 180 of

[2]In this chapter, we follow Chapter 2 in using "(typically-) privative / subsective adjective" to refer to adjectives historically classified as such, which often but not always result in the respective inference.

the 798 bigrams are zero frequency in the C4 pretraining corpus (Raffel et al., 2020), which Chapter 4 takes as a proxy for the undisclosed pretraining corpora of the models they study. These bigrams are assumed to be novel to both humans and LLMs. A bigram is referred to as high frequency if it is in the top quartile of bigrams they study.

Chapter 2 shows that the membership inference in question depends on both the adjective and the noun, with bigrams with "subsective" adjectives usually yielding subsective inferences (e.g., "a homemade N is an N", but not always: consider *homemade cat*), while bigrams with "privative" adjectives such as *fake crowd* elicit a wide distribution of ratings from subsective ("is") to privative ("is not"), with high variance for many (but not all) bigrams. Varying ratings between participants are expected in this setting, since we are dealing not only with the lexicon but also with a broad question (a linguistic generic) which may depend on participants' world knowledge. Participants nonetheless show convergent ratings for many zero-frequency bigrams, demonstrating their ability to generalize and implying a shared underlying mechanism.

5.4 Analogy model

5.4.1 Algorithm

We implement a computational model of analogy which is "trained" on the human ratings from Chapter 2 for a set of common (high-frequency) bigrams, which are stored in the model's memory. This is intended to imitate human prior experience with certain bigrams, where they may have learned that, for instance, a *counterfeit watch* is still a *watch*. Humans are known to store frequent multi-word expressions even when those expressions are compositional, not just when they are idiomatic (Arnon and Snider, 2010; Caldwell-Harris, Berant, and Edelman, 2012; Tremblay and Baayen, 2010, i.a.), so it is plausible to assume that they can also store the associated inferences. Specifically, we consider the top quartile of bigrams in Chapter 2 as "known", i.e., in the training set. (Section C.1 also explores an alternative approach where the training set is balanced evenly across adjectives.)

Given these known bigrams, the model predicts the ratings for the remainder of the bigrams by analogy to similar bigrams in its training set, via the algorithm in Figure 5.2. The setting `mem` configures whether this algorithm is also applied to bigrams in the training set, as if they were not known; we discuss in Section 5.4.4 what is more human-like.

The model stores and predicts the entire rating distribution for each bigram, rather than a single rating. As Chapter 4 discussed in the context of LLMs, it is not clear how to evaluate the alignment of a single rating against high variance distributions like the human data we are taking as the evaluation target. As discussed in Section 5.3, such high variation is a natural consequence of working with the lexicon, but does necessitate a more complex metric than just accuracy to assess model fit. We use same metric that Chapter 4 uses for LLMs: the Jensen-Shannon divergence between the model-predicted rating distribution and the human rating distribution for each bigram. We compute an aggregate score by averaging across all bigrams. We report this aggregate score as well as the average score over zero-frequency bigrams (presumed to be novel to both humans and LLMs) to measure its ability to generalize. These zero-frequency bigrams are always held out from the model.

Implementing analogical reasoning in a computational model allows us to define precisely what we mean by analogy and test the effects of these implementation choices. We explore two types of analogy: either just over nouns (*counterfeit scarf* → *counterfeit watch*),[3] or allowing analogy over both noun and up to one additional adjective (*counterfeit scarf* → *fake watch*; `N+A` setting). We allow the model to retain $k \leq 5$ nearby bigrams (after filtering to bigrams in the training set) to impose constraints akin to human working memory (Adam, Vogel, and Awh, 2017; Cowan, 2001). The exact value of k is a hyperparameter optimized on the training set (with memorization disabled). Section C.1 also discusses the case where $k = 1$, i.e. where the model only considers the most similar bigram, which is a plausible route for humans.

[3]We see in Section 5.5 that this is a popular human strategy: humans choose an analogy over just nouns 58% of the time.

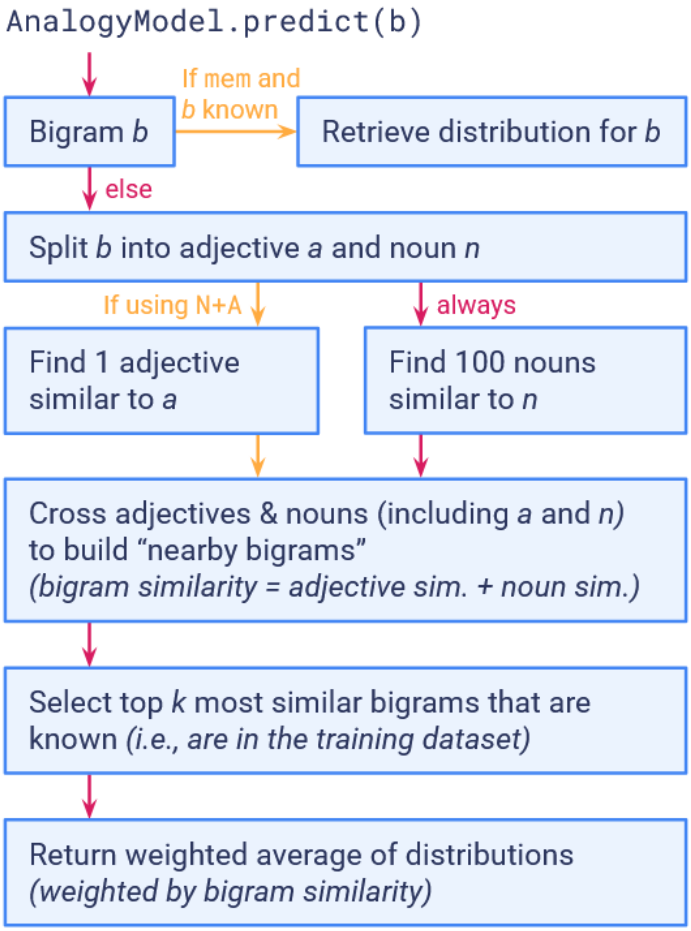


Figure 5.2: *Algorithm for the analogy model. Yellow paths are dependent on the configuration options* `mem` *and* `N+A` *(Noun + Adjective). k is a hyperparameter.*

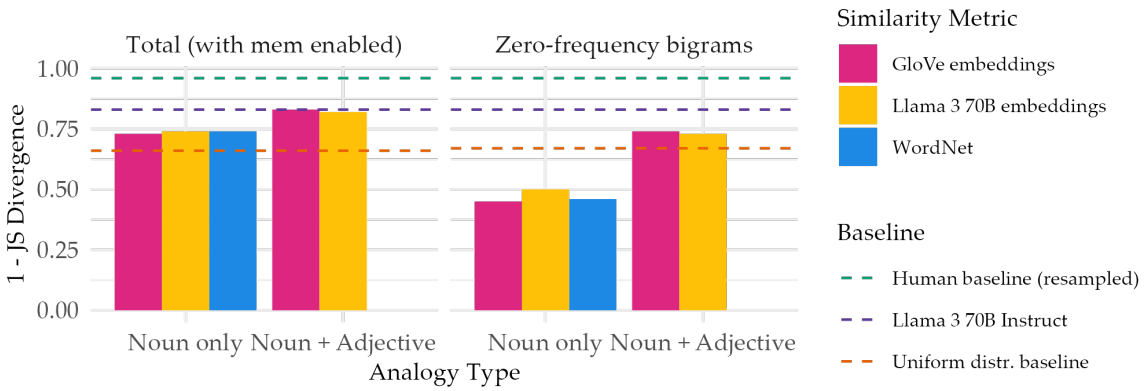


Figure 5.3: *Average JS divergence between distributions produced by the analogy model and human distributions from Chapter 2 on zero-frequency bigrams and on the whole dataset (with memorization of the training set). Additional results are given in Table C.1 in the Appendix.*

We calculate word similarity in three ways: (1) cosine similarity over GloVe embeddings (Pennington, Socher, and Manning, 2014); (2) cosine similarity over embeddings from Llama 3 70B Instruct (Dubey et al., 2024) and (3) Wu-Palmer similarity over the WordNet taxonomy (Miller, 1995; Wu and Palmer, 1994). Llama 3 70B Instruct was selected as the source for LLM embeddings because this was the model with the highest performance in Chapter 4. To derive word embeddings from Llama, we pass each word individually to the LLM and average the hidden states of the subword tokens in the final layer.[4] Wu-Palmer similarity groups nouns[5] that share common hypernyms in WordNet, penalized by how broad that hypernym is. Using WordNet allows us to measure similarity based solely on a human-created dataset, as opposed to distributionally derived embeddings. Since WordNet does not provide a taxonomy of adjectives, this approach is limited to noun-only analogies.

5.4.2 Results

Figure 5.3 shows the performance of the different analogy model configurations on the whole dataset (allowing memorization of the training set) and on held-out, zero-frequency bigrams

[4]We could alternatively pool the embeddings from the initial embedding layer, but the absence of contextualization in this approach may degrade results for multi-token words (~40% of our dataset). Nevertheless, we show in Section C.1 that results are similar in this setting.

[5]Strictly, the metric groups noun synsets ("senses"); we use the 2 most common synsets per noun.

(assumed to be novel to humans and LLMs). More details, including results for privative adjectives only and for single-bigram analogies ($k = 1$), are given in Section C.1 (Table C.1).

GloVe embeddings. Both the noun-only and N+A setting perform well overall, with the N+A setting appearing to be on par with LLM performance. However, we find that this is reliant on memorizing the training set; neither setting generalizes well to zero-frequency bigrams. In particular, noun-only analogies perform below a uniform distribution baseline on zero-frequency bigrams.

WordNet. Perhaps surprisingly, we find that this qualitatively different similarity metric yields very similar results to using GloVe embeddings, at least in the noun-only case where this metric is defined. We discuss the implication further in Section 5.4.3.

Llama embeddings. Using the embeddings derived from Llama 3 70B Instruct also does not improve performance significantly compared to using GloVe, though we see a small increase for the noun-only setting—see also the discussion in Section 5.4.3.

Error analysis. To investigate where the analogy model fails, we fit a linear regression in R (R Core Team, 2023) that predicts the JS divergence of the best-performing model from the adjective class (subsective vs. privative), human rating mean and human rating SD, with an interaction between adjective class and mean. Including the human SD allows us to target bigrams with divergent ratings; including an interaction of adjective class and mean allows us to pick out e.g. bigrams with subsective adjectives but privative ratings.

All main effects and the interaction are significant: JS divergence is lower for privative-class adjectives, higher for bigrams with subsective-class adjectives with privative ratings (i.e., low mean ratings, such as *homemade money* or *tiny abundance*), higher for privative-class bigrams with subsective ratings (i.e., high mean ratings, such as *false rumor* or *counterfeit watch*), and lower for bigrams with a high human standard deviation. The fact that it struggles on bigrams like *homemade money* (JS = 0.81) and *tiny abundance* (JS = 0.58) in particular

is not surprising, given that these adjectives are subsective for all except two bigrams in the model's pool of analogy candidates.

5.4.3 Discussion: Effect of similarity metric

The similarity metric used is not a main modulator of model performance. One possible explanation is that the analogies found by our model may often be suboptimal or inadequate, regardless of the similarity metric used. There are two potential sources of this inadequacy: first, analogical reasoning may inherently be a flawed approach for some bigrams. Second, the training set may be so sparse that the model cannot retrieve sufficiently similar nouns or bigrams to adequately support analogical reasoning. After all, our training set contains ratings for only 279 bigrams using 89 nouns (of 102 nouns in the original dataset).[6] While we cannot fully tease these two possibilities apart with our current experiments, Section C.3 explores adding data from the human rating experiment in Section 5.5.

5.4.4 Discussion: Humans

Working with lexical semantics requires us to deal with per-bigram distributions and a distribution comparison metric, rather than proportions of correct answers or significant effects in a regression. This makes interpretation of the results more complicated. It is not clear at what threshold to conclude that the model captures human performance, versus what amount of JS divergence represents noise/artifacts generated by the relatively small distribution sample size in the human experiment ($n = 12$ per bigram). Short of replicating the human experiment in Chapter 2 and calculating the JS divergence between the two, we have three points of reference: (1) We can approximate a human JS divergence by resampling from the human distribution. This yields an average JS divergence of just 0.05; (2) The best LLM performance that achieves JS divergence of 0.17 both overall and on zero-frequency bigrams (Ross, Davidson, and Kim, 2024); (3) The ratings collected from the experiment in Section

[6]The 102 nouns were selected by Ross, Kim, and Davidson such that each noun has at least one closely related other noun.

5.5, where humans are asked to perform the same task as the analogy model, yield an overall JS divergence of 0.16 compared to the original distributions.

Our analogy model achieves a JS divergence of 0.17 at best, when allowed to memorize its training data; 0.25 when it does not memorize it. On zero-frequency bigrams, the best score is 0.25. While the results are impressive with memorization, its ability to generalize to zero-frequency bigrams is 8 points worse than LLMs and 11 points worse than humans. This suggests that our analogy model does not fully capture human behavior. While a key part of the modeling assumption is that the training data represents humans' known and memorized bigrams, it is still unclear whether it is human-like to return the exact perfect distribution—all the more so considering that we typically ask humans to give single ratings, not entire distributions.

As an alternative metric, we conduct per-bigram Kolmogorov-Smirnoff tests (with Holm-Bonferroni adjustment) comparing the distributions predicted by the analogy model to the human distributions. We find that with memorization of the training set, 10 of the predicted distributions are significantly different ($p < 0.05$), of which 3 are zero-frequency bigrams; without memorization, this rises to 20. Since we only have a sample size of $n = 12$ for humans, this is a conservative estimate. Figure 5.4 shows a selection of such distributions. The fact that the analogy model significantly deviates from the correct distribution for these cases supports our conclusion that while analogy is successful in most cases, it does not offer a full explanation.

5.4.5 Discussion: LLMs

It may seem striking that the analogy model can achieve the same overall JS divergence as Llama 3 70B Instruct, the best model studied in Chapter 4, when we allow training set memorization. However, comparing results on the zero-frequency bigrams (and also on performance without `mem`, see Table C.1) shows that Llama 3 70B Instruct generalizes much better than our analogy model. Further, fitting a linear regression to predict the LLM's JS divergence per-bigram from the Llama embedding analogy model's divergence shows that al-

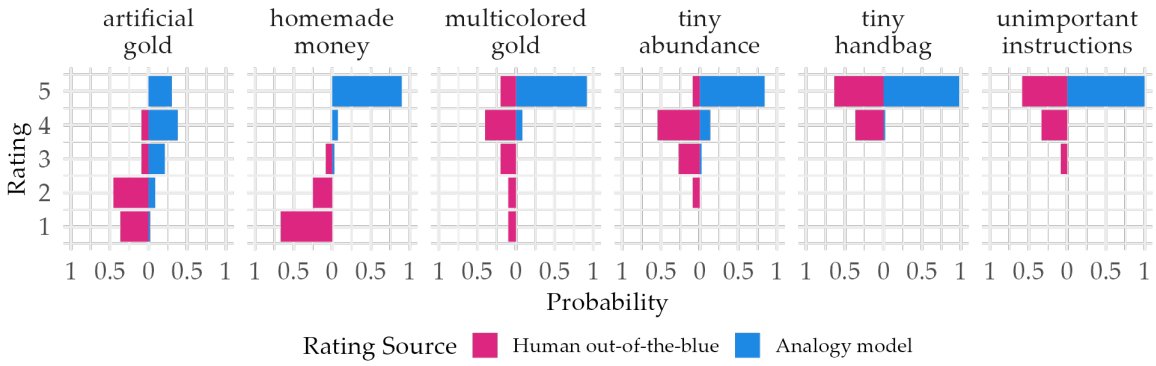


Figure 5.4: *Difference between distributions for 6 of the 10 bigrams which are significantly different between the analogy model (even with* `mem`*) and the original human distributions. In each case, the model predicts more subsective ratings than humans.*

though the effect is significant ($p < 0.001$), this only explains 12% of the variance in the LLM's ratings ($R^2 = 0.12$; $R^2 = 0.04$ with `mem` enabled). In other words, the LLM's behavior is not particularly well explained by the analogy model, and it does not succeed and fail in the same places.

5.5 Experiment 8: Human analogical reasoning

While the analogy model allows us to precisely control the mechanism and data used for analogical reasoning, it also suffers from an artificial restriction on the bigrams to which it can draw an analogy: its training dataset is strictly limited to the bigrams that Chapter 2 gathered human ratings for. Actual human analogical reasoning would not be limited in the same way, and is likely to involve a much wider range of analogy targets. In this experiment on human participants, we expand the definition of analogy to whatever our participants construe as analogy (given our instructions and training examples), enabling access to whatever bigrams they are able to come up with as suitable analogies. This allows us to measure two things: (1) how easy it is for people to come up with analogies at all, and (2) what effect analogical reasoning has on the resulting rating distributions.

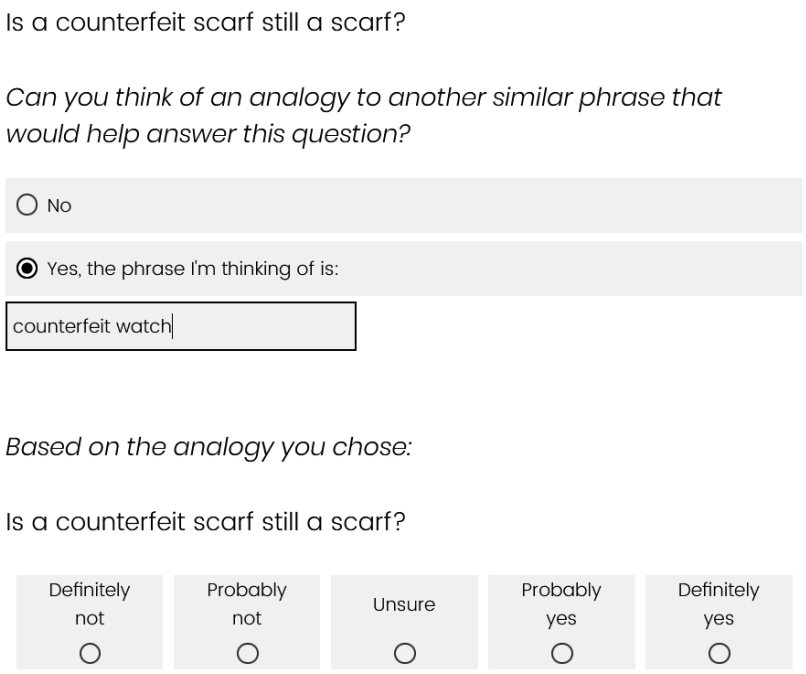
5.5.1 Method

We select 96 bigrams from the 798 bigrams from Chapter 2 such that they are evenly balanced by adjective and by zero vs. top quartile frequency, and all have convergent human rating distributions ($\mu \leq 2$ or $\mu \geq 4$ on the 5-point scale). We also attempt to include a high proportion of bigrams where analogy might be hard, though our criteria turn out not to be predictive of the actual results—see Figure 5.6 for illustration and Section C.2 for further discussion. For example, we adversarially pick some nouns for *homemade* which are likely to yield privative judgments, such as *homemade money*.

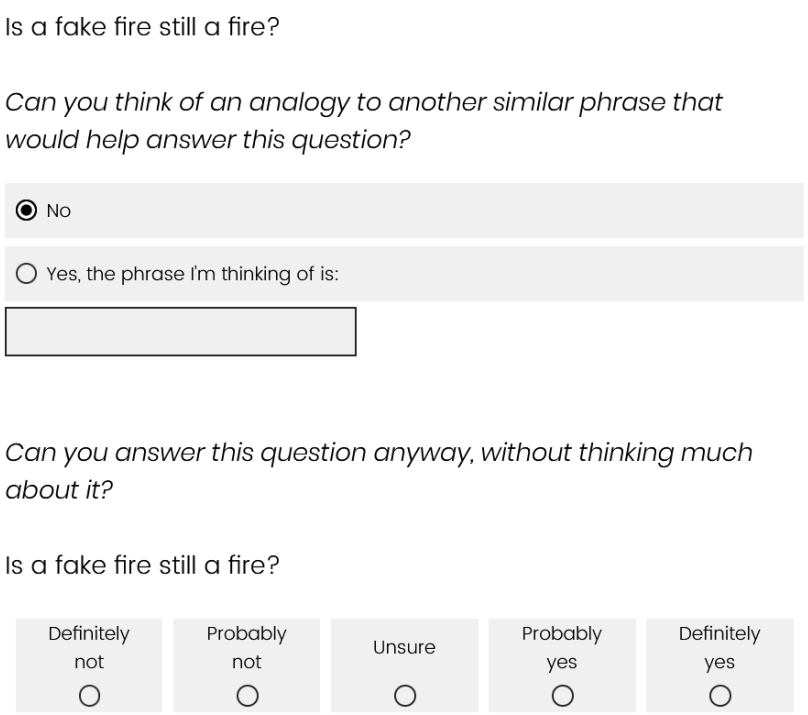
For each bigram, we show participants the question "Is an {adjective} {noun} still a {noun}?" and first ask them whether they are able to come up an analogy that helps them answer the question. We then ask them to answer the question, either using the analogy or not, depending on their first answer. Screenshots of each path are shown in Figure 5.5. Participants first see an explanation of what we mean by analogy, including an example (*toy hippo* → *toy elephant*), followed by three training examples which include another example of an analogy (*melted plastic* → *melted wax/chocolate*). The full instructions, including our description of "analogy", are given in Tables 5.1 and 5.2. The analogy text field is limited to 1-3 words to encourage analogy to adjective-noun phrases (pilot participants sometimes typed a reasoning process into the field).

We recruited 176 native American English speakers[7] on Prolific, of which we excluded 33 for not meeting our native speaker criteria, failed attention checks, or failing to adequately follow our instructions for analogical reasoning (verified based on manual inspection and regular expression searches on the free text entry fields).

[7]We recruit people who self-report English as their first and primary language and are located in the US. We further ask them at the end of the study whether they learned English before the age of 5 and whether they speak American English—if not, they are paid but excluded from the analysis. This implementation of "native speaker" is merely intended as a practical way to expect shared language experiences among our participant sample (Cheng et al., 2021).



(**a**) *Path when analogy found.*



(**b**) *Path when no analogy found.*

Figure 5.5: *Screenshots of questions in Experiment 8. Responses to the follow-up questions were only analyzed in the case when an analogy was found.*

Table 5.1: *Training instructions shown to participants to demonstrate what we intend by "analogy" (subsequent examples in Table 5.2).*

This survey involves questions of the form "Is a toy hippo still large?" We're interested in whether it's possible to solve these kinds of questions by reasoning using a similar phrase that you already know the answer for ("by analogy"), such as "toy hippo" → "toy elephant" (toy elephants are usually not large). For the purposes of this survey, the similar phrase / analogy can be another similar thing, or a class of things (like animals or gadgets). The important part is that you know the answer for the new phrase without having to think about it. Let's start with three examples that demonstrate how the survey works and what we mean by analogy. Each question consists of two parts. First you will answer whether you can think of a suitable analogy (yes/no), and type in the similar phrase if you answered yes. The phrase should consist of 1-3 words and will typically be of the form "[adjective] [noun]". Then you will attempt to answer the original question (e.g. "Is a toy hippo still large?") using the phrase you chose, or without it if you couldn't think of one. Please pay close attention to the following examples, as we will ask you to follow this style of reasoning in the rest of the survey.

Table 5.2: *Training examples shown to participants to demonstrate what we intend by "analogy".*

Is melted plastic still plastic? *Can you think of an analogy to another similar phrase that would help answer this question?* You can think of an analogy from "melted plastic" → "melted wax" or "melted chocolate." This is useful because you immediately know the answer to "Is melted wax still wax?" or "Is melted chocolate still chocolate?" So, you would answer "yes" to this question and type "melted wax" or "melted chocolate" in the text box below. *Based on the analogy you chose:* Is melted plastic still plastic? Because melted wax is still wax (or melted chocolate is still chocolate), you conclude that melted plastic is still plastic, or probably still plastic. So, you would answer "Definitely yes" or "Probably yes" depending on your interpretation.
Is a hard-boiled egg still runny? *Can you think of an analogy to another phrase that would help answer this question?* You probably find it hard to quickly think of an analogy that can help answer the question. While you may be able to come up with similar phrases, they don't immediately provide an obvious answer. So, you would answer "No" to this question. [*Instructions for second part irrelevant, omitted*]
Is a decorative pumpkin still edible? *Can you think of an analogy to another similar phrase that would help answer this question?* As in the previous example, it is hard to quickly think of an analogy that can help answer the question. While you may be able to come up with similar phrases, they don't immediately provide an obvious answer. So, you would answer "No" to this question. [*Instructions for second part irrelevant, omitted*]

5.5.2 Results

Overall, participants self-reported that they could find an analogy for 56.4% of responses. For every bigram except *fake impression*, at least one person was able to find an analogy, although 13 of 143 participants never produced an analogy. A plot of analogy availability for each bigram is shown in Figure 5.6, colored by the estimation of analogy difficulty discussed in Section C.2, which does not correlate with actual availability.[8]

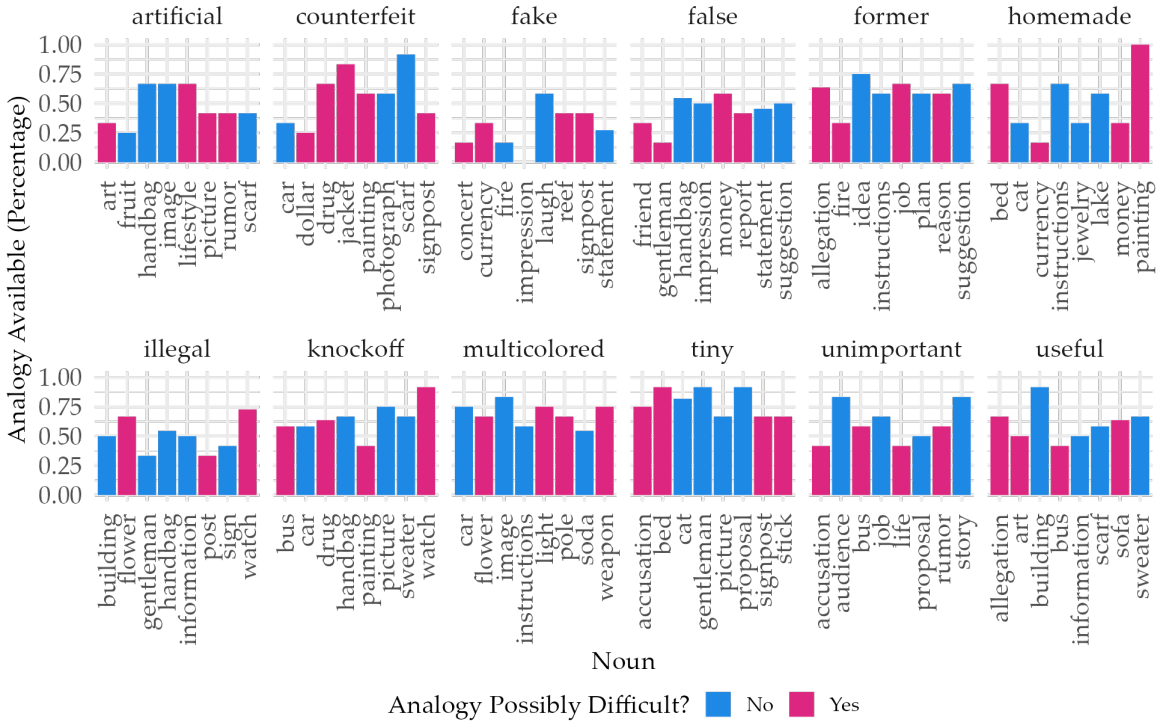


Figure 5.6: *Analogy availability for all 96 bigrams in the analogy prompting experiment. Color indicates whether it was predicted in advance that it might be difficult to find an analogy, based on the ratings from Chapter 2 in conjunction with noun frequencies and WordNet-based distance measures (see Section C.2).*

Type of analogy. Figure 5.7 shows statistics for the types of analogy drawn. We find that 58.4% of analogies use the same adjective as the original bigram, such as *knockoff watch* → *knockoff purse*, while only 10% change the adjective and use the same noun, such as *homemade*

[8]We attempted a regression to predict analogy availability but found nothing of interest; see Section C.2.

money → *counterfeit money*. A further 6.2% of analogies use a single noun. While a number of these examples seem intended as same-adjective (such as *tiny bed* → *chair*), we do see some interesting cases such as *artificial rumor* → *lie*, which may not be an analogy in the strict sense but are still solving the task by mapping to a known phrase. The remaining 25.4% use a different adjective/modifier and noun.

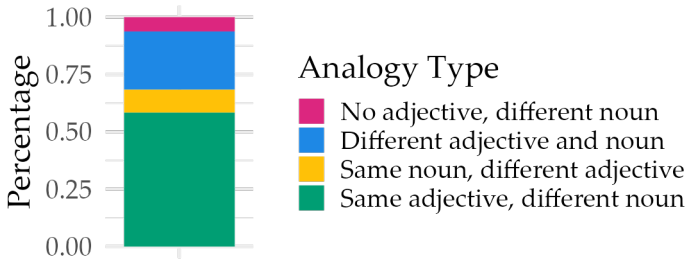


Figure 5.7: *Types of analogy chosen by participants.*

Qualitatively, we see that our participants reach for a much wider set of concepts than our analogy model when drawing analogies; choices such as *homemade lake* → *homemade cookies*, *false impression* → *wrong interpretation* or even *multicolored weapon* → *painted nails* are common. Participants are more likely than our model to reach for nouns that are not that similar to the original noun but are highly associated with the adjective, such as *knockoff purse* (11 occurrences as analogy), *counterfeit money* (10 occurrences), *homemade cookies* or *illegal immigrant* (3 occurrences each).

Distribution shift. Does analogical reasoning shift the distribution compared to the original ratings gathered in Chapter 2, where no instructions on how to reason were provided? In the cases where an analogy was found, we find an average JS divergence of 0.16 overall between bigram distributions in this experiment vs. in Chapter 2, with 0.21 on privative-type adjectives (0.32 for *fake*), 0.35 on *homemade* (recall that nouns for *homemade* were picked adversarially to be more likely to be privative) and 0.14 on zero-frequency (presumed novel) bigrams.

We also conduct Kolmogorov-Smirnoff tests per-bigram (with Holm-Bonferroni adjustment) to determine which of the distributions are significantly different. Since our n per bigram is quite small for statistical purposes (at best $n = 12$, lower if not all participants found an analogy for the bigram), no bigrams are significantly different. We cannot conclude from this that the distributions are indeed the same when analogy is used; the sample size is just too limited. Instead, we plot the distributions for 6 bigrams with the highest JS divergences in Figure 5.8. The divergence for *homemade currency* and *homemade money* (and to a lesser extent *false friend*) is particularly striking: analogy leads people to dramatically different inferences in these cases, since most *homemade* and many *false* items (such as *false rumor*) still clearly qualify as an instance of the noun.

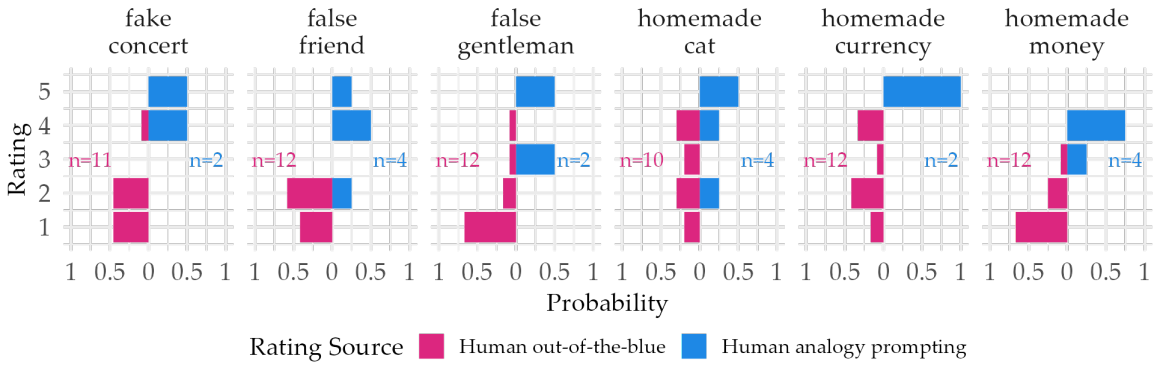


Figure 5.8: *Difference between distributions for bigrams which are significantly different when an analogy is used. n = number of ratings in each distribution; for analogy prompting, this is however many people found an analogy.*

Correlation between analogy availability and distribution shift. We fit a beta regression in R (Brooks et al., 2017) that predicts JS divergence as a function of analogy availability. We find a strong negative correlation: JS divergence decreases as analogy availability increases ($p < 0.001$). In other words, the harder it is to find an analogy, the more likely any analogies that are found will lead people astray from the original distribution.

5.5.3 Discussion

This experiment shows that analogy is a viable approach for many bigrams, and in many cases results in similar judgments as in Chapter 2, where participants could reason freely. However, for several bigrams such as *homemade money*, using an analogy yields dramatically different inferences, suggesting that analogy was not used to derive the original distribution. We also see bigrams where people struggle to come up with any analogy at all, such as *fake impression* ($n = 0$). This was the case for 10 of our 35 zero-frequency bigrams ($n \leq 50\%$), putting into question the viability of analogical reasoning for generalization. Our analogy model also shows a higher-than-average JS divergence for all bigrams where analogical reasoning substantially shifts human ratings except for *tiny proposal*. It also shows a higher-than average JS divergence for over half the ones where humans struggle to come up with an analogy. Overall, a linear regression predicting human JS divergence from the analogy model's JS divergence explains 40% of variation ($R^2 = 0.40$, $p < 0.001$), suggesting that analogy serves as a viable explanation for some, but not all of the variation in human inferences. For explaining LLM behavior, human analogy availability and human-human JS divergence when using analogies both correlate poorly with LLM-human JS divergence per-bigram, with $R^2 = 0.05$ in both cases ($p = 0.03$ and $p = 0.04$ respectively). A similar regression with our analogy model in Section 5.4.5 also showed low correlation. This suggests that analogical reasoning poorly explains LLM behavior, corroborating our previous conclusion in Section 5.4.5.

Finally, we observe that our participants use a much broader definition of "analogy" than our analogy model (or the examples we gave during training), suggesting that our analogy model adheres to adjective and noun similarity overly strictly. On the other hand, our analogy model is strictly non-compositional, whereas some human analogies such as *false impression* → *wrong interpretation* may well be arising from the participants first composing the meaning of *false impression* and then looking for phrases with a similar meaning.[9]

[9] *False* may mean *not truthful/insincere* or just *fake* (as in *false teeth*); the choice of meaning depends on the noun.

5.6 Conclusion

Chapter 2 claimed that humans must be handling adjective-noun bigrams compositionally, since they draw consistent inferences about novel bigrams, and Chapter 4 takes LLMs' capacity to draw reasonably human-like inferences on the same novel bigrams as evidence for composition. We explored the possibility that this generalization might be explained without composition in either or both cases, specifically by analogical reasoning over adjective and nouns using previously encountered and memorized inferences.

Composition in humans. We find that while many of the novel bigrams in the dataset can indeed be handled successfully by analogy, analogy is not sufficient to explain human behavior fully. Our analogy model diverges significantly from human distributions on 20 bigrams and shows insufficient generalization to zero-frequency bigrams, with a JS divergence of 0.25 from humans. Humans both struggle to come up with analogies for 24% of bigrams tested and are led astray when they do for several bigrams, such as *homemade currency*. We thus conclude that analogical reasoning is a successful strategy for generalization in a remarkable proportion of the dataset of Chapter 2, but analogy does not suffice to handle the full data. Thus, their conclusion that some mechanism of composition seems necessary to handle the whole range, *homemade currency* and all, is supported—even if humans need not (and judging by our data, quite possibly do not) invoke it in every case. This conclusion is similar to the result of Albright and Hayes (2003), who found that an analogical model of English past tense morphology did not explain participant behavior well, and concluded that speakers used abstract rules to generalize rather than analogy.

Composition in LLMs. We likewise find that LLM behavior can be partially, but not fully explained by analogical reasoning. Our analogy model is unable to reach the performance of the most successful LLMs in Chapter 4, in particular when generalizing to zero-frequency bigrams. Moreover, a linear model predicting LLM JS divergence as a function of analogy model JS divergence only explains 16% of the variance. While this does not prove that Llama

3 70B Instruct is conducting bona fide composition, it provides exciting indications that it might—at minimum, Llama 3 70B Instruct is better able to incorporate the interaction between the adjective meaning and noun meaning than our purely word analogy-based model. Investigating how composition, typically conceptualized as abstract rules, can be implemented in LLMs would be an interesting avenue of future research—the *abstraction-via-exemplars* account discussed in Misra and Kim (2023) may provide a promising starting point.

Standards of evidence for composition This chapter contributes to a broader discussion about the standards of evidence required for composition (McCurdy et al., 2024; Pavlick, 2025). If behavioral experiments about generalization can provide evidence about composition (and not all researchers believe they can), we must be sure to rule out other methods of generalization such as analogy. We further need to ensure we have a precise enough definition of compositionality to capture our intuition that analogy, by virtue of referring to information not (obviously) included in the meanings of the parts, is not a kind of composition (Szabó, 2012). By making an explicit model of analogical reasoning, we can both show the way in which it requires this additional information and show that analogical reasoning fails to generalize in the expected way, relative to our human data.

Chapter 6

Conclusion

6.1 Main findings of this dissertation

In this dissertation, we studied compositionality of meaning in adjective-noun modification through the lens of the subsective and privative inferences that this composition yields.

There is no class of privative adjectives In Chapter 2, we found that privative inferences depend not just on the adjective, as assumed in much previous theoretical work, but also on the noun and on the context that the adjective-noun bigram is presented in. All so-called privative adjectives like *fake* and *counterfeit* have nouns with which they are subsective. Moreover, for every adjective that we tested, we found some nouns that at least called the inference into question (as opposed to being clearly subsective), and we found (compound) nouns like *€50 note* that made even color and shape adjectives privative.

"Responsible" adjective-noun modification Needing to account for potential privativity for every adjective led us to propose a new theory of adjective-noun modification in Chapter 3 which makes adjectives directly responsible for resolving contradictions between adjective and noun meaning, since such contradictions are what ultimately drive privative inferences. While our data can be accounted for by noun widening, it is explained most transparently when this notion of conflict resolution is incorporated directly at the level of adjective-noun

modification, rather than when it is handed through a separate rescue mechanism. Under this view, privativity arises naturally as a result of the meaning composition process.

Privativity does not derive from context-sensitive adjective meaning While Chapter 2 finds that privative inferences are context-sensitive even for a fixed bigram, Chapter 3 argues that the adjectives themselves are not, and neither is the composition mechanism. Instead, this is reduced to (a) a standard effect of the discourse contributing more information and (b) the discourse context refining the standards for what counts as an instance of a noun in that context, independent of the composition.

***Fake* as a window into conceptual structure** Unlike other so-called privative adjectives, Chapter 2 shows *fake* has a more complex meaning, which may target any number of properties "core" to the noun, depending on the context. Besides posing an interesting challenge for how to incorporate this limited context-sensitivity into its lexical entry (Chapter 3), the felicity conditions for *fake* give us an exciting reflex of language onto the conceptual system, alongside kinds (Prasada and Dillingham, 2006), and opens avenues for future research in this space.

(Compositional) generalization in humans Despite the rich paradigm of data, Chapter 2 shows that humans still generalize to many novel adjective-noun bigrams involving subsective and privative adjectives alike. Moreover, Chapter 5 demonstrates that this can't be done by analogy in all cases: at least some bigrams whose inferences humans converge on defy analogical generalization and are best explained by a compositional mechanism.

Privativity and generalization in LLMs Somewhat surprisingly, Chapter 4 shows that LLMs trained only on distributional from large internet corpora are also able to capture a large amount of these inference phenomena. Given a context, they are successfully able to determine whether a given adjective-noun bigram has a privative inference or not, and they capture around three-quarters of the dataset from Chapter 2 in the out-of-the-blue case as

well. As part of this finding, we also tackle the challenges of evaluating LLMs on a gradient phenomenon that yields distributions of human answers rather than a single correct answer. Most surprisingly, we find that LLMs are equally able to draw these inferences for novel bigrams as they are for bigrams we assume they have seen before, indicating that the process that they use to combine adjectives and nouns can generalize. This reasonable success on modeling human behavior allows us to use LLMs to model hypotheses about the effect of context that are difficult to capture in a formal semantic theory, in a complementary fashion.

Standards of evidence for composition Whether we should conclude from the results in Chapter 4 that LLMs generalize compositionally (as opposed to by some other mechanism) raises important questions about the standards of evidence for composition and other symbolically-defined processes in both humans and LLMs. Accounting for alternative mechanisms such as memorization, frequency and analogy which are salient for LLMs also informed our human experiments, allowing us to test those paths for humans as well. In particular, discounting analogical generalization forms an entire chapter of this dissertation (Chapter 5), a step not normally taken explicitly by linguists who argue that humans must be composing meanings since they can understand novel sentences (Szabó, 2012). More broadly, the debate around standards of evidence for compositionality in LLMs (and any non-humans) suggests avenues for future research for both humans and computational systems that use language like LLMs.

6.2 Avenues for future work

As with any dissertation, there remain many avenues for future research.

Dual character concepts and other theories of conceptual combination Experiment 4 studies the acceptability of *fake* and *real* with respect to a selection of nouns and their k-properties from Prasada and Dillingham (2006) and from our Experiment 3. Some of these concepts, like arguably *lifeguard*, are *dual character concepts* which have both concrete and abstract prop-

erties and have their own special relationship with *real*, as well as with *true* (Knobe, Prasada, and Newman, 2013). More experiments are needed to understand how our thesis that *real* targets k-properties interacts with the dual character of concepts like these, and whether the theory proposed in Chapter 3 can help explain any of the phenomena we see in those studies. More broadly, this dissertation proposes a structured representation of concepts (C-structure) that is sufficient to meet our linguistic needs for adjective-noun modification, but this is not intended to be a complete representation of concepts. In particular, we do not commit to a theory of concept representation such as prototypes or knowledge/theory-based representations, although Prasada and Dillingham (2006) lives within the theory tradition. There is scope for significantly more work to link C-structure to the rich literature on concept representation in the mind and see if there are any other effects of conceptual structure on language that should be tracked in the semantics.

Integration with syntactic and cross-linguistic theories of adjectives As discussed at the end of Chapter 3, the theory in this dissertation covers only the semantics of adjective-noun modification. It should be possible to slot this theory into existing syntactic accounts of how adjectives and nouns come to be in a modifier relation, with different effects if the noun is the overt noun or a covert noun, different effects if we have reduced relative clause vs. direct modification, locality effects and so forth. Actually carrying out this integration and examining any new predictions that this makes is an obvious next step. Likewise, this theory should in principle hold equally across languages, since it governs how an adjective concept and a noun concept combine in the semantics, although of course what lexical items exist and what properties are said to hold of a noun will vary across languages and cultures. Accounts of the meaning of particular adjectives in particular languages can in principle be slotted in to the core meaning that our adjective schema provides. Again, this actual integration needs to be done, and it remains to be seen whether it sheds any new light on existing puzzles. In particular, Guerrini (2024b) raises an interesting puzzle for Italian involving prenominal

and postnominal adjective-noun modification and the associated potential differences in privativity, which is not obviously explained by this account.[1]

Theory of adverb-verb modification and dual content semantics beyond the adjective-noun phrase Privativity also appears to occur in adverb-verb modification, such as the parallel between *Sam had a false proof of the theorem* (not a proof) and *Sam falsely proved the theorem* (not a proving event). It could be very fruitful to extend our theory's machinery for handling privativity (conflicts between adjective and noun meaning) while being as faithful to the noun meaning as possible to adverb-verb modification. Two ingredients are needed for this: the first is a suitable theory of the representation of verbal/event concepts that can parallel k-properties and t-properties for nouns (see Casati and Varzi, 2008 for an overview of approaches to event representations). The second is a compositional dual content semantics that extends beyond adjective-noun phrases and can handle all the material we might expect to syntactically intervene between the verb and its modifying adverb. This includes function words like determines which might occur in the DP argument of the VP, and which do not obviously have a (contentful) C-structure.

Better behavioral experiments for linguistic compositionality The experiments in this dissertation test adjective-noun composition by virtue of the inferences that you can draw about the composed bigram. This captures only one aspect of the composition, albeit an interesting and distinctive one. Song et al. (2025) (concurrent with this dissertation) study whether LLMs can capture emergent properties (Springer and Murphy, 1992) that arise from adjective-noun composition, such as a *peeled apple* being white, not red or green, albeit using quite different methodology that may not extend so well back to humans. While we tentatively conclude in this dissertation that our experiments provide evidence for compositional generalization in humans and LLMs, this rests on the argument that we can't think of any other way to generalize that yields these results (having excluded memorization, random

[1] I was unable to replicate the patterns found by Guerrini (2024b) with my own native (mid/southern) Italian-speaking consultant, and hence leave the establishment of the exact patterns that we need to explain to future experimental work.

choice and analogy). It remains possible that humans and LLMs are somehow able to generalize these inferences without doing full meaning composition of the sort proposed by our linguistic theory. If we believe that behavioral experiments are sufficient evidence for (linguistic meaning) compositionality (McCurdy et al., 2024; Nefdt, 2020) then more research is needed to design experiments that can answer this question more fully, either in the domain of adjective-noun composition or elsewhere.

Process-level / model-internal evidence for linguistic compositionality in humans and LLMs While behavioral experiments can at minimum provide negative evidence for compositionality (if humans/LLMs cannot generalize, they they are not doing composition), some researchers believe that evidence of symbolic and/or compositional representations and processes are necessary for compositionality as they define it (McCurdy et al., 2024; Nefdt, 2020). If this is the case, more research is needed in neuroscience and LLM (mechanistic) interpretability to look for evidence of such processes. While there is a lot of current work on composition in LLMs, much of it currently focuses on function composition or circuit composition rather than the composition of linguistic meaning (McCurdy et al., 2024).

Synthetic data for analogical reasoning Chapter 5 concludes that analogical reasoning does not explain the human generalization results found in Chapter 2 in part because our computational model of analogy is unable to reproduce the human results. However, our model is limited—severely, perhaps—by its lack of bigrams to which it can to draw an analogy, since it only has access to (the high-frequency portion of) the 798 bigrams we gathered human judgments for in Chapter 2. Since Chapter 4 shows that LLMs are reasonable, if not excellent, predictors of human judgments, one way to extend the analogy model's dataset would be to generate a large quantity of synthetic, "silver standard" bigram rating distributions using a capable LLM to stand in for the more expensive human judgments. This would allow us to better investigate whether the analogical reasoning that the model implements is inherently flawed (as it seems to be for at least a small number of cases such as *home-*

made money) or whether its low correlation with human performance is partially or largely explained by lack of sufficient data (that humans can be assumed to have access to).

Conclusion This covers some of the largest areas for future work; there are certainly many other smaller gaps to be filled, including but not limited to many of the footnotes throughout this dissertation. I hope that this dissertation inspires more research in these areas, and that any reader who has made it this far in the dissertation may feel inspired to explore one or more of these outstanding questions.

References

Abeysinghe, Bhashithe and Ruhan Circi (2024). *The Challenges of Evaluating LLM Applications: An Analysis of Automated, Human, and LLM-Based Approaches*. URL: http://arxiv.org/abs/2406.03339.

Adam, Kirsten C. S., Edward K. Vogel, and Edward Awh (2017). "Clear evidence for item limits in visual working memory". In: *Cognitive Psychology* 97, pp. 79–97. URL: https://www.sciencedirect.com/science/article/pii/S0010028517300634.

Albright, Adam and Bruce Hayes (2003). "Rules vs. analogy in English past tenses: a computational/experimental study". In: *Cognition* 90.2, pp. 119–161. URL: https://www.sciencedirect.com/science/article/pii/S001002770300146X.

Ambridge, Ben (2020). "Against stored abstractions: A radical exemplar model of language acquisition". In: *First Language* 40.5, pp. 509–559. URL: https://doi.org/10.1177/0142723719869731.

Apidianaki, Marianna and Aina Garí Soler (2021). "ALL Dolphins Are Intelligent and SOME Are Friendly: Probing BERT for Nouns' Semantic Properties and their Prototypicality". In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. BlackboxNLP 2021. Ed. by Jasmijn Bastings, Yonatan Belinkov, Emmanuel Dupoux, Mario Giulianelli, Dieuwke Hupkes, Yuval Pinter, and Hassan Sajjad. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 79–94. URL: https://aclanthology.org/2021.blackboxnlp-1.7.

Argyle, Lisa P., Ethan C. Busby, Nancy Fulda, Joshua R. Gubler, Christopher Rytting, and David Wingate (2023). "Out of One, Many: Using Language Models to Simulate Human Samples". In: *Political Analysis* 31.3, pp. 337–351. URL: https://www.cambridge.org/core/journals/political-analysis/article/out-of-one-many-using-language-models-to-simulate-human-samples/035D7C8A55B237942FB6DBAD7CAA4E49.

Arnon, Inbal and Neal Snider (2010). "More than words: Frequency effects for multi-word phrases". In: *Journal of Memory and Language* 62.1, pp. 67–82. URL: https://www.sciencedirect.com/science/article/pii/S0749596X09000965.

Baggio, Giosuè (2021). "Compositionality in a Parallel Architecture for Language Processing". In: *Cognitive Science* 45.5, e12949. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12949.

Baker, Mark C. (2003). *Lexical categories: verbs, nouns, and adjectives*. Cambridge studies in linguistics 102. Cambridge, UK ; New York: Cambridge University Press. xvi+353. ISBN: 978-0-521-80638-1.

Baroni, Marco and Roberto Zamparelli (2010). "Nouns are Vectors, Adjectives are Matrices: Representing Adjective-Noun Constructions in Semantic Space". In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2010. Cambridge, MA: Association for Computational Linguistics, pp. 1183–1193. URL: https://aclanthology.org/D10-1115.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4". In: *Journal of Statistical Software* 67.1. URL: http://www.jstatsoft.org/v67/i01/.

Bertolini, Lorenzo, Julie Weeds, and David Weir (2022). "Testing Large Language Models on Compositionality and Inference with Phrase-Level Adjective-Noun Entailment". In: *Proceedings of the 29th International Conference on Computational Linguistics*. COLING 2022. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 4084–4100. URL: https://aclanthology.org/2022.coling-1.359.

Bigley, Katherine (2025). "The Adjective Formerly Known as Privative: An Analysis of the Formal Semantics of Former through Implicature and Qualia Theory". Bachelor's Thesis. Cambridge, MA: Harvard University.

Bisbee, James, Joshua D. Clinton, Cassy Dorff, Brenton Kenkel, and Jennifer M. Larson (2024). "Synthetic Replacements for Human Survey Data? The Perils of Large Language Models". In: *Political Analysis*, pp. 1–16. URL: https://www.cambridge.org/core/journals/political-analysis/article/synthetic-replacements-for-human-survey-data-the-perils-of-large-language-models/B92267DC26195C7F36E63EA04A47D2FE.

Boleda, Gemma, Marco Baroni, The Nghia Pham, and Louise McNally (2013). "Intensionality was only alleged: On adjective-noun composition in distributional semantics". In: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013) – Long Papers*. IWCS 2013. Potsdam, Germany: Association for Computational Linguistics, pp. 35–46. URL: https://aclanthology.org/W13-0104.

Boleda, Gemma, Eva Maria Vecchi, Miquel Cornudella, and Louise McNally (2012). "First Order vs. Higher Order Modification in Distributional Semantics". In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*. EMNLP 2012. Jeju Island, Korea: Association for Computational Linguistics, pp. 1223–1233. URL: https://aclanthology.org/D12-1112.

Brooks, Mollie E., Kasper Kristensen, Koen J. van Benthem, Arni Magnusson, Casper Berg W., Anders Nielsen, Hans J. Skaug, Martin Mächler, and Benjamin M. Bolker (2017). "glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling". In: *The R Journal* 9.2, p. 378. URL: https://journal.r-project.org/archive/2017/RJ-2017-066/index.html.

Brown, Tom B. et al. (2020). *Language Models are Few-Shot Learners*. URL: http://arxiv.org/abs/2005.14165.

Bybee, Joan L. (2010). *Language, usage and cognition*. Cambridge ; New York: Cambridge University Press. ix+252. ISBN: 978-0-521-85140-4.

Caldwell-Harris, Catherine, Jonathan Berant, and Shimon Edelman (2012). "Measuring Mental Entrenchment of Phrases with Perceptual Identification, Familiarity Ratings, and Corpus Frequency Statistics". In: *Measuring Mental Entrenchment of Phrases with Perceptual Identification, Familiarity Ratings, and Corpus Frequency Statistics*. De Gruyter Mouton, pp. 165–194. ISBN: 978-3-11-027407-3. URL: https://www.degruyter.com/document/doi/10.1515/9783110274073.165/html.

Călinescu, Lia, Gillian Ramchand, and Giosuè Baggio (2025). "Lexical semantics trumps syntax during noun composition in predication and modification contexts: insights from the N400 and alpha and beta band synchronisation". In: *Language, Cognition and Neuroscience* 0.0, pp. 1–29. URL: https://doi.org/10.1080/23273798.2025.2457975.

Cappelle, Bert, Denis Pascal, and Mikaela Keller (2018). "Facing the facts of fake: A distributional semantics and corpus annotation approach". In: *Yearbook of the German Cognitive Linguistics Association* 6.1, pp. 9–42. URL: https://www.proquest.com/docview/2167950985/abstract/A9E210AB57764943PQ/1.

Casati, Roberto and Achille C. Varzi (2008). "Event Concepts". In: *Understanding Events: From Perception to Action*. Ed. by Thomas F. Shipley and Jeffrey M. Zacks. Oxford University Press. ISBN: 978-0-19-518837-0. URL: https://doi.org/10.1093/acprof:oso/9780195188370.003.0002.

Chan, Xin, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu (2024). *Scaling Synthetic Data Creation with 1,000,000,000 Personas*. URL: http://arxiv.org/abs/2406.20094.

Cheng, Lauretta S. P., Danielle Burgess, Natasha Vernooij, Cecilia Solís-Barroso, Ashley McDermott, and Savithry Namboodiripad (2021). "The Problematic Concept of Native Speaker in Psycholinguistics: Replacing Vague and Harmful Terminology With Inclusive and Accurate Measures". In: *Frontiers in Psychology* 12. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2021.715843/full.

Chierchia, Gennaro (1995). "Individual-Level Predicates as Inherent Generics". In: *The Generic Book*. Ed. by Gregory N. Carlson and Francis Jeffry Pelletier. The U of Chicago Press, pp. 176–223. ISBN: 978-0-226-09292-8.

Chierchia, Gennaro, Danny Fox, and Benjamin Spector (2012). "Scalar implicature as a grammatical phenomenon". In: *Handbooks of Linguistics and Communication Science Semantics*. Vol. 3. de Gruyter.

Christensen, R. H. B. (2022). *ordinal—Regression Models for Ordinal Data*.

Chuang, Yun-Shiuan, Zach Studdiford, Krirk Nirunwiroj, Agam Goyal, Vincent V. Frigo, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T. Rogers (2024). *Beyond Demographics: Aligning Role-playing LLM-based Agents Using Human Belief Networks*. URL: http://arxiv.org/abs/2406.17232.

Cinque, Guglielmo (2010). *The Syntax of Adjectives: A Comparative Study*. MIT Press. 221 pp. ISBN: 978-0-262-01416-8. DOI: 10.7551/mitpress/9780262014168.001.0001.

Coil, Albert and Vered Shwartz (2023). "From chocolate bunny to chocolate crocodile: Do Language Models Understand Noun Compounds?" In: *Findings of the Association for Computational Linguistics: ACL 2023*. Findings 2023. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 2698–2710. URL: https://aclanthology.org/2023.findings-acl.169.

Coppock, Elizabeth and Lucas Champollion (2023). "Invitation to Formal Semantics". URL: https://eecoppock.info/semantics-boot-camp.pdf.

Cowan, Nelson (2001). "The magical number 4 in short-term memory: A reconsideration of mental storage capacity". In: *Behavioral and Brain Sciences* 24.1, pp. 87–114. URL: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/magical-number-4-in-shortterm-memory-a-reconsideration-of-mental-storage-capacity/44023F1147D4A1D44BDC0AD226838496.

Del Pinal, Guillermo (2015). "Dual Content Semantics, privative adjectives, and dynamic compositionality". In: *Semantics and Pragmatics* 8, 7:1–53. URL: https://semprag.org/index.php/sp/article/view/sp.8.7.

— (2018). "Meaning, modulation, and context: a multidimensional semantics for truth-conditional pragmatics". In: *Linguistics and Philosophy* 41.2, pp. 165–207. URL: https://doi.org/10.1007/s10988-017-9221-z.

Demonte, Violeta (2011). "Adjectives". In: *Handbooks of Linguistics and Communication Science*. In collab. with Paul Portner, Claudia Maienborn, and Klaus von Heusinger. Vol. 33/2. Berlin, Boston: DE GRUYTER, pp. 1314–1340. ISBN: 978-3-11-018523-2. DOI: 10.1515/9783110255072.1314.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. NAACL-HLT

2019. Ed. by Jill Burstein, Christy Doran, and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. URL: https://aclanthology.org/N19-1423.

Do, Nam and Ellie Pavlick (2021). "Are Rotten Apples Edible? Challenging Commonsense Inference Ability with Exceptions". In: *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Findings 2021. Ed. by Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli. Online: Association for Computational Linguistics, pp. 2061–2073. URL: https://aclanthology.org/2021.findings-acl.181.

Dodge, Jesse, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner (2021). *Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus*. URL: http://arxiv.org/abs/2104.08758.

Dubey, Abhimanyu et al. (2024). *The Llama 3 Herd of Models*. URL: http://arxiv.org/abs/2407.21783.

Ettinger, Allyson (2020). "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models". In: *Transactions of the Association for Computational Linguistics* 8, pp. 34–48. URL: https://doi.org/10.1162/tacl_a_00298.

Fodor, Jerry A. (1998). *Concepts: Where Cognitive Science Went Wrong*. Oxford Cognitive Science Series. Oxford: Clarendon Press. ISBN: 978-0-19-151906-2. URL: https://doi.org/10.1093/0198236360.001.0001.

Fritz, Isabella and Giosuè Baggio (2020). "Meaning composition in minimal phrasal contexts: distinct ERP effects of intensionality and denotation". In: *Language, Cognition and Neuroscience* 35.10, pp. 1295–1313. URL: https://doi.org/10.1080/23273798.2020.1749678.

García-Ferrero, Iker, Begoña Altuna, Javier Álvez, Itziar Gonzalez-Dios, and German Rigau (2023). *This is not a Dataset: A Large Negation Benchmark to Challenge Large Language Models*. URL: http://arxiv.org/abs/2310.15941.

Goodale, Michael and Salvador Mascarenhas (2023). *Systematic polysemy in adjective-noun combination in contextual word embeddings*. URL: https://lingbuzz.net/lingbuzz/007644.

Greenberg, Yael (2003). "Manifestations of genericity". PhD thesis. New York: Routledge.

Groenendijk, Jeroen and Martin Stokhof (2005). "Why Compositionality?" In: *Reference and Quantification: The Partee Effect*. Ed. by Greg N. Carlson and Francis Jeffry Pelletier. CSLI Publications, pp. 83–106.

Guerrini, Janek (2022). "'Like a N' constructions: genericity in similarity". In: *Proceedings of the ESSLLI 2022 Student Session*. ESSLLI 2022 Student Session. Ed. by Dean McHugh and Alexandra Mayn. Galway. URL: https://doi.org/10.21942/uva.20368104.

Guerrini, Janek (2024a). "Conceptual representations in grammar: complex meanings, simple composition". PhD thesis. Paris: Ecole Normale Supérieure.

— (2024b). "Keeping Fake Simple". In: *Journal of Semantics* 41.2, pp. 175–210. URL: https://doi.org/10.1093/jos/ffae010.

— (2024c). *Pre-nominal modification in Italian as a window into the core meaning of color adjectives.* To appear in Proceedings of NELS 55.

Guevara, Emiliano (2010). "A Regression Model of Adjective-Noun Compositionality in Distributional Semantics". In: *Proceedings of the 2010 Workshop on GEometrical Models of Natural Language Semantics*. GEMS 2010. Uppsala, Sweden: Association for Computational Linguistics, pp. 33–37. URL: https://aclanthology.org/W10-2805.

Halle, M and Alec Marantz (1993). "Distributed morphology and the pieces of inflection". In: *The view from building 20*. The MIT Press, pp. 111–176.

Hamann, Cornelia (1991). "Adjectives (Adjektive)". In: *Semantik / Semantics*. Ed. by Dieter Wunderlich and Arnim von Stechow. Vol. 6. Germany: De Gruyter, Inc. ISBN: 978-3-11-012696-9.

Hanna, Michael and David Mareček (2021). "Analyzing BERT's Knowledge of Hypernymy via Prompting". In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. BlackboxNLP 2021. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 275–282. URL: https://aclanthology.org/2021.blackboxnlp-1.20.

Hartung, Matthias, Fabian Kaupmann, Soufian Jebbara, and Philipp Cimiano (2017). "Learning Compositionality Functions on Word Embeddings for Modelling Attribute Meaning in Adjective-Noun Phrases". In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. EACL 2017. Valencia, Spain: Association for Computational Linguistics, pp. 54–64. URL: https://aclanthology.org/E17-1006.

Heim, Irene and Angelika Kratzer (1998). *Semantics in generative grammar*. Blackwell textbooks in linguistics 13. Malden, Mass., USA: Blackwell. viii+324. ISBN: 978-0-631-19712-6.

Hendriks, Petra and Helen de Hoop (2001). "Optimality Theoretic Semantics". In: *Linguistics and Philosophy* 24.1, pp. 1–32. URL: https://doi.org/10.1023/A:1005607111810.

Hofstadter, Douglas R (2001). "Epilogue: Analogy as the Core of Cognition". In: *The Analogical Mind*, pp. 499–538. DOI: 10.7551/mitpress/1251.003.0020.

Hogeweg, Lotte (2012). "Rich Lexical Representations and Conflicting Features". In: *International Review of Pragmatics* 4.2, pp. 209–231. URL: https://brill.com/view/journals/irp/4/2/article-p209_4.xml.

Honari-Jahromi, Maryam, Brea Chouinard, Esti Blanco-Elorrieta, Liina Pylkkänen, and Alona Fyshe (2021). "Neural representation of words within phrases: Temporal evolution of color-adjectives and object-nouns during simple composition". In: *PLOS ONE* 16.3, e0242754. URL: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0242754.

Hupkes, Dieuwke et al. (2023). "A taxonomy and review of generalization research in NLP". In: *Nature Machine Intelligence* 5.10, pp. 1161–1174. URL: https://www.nature.com/articles/s42256-023-00729-y.

Jiang, Albert Q. et al. (2024). *Mixtral of Experts*. URL: http://arxiv.org/abs/2401.04088.

Kamp, Hans (1975). "Two Theories about Adjectives". In: *Formal Semantics of Natural Language*. Cambridge University Press, pp. 123–155. ISBN: 978-0-521-11111-9. URL: https://doi.org/10.1017/CBO9780511897696.011.

Kamp, Hans and Barbara H. Partee (1995). "Prototype theory and compositionality". In: *Cognition* 57.2, pp. 129–191. DOI: 10.1016/0010-0277(94)00659-9.

Kennedy, Christopher (2007). "Vagueness and grammar: the semantics of relative and absolute gradable adjectives". In: *Linguistics and Philosophy* 30.1, pp. 1–45. URL: http://link.springer.com/10.1007/s10988-006-9008-0.

— (2012). "Adjectives". In: *Routledge Companion to Philosophy of Language*. Ed. by Delia Graff Fara and Gillian Russell. Routledge Philosophy Companions. Abingdon: Routledge, pp. 347–360. ISBN: 978-0-415-99310-4. DOI: 10.4324/9780203206966-36.

Kennedy, Christopher and Louise McNally (2005). "Scale Structure, Degree Modification, and the Semantics of Gradable Predicates". In: *Language* 81.2, pp. 345–381. DOI: 10.1353/lan.2005.0071.

— (2010). "Color, context, and compositionality". In: *Synthese* 174.1, pp. 79–98. DOI: 10.1007/s11229-009-9685-7.

Kim, Najoung (2021). "Compositional Linguistic Generalization in Artificial Neural Networks". PhD thesis. Johns Hopkins University. URL: http://jhir.library.jhu.edu/handle/1774.2/66745.

Kim, Najoung, Tal Linzen, and Paul Smolensky (2022). *Uncontrolled Lexical Exposure Leads to Overestimation of Compositional Generalization in Pretrained Models*. URL: http://arxiv.org/abs/2212.10769.

Klein, Ewan (1980). "A Semantics for Positive and Comparative Adjectives". In: *Linguistics and Philosophy* 4.1, pp. 1–45. DOI: 10.1007/BF00351812.

Kluck, Marlies (2007). "Optimizing Interpretation from a Generative Lexicon: a case study of Metonymic Type Coercion in modified nouns". In: *Proceedings of the Fourth International Workshop on Generative Approaches to the Lexicon*. Ed. by K. Kanzani, P. Bouillon, and L. Danlos.

Knobe, Joshua, Sandeep Prasada, and George E. Newman (2013). "Dual character concepts and the normative dimension of conceptual representation". In: *Cognition* 127.2, pp. 242–257. URL: https://www.sciencedirect.com/science/article/pii/S0010027713000139.

Krifka, Manfred, Francis Jeffry Pelletier, Gregory N. Carlson, Alice Ter Meulen, Godehard Link, and Gennaro Chierchia (1995). "Genericity: An Introduction". In: *THE GENERIC BOOK, Carlson, Gregory N., & Pelletier, Francis Jeffry [Eds], Chicago, IL: The U of Chicago Press, 1995, pp 1-124*. ISBN: 978-0-226-09292-8.

Lakoff, George (1987). *Women, fire, and dangerous things: what categories reveal about the mind*. Chicago: University of Chicago Press. xvii+614. ISBN: 978-0-226-46803-7.

Lalisse, Matthias and Ash Asudeh (2015). "Distinguishing intersective and non-intersective adjectives in compositional distributional semantics". Master's Thesis. Oxford: University of Oxford. URL: https://pages.jh.edu/mlaliss1/media/MLalisse_thesis.pdf.

Lasersohn, Peter (2012). "Contextualism and compositionality". In: *Linguistics and Philosophy* 35.2, pp. 171–189. URL: https://doi.org/10.1007/s10988-012-9115-z.

Leech, Geoffrey, Paul Rayson, and Andrew Wilson (2014). *Word Frequencies in Written and Spoken English: based on the British National Corpus*. London: Routledge. 320 pp. ISBN: 978-1-315-84016-1. DOI: 10.4324/9781315840161.

Link, Godehard (1983). "The Logical Analysis of Plurals and Mass Terms: A Lattice-Theoretical Approach". In: *Meaning, Use, and Interpretation of Language*.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. URL: http://arxiv.org/abs/1907.11692.

Lyu, Qing, Hua Zheng, Daoxin Li, Li Zhang, Marianna Apidianaki, and Chris Callison-Burch (2022). *Is "My Favorite New Movie" My Favorite Movie? Probing the Understanding of Recursive Noun Phrases*. URL: http://arxiv.org/abs/2112.08326.

Maienborn, Claudia (2020). "Revisiting Olga, the beautiful dancer: An intersective A-analysis". In: *Semantics and Linguistic Theory*, pp. 63–82. URL: https://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/30.63.

Martin, Joshua (2022). "Compositional Routes to (Non)Intersectivity". PhD thesis. United States – Massachusetts: Harvard University. 252 pp. URL: https://www.proquest.com/docview/2681380157/abstract/58A63B8C3E6548AEPQ/1.

Matushansky, Ora (2008). "On the Attributive Nature of Superlatives". In: *Syntax* 11.1, pp. 26–90. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9612.2008.00101.x.

McCurdy, Kate, Paul Soulos, Paul Smolensky, Roland Fernandez, and Jianfeng Gao (2024). "Toward Compositional Behavior in Neural Models: A Survey of Current Views". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. EMNLP 2024. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Miami, Florida, USA: Association for Computational Linguistics, pp. 9323–9339. URL: https://aclanthology.org/2024.emnlp-main.524/.

Miller, George A. (1995). "WordNet: a lexical database for English". In: *Commun. ACM* 38.11, pp. 39–41. URL: https://dl.acm.org/doi/10.1145/219717.219748.

Misra, Kanishka and Najoung Kim (2023). "Abstraction via exemplars? A representational case study on lexical category inference in BERT". In: *BUCLD 48: Proceedings of the 48th Annual Boston University Conference on Language Development*. BUCLD 48. Boston, USA: Cascadilla Press.

Montague, Richard (1970). "English as a formal language". In: *Linguaggi nella Societá e nella Tecnica*. Ed. by Bruno Visentini. Milan: Edizioni di Comunitá, pp. 189–224.

Morzycki, Marcin (2016). *Modification*. Cambridge University Press. 347 pp. ISBN: 978-1-107-00975-2.

Moskvoretskii, Viktor, Alexander Panchenko, and Irina Nikishina (2024). "Are Large Language Models Good at Lexical Semantics? A Case of Taxonomy Learning". In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. LREC-COLING 2024. Ed. by Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue. Torino, Italia: ELRA and ICCL, pp. 1498–1510. URL: https://aclanthology.org/2024.lrec-main.133.

Murphy, Gregory L. (2002). *The Big Book of Concepts*. Cambridge, Mass: The MIT Press. ISBN: 978-0-262-28035-8. DOI: 10.7551/mitpress/1602.001.0001.

Nayak, Neha, Mark Kowarsky, Gabor Angeli, and Christopher D. Manning (2014). *A Dictionary of Nonsubsective Adjectives*. CSTR 2014-04. Department of Computer Science, Stanford University. URL: https://www-cs.stanford.edu/~angeli/papers/2014-tr-adjectives.pdf.

Nefdt, Ryan M. (2020). "A Puzzle concerning Compositionality in Machines". In: *Minds and Machines* 30.1, pp. 47–75. URL: https://doi.org/10.1007/s11023-020-09519-6.

Nikishina, Irina, Polina Chernomorchenko, Anastasiia Demidova, Alexander Panchenko, and Chris Biemann (2023). "Predicting Terms in IS-A Relations with Pre-trained Trans-

formers". In: *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*. Findings 2023. Ed. by Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi. Nusa Dua, Bali: Association for Computational Linguistics, pp. 134–148. URL: https://aclanthology.org/2023.findings-ijcnlp.12.

O'Donnell, Timothy J. (2015). *Productivity and Reuse in Language: A Theory of Linguistic Computation and Storage*. Cambridge, Massachusetts, London, England: The MIT Press. ISBN: 978-0-262-02884-4. URL: https://doi.org/10.7551/mitpress/9780262028844.001.0001.

Oliver, Michael A. (2014). "Interpretation as Optimization: Constitutive material adjectives". In: *Lingua* 149, pp. 55–73. DOI: 10.1016/j.lingua.2014.05.004.

OpenAI et al. (2024). *GPT-4 Technical Report*. URL: http://arxiv.org/abs/2303.08774.

Ormerod, Mark, Jesús Martínez del Rincón, and Barry Devereux (2024). "How Is a "Kitchen Chair" like a "Farm Horse"? Exploring the Representation of Noun-Noun Compound Semantics in Transformer-based Language Models". In: *Computational Linguistics* 50.1, pp. 49–81. URL: https://aclanthology.org/2024.cl-1.3.

Pagin, Peter and Francis Jeffry Pelletier (2007). "Context, Content, and Composition". In: *Context-Sensitivity and Semantic Minimalism: new essays on semantics and pragmatics*. Ed. by Gerhard Preyer and Georg Peter. Oxford University Press, pp. 25–62. URL: https://id.lib.harvard.edu/alma/990113196480203941/catalog.

Parsons, Terence (1972). "Some Problems Concerning the Logic of Grammatical Modifiers". In: *Semantics of Natural Language*. In collab. with Donald Davidson and Gilbert Harman. Synthese Library. Dordrecht: Springer Netherlands, pp. 127–141. ISBN: 978-90-277-0310-1. DOI: 10.1007/978-94-010-2557-7_5.

Partee, Barbara H. (1984). "Compositionality". In: *Varieties of Formal Semantics*. Ed. by Fred Landman and Frank Veltman. GRASS 3. Dordrecht: Foris, pp. 281–311.

— (2007). "Compositionality and coercion in semantics: The dynamics of adjective meaning". In: *Cognitive foundations of interpretation*, pp. 145–161.

— (2008). *Compositionality in Formal Semantics: Selected Papers*. John Wiley & Sons. 347 pp. ISBN: 978-0-470-75129-9.

— (2009). "Formal semantics, lexical semantics, and compositionality: The puzzle of privative adjectives". In: *Philologia* 7.1, pp. 11–21. URL: http://www.philologia.org.rs/index.php/ph/article/view/216.

— (2010). "Privative adjectives: Subsective plus coercion". In: *Presuppositions and discourse: Essays offered to Hans Kamp*. Ed. by Thomas Zimmermann, Rainer Bauerle, and Uwe Reyle. Brill, pp. 273–285. URL: https://brill.com/downloadpdf/book/edcoll/9789004253162/B9789004253162-s011.pdf.

Pavlick, Ellie (2022). "Semantic Structure in Deep Learning". In: *Annual Review of Linguistics* 8.1, pp. 447–471. DOI: 10.1146/annurev-linguistics-031120-122924.

— (2025). "Not-Your-Mother's-Connectionism: LLMs as Cognitive Models". Harvard University. URL: https://www.youtube.com/watch?v=qroaay8Xukk.

Pavlick, Ellie and Chris Callison-Burch (2016a). "Most "babies" are "little" and most "problems" are "huge": Compositional Entailment in Adjective-Nouns". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. ACL 2016. Berlin, Germany: Association for Computational Linguistics, pp. 2164–2173. URL: https://aclanthology.org/P16-1204.

— (2016b). "So-Called Non-Subsective Adjectives". In: *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. *SEM 2016. Berlin, Germany: Association for Computational Linguistics, pp. 114–119. URL: https://aclanthology.org/S16-2014.

Pelletier, Francis Jeffry (2012). "Holism And Compositionality." In: *The Oxford Handbook of Compositionality*. Ed. by Wolfram Hinzen, Edouard Machery, and Markus Werning. Oxford University Press. ISBN: 978-0-19-954107-2. URL: https://doi.org/10.1093/oxfordhb/9780199541072.013.0007.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "GloVe: Global Vectors for Word Representation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2014. Ed. by Alessandro Moschitti, Bo Pang, and Walter Daelemans. Doha, Qatar: Association for Computational Linguistics, pp. 1532–1543. URL: https://aclanthology.org/D14-1162.

Prasada, Sandeep and Elaine M. Dillingham (2006). "Principled and statistical connections in common sense conception". In: *Cognition* 99.1, pp. 73–112. URL: https://www.sciencedirect.com/science/article/pii/S0010027705000247.

— (2009). "Representation of Principled Connections: A Window Onto the Formal Aspect of Common Sense Conception". In: *Cognitive Science* 33.3, pp. 401–448. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1551-6709.2009.01018.x.

Prince, Alan and Paul Smolensky (2004). *Optimality theory: constraint interaction in generative grammar*. Malden, MA: Blackwell Pub. xi+289. ISBN: 978-1-4051-1932-0.

Pustejovsky, James (1991). "The Generative Lexicon". In: *Computational linguistics - Association for Computational Linguistics* 17.4, pp. 409–441.

— (2013). "Inference Patterns with Intensional Adjectives". In: *Proceedings of the 9th Joint ISO - ACL SIGSEM Workshop on Interoperable Semantic Annotation*. ISA 2013. Potsdam, Germany: Association for Computational Linguistics, pp. 85–89. URL: https://aclanthology.org/W13-0509.

Pylkkänen, Liina (2019). "The neural basis of combinatory syntax and semantics". In: *Science* 366.6461, pp. 62–66. URL: https://www.science.org/doi/10.1126/science.aax0050.

Quilty-Dunn, Jake, Nicolas Porot, and Eric Mandelbaum (2023). "The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences". In: *Behavioral and Brain Sciences* 46, e261. URL: https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/best-game-in-town-the-reemergence-of-the-languageofthought-hypothesis-across-the-cognitive-sciences/76F46784C6C07FF52FF45B934D6D3542.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. URL: https://www.R-project.org/.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019). "Language Models are Unsupervised Multitask Learners". In: p. 24. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu (2020). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". In: *Journal of Machine Learning Research* 21.140, pp. 1–67. URL: http://jmlr.org/papers/v21/20-074.html.

Rambelli, Giulia, Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci (2024). "Compositionality as an Analogical Process: Introducing ANNE". In: *The First Workshop on Analogical Abstraction in Cognition, Perception, and Language* (*Analogy-ANGLE*). URL: https://analogy-angle.github.io/assets/Rambelli_1.pdf.

Rambelli, Giulia, Emmanuele Chersoni, Claudia Collacciani, and Marianna Bolognesi (2024). "Can Large Language Models Interpret Noun-Noun Compounds? A Linguistically-Motivated Study on Lexicalized and Novel Compounds". In: *The First Workshop on Analogical Abstraction in Cognition, Perception, and Language* (*Analogy-ANGLE*). URL: https://analogy-angle.github.io/assets/Rambelli_2.pdf.

Ravichander, Abhilasha, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung (2020). "On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT". In: *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*. *SEM 2020. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 88–102. URL: https://aclanthology.org/2020.starsem-1.10.

Recanati, François (2010). *Truth-Conditional Pragmatics*. Oxford University Press. ISBN: 978-0-19-171022-3. DOI: 10.1093/acprof:oso/9780199226993.001.0001.

Reinert, Carolin (2024). "The compositionality of adjective noun constructions". PhD thesis. Frankfurt am Main: Johann Wolfgang Goethe-Universität zu Frankfurt am Main. URL: http://publikationen.ub.uni-frankfurt.de/frontdoor/index/index/docId/88057.

Ross, Hayley, Kathryn Davidson, and Najoung Kim (2024). "Is artificial intelligence still intelligence? LLMs generalize to novel adjective-noun pairs, but don't mimic the full human distribution". In: *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*. GenBench 2024. Ed. by Dieuwke Hupkes, Verna Dankers, Khuyagbaatar Batsuren, Amirhossein Kazemnejad, Christos Christodoulopoulos, Mario Giulianelli, and Ryan Cotterell. Miami, Florida, USA: Association for Computational Linguistics, pp. 131–153. URL: https://aclanthology.org/2024.genbench-1.9/.

Ross, Hayley, Najoung Kim, and Kathryn Davidson (2025). "Fake reefs are sometimes reefs and sometimes not, but are always compositional". In: *Experiments in Linguistic Meaning* 3, pp. 332–343. URL: https://journals.linguisticsociety.org/proceedings/index.php/ELM/article/view/5813.

Sathe, Aalok, Evelina Fedorenko, and Noga Zaslavsky (2024). "Language use is only sparsely compositional: The case of English adjective-noun phrases in humans and large language models". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 46.0. URL: https://escholarship.org/uc/item/0qd3662b.

Schumacher, Petra B., Patrick Brandt, and Hanna Weiland-Breckle (2018). "Online Processing of "Real" and "Fake": The Cost of Being Too Strong". In: *The Semantics of Gradability, Vagueness, and Scale Structure: Experimental Perspectives*. Ed. by Elena Castroviejo, Louise McNally, and Galit Weidman Sassoon. Language, Cognition, and Mind. Cham: Springer International Publishing, pp. 93–111. ISBN: 978-3-319-77791-7. URL: https://doi.org/10.1007/978-3-319-77791-7_4.

Siegel, Muffy E. A. (1980). *Capturing the adjective*. Outstanding dissertations in linguistics. New York: Garland Pub. xi+191. ISBN: 978-0-8240-4563-0.

Smith, Edward E. and Douglas L. Medin (1981). *Categories and concepts*. Cognitive science series 4. Cambridge, Mass.: Harvard University Press. viii+203. ISBN: 978-0-674-10275-0.

Smith, Edward E., Daniel N. Osherson, Lance J. Rips, and Margaret Keane (1988). "Combining Prototypes: A Selective Modification Model". In: *Cognitive science* 12.4, pp. 485–527. DOI: 10.1207/s15516709cog1204_1.

Song, Seokwon, Taehyun Lee, Jaewoo Ahn, Jae Hyuk Sung, and Gunhee Kim (2025). *Is a Peeled Apple Still Red? Evaluating LLMs' Ability for Conceptual Combination with Property Type*. URL: http://arxiv.org/abs/2502.06086.

Springer, Ken and Gregory L. Murphy (1992). "Feature Availability in Conceptual Combination". In: *Psychological Science* 3.2, pp. 111–117. URL: https://doi.org/10.1111/j.1467-9280.1992.tb00008.x.

Szabó, Zoltán Gendler (2012). "The case for compositionality". In: *The Oxford Handbook of Compositionality*. Ed. by Wolfram Hinzen, Edouard Machery, and Markus Werning. Oxford University Press. ISBN: 978-0-19-954107-2. URL: https://doi.org/10.1093/oxfordhb/9780199541072.013.0003.

Tat, Deniz (2011). "APs as reduced relatives: The case of Bir in some varieties of Turkic". In: *Proceedings of the 7th Workshop on Altaic Formal Linguistics (WAFL7)*. MIT Working Papers in Linguistics.

Team, MosaicML NLP (2023). *MPT-30B: Raising the bar for open-source foundation models*. Databricks. URL: https://www.databricks.com/blog/mpt-30b.

Ten Hacken, Pius (2016). *The Semantics of Compounding*. Cambridge University Press. 265 pp. ISBN: 978-1-107-09970-8.

Touvron, Hugo et al. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. URL: http://arxiv.org/abs/2307.09288.

Tremblay, Antoine and Harald Baayen (2010). "Holistic Processing of Regular Four-word Sequences: A Behavioural and ERP Study of the Effects of Structure, Frequency, and Probability on Immediate Free Recall". In: *Perspectives on Formulaic Language: Acquisition and Communication*, p. 151.

Truong, Thinh Hung, Timothy Baldwin, Karin Verspoor, and Trevor Cohn (2023). *Language models are not naysayers: An analysis of language models on negation benchmarks*. URL: http://arxiv.org/abs/2306.08189.

Vecchi, Eva Maria, Marco Marelli, Roberto Zamparelli, and Marco Baroni (2017). "Spicy Adjectives and Nominal Donkeys: Capturing Semantic Deviance Using Compositionality in Distributional Spaces". In: *Cognitive Science* 41.1, pp. 102–136. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/cogs.12330.

Walter, Elizabeth (2005). *Cambridge advanced learner's dictionary*. Ed. by Cambridge University Press. [Rev. ed.]. Cambridge ; New York: Cambridge University Press. xii+1572. ISBN: 978-0-521-84378-2.

Webb, Taylor, Keith J. Holyoak, and Hongjing Lu (2023). "Emergent analogical reasoning in large language models". In: *Nature Human Behaviour* 7.9, pp. 1526–1541. URL: https://www.nature.com/articles/s41562-023-01659-w.

Wu, Zhibiao and Martha Palmer (1994). "Verb Semantics and Lexical Selection". In: *32nd Annual Meeting of the Association for Computational Linguistics*. ACL 1994. Las Cruces, New Mexico, USA: Association for Computational Linguistics, pp. 133–138. URL: https://aclanthology.org/P94-1019.

Yang, An et al. (2024). *Qwen2 Technical Report*. URL: http://arxiv.org/abs/2407.10671.

Yu, Lang and Allyson Ettinger (2020). "Assessing Phrasal Representation and Composition in Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. EMNLP 2020. Ed. by Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu. Online: Association for Computational Linguistics, pp. 4896–4907. URL: https://aclanthology.org/2020.emnlp-main.397.

Zehr, Jeremy and Florian Schwarz (2018). *PennController for Internet Based Experiments (IBEX)*. URL: https://doi.org/10.17605/OSF.IO/MD832.

Appendix A

Appendix to Chapter 3

A.1 Lexical entries for additional adjectives

A key argument of Chapter 3 is that the meaning of "privative" adjectives and the fact that they yield different inferences with different nouns can be explained without recourse to adjective-internal context-sensitivity or reference to qualia dimensions for all such adjectives except *fake*. This is illustrated with lexical entries for *stone* and *counterfeit*. This section provides additional lexical entries for *homemade*, *artificial*, *false* and *knockoff* to further make this point. We do not provide a lexical entry for *former*, which shows additional complexity with regards to e.g. scoping over possessives (Bigley, 2025; Morzycki, 2016; Reinert, 2024) and may yield privativity only as an implicature (Bigley, 2025).

A.1.1 *Homemade*

(62) Lexical entry for *homemade*

$[\![\text{homemade}]\!]^w$ =

E-structure: $\lambda N.\lambda x.\ \textsc{homemade}_w(N)(x)$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{hm}_w(x)) +$

$$\sum_{\textsc{p}\in\mathbb{P}(N)} \lambda x.(\neg(\textsc{hm}_w(x) \rightarrow \neg \textsc{p}_w(x)) \rightarrow \textsc{p}_w(x))$$

where $\text{HM}_w(x) = \exists y \exists z \exists e_1 [\text{PERSON}_w(y) \wedge \text{HOME}_w(z)(y) \wedge \text{MAKING}_w(e_1) \wedge \text{AGENT}(e_1)(y) \wedge \text{THEME}(e_1)(x) \wedge \text{LOCATION}(e_1)(z)]$

This lexical entry may be slightly too strict, as *homemade* may also have a meaning more like *handmade* and need not strictly all happen in a home (I can do some of my knitting in the park and still call the result a *homemade sweater*). Nevertheless it is sufficient to make the point that *homemade* contributes a meaning corresponding to a type $\langle s, \langle e, t \rangle \rangle$ property and can be either privative or subsective depending on whether it is possible to make the relevant noun at home.

(63) Derivation of privative inference for *homemade cat*, outlined:

a. Just as for *lion*, *cat* has a k-property of being biologically born of a cat.

b. This contradicts the meaning contributed by homemade which says that the *homemade cat* was made by a person.

c. By the Genericity Constraint applied to *cat* and *homemade cat*, a *homemade cat* is not a *cat* in all "normal" worlds.

A.1.2 *False*

False has multiple senses, as evidenced both by the items in our experiment and any dictionary entry for *false* (Walter, 2005 i.a.). The senses we will focus on here are the ones where *false* means untrue or incorrect with respect to the information content of the noun, such as in *false information*, *false rumor*, *false fact* etc., and the sense(s) where it means something like *insincere* or *unfounded*, as in *false laugh*, *false friend*, *false hope* and *false alarm*. There is a third sense of *false* that appears to simply mean *fake*, as used in *false teeth*, *false bottom* (in a container) and other seemingly fixed expressions. I will assume that that receives the same analysis as *fake*.

For the first sense, I will make use of a CONTENT operator to access the propositional content of nouns like *fact*; this is standing in for a fuller theory of such nouns.

(64) Lexical entry for *false*$_1$, untrue information content

$[\![\text{false}_1]\!]^w$ =

E-structure: $\lambda N.\lambda x.\ \textsc{false-1}_w(N)(x)$

C-structure: $\lambda N.\ (\lambda x.\ \neg\textsc{content}(x)) +$

$$\sum_{\textsc{p}\in\mathbb{P}(N)} \lambda x.(\neg(\neg\textsc{content}(x) \rightarrow \neg\textsc{p}_w(x)) \rightarrow \textsc{p}_w(x))$$

We expect this to yield a subsective inference for most such nouns, such as *rumor*, *impression* and *sign*; exceptions are *fact*, *proof* and some people's interpretation of *information*, which require their content to be true (i.e., *facts are true* is a k-property of *fact*).

For the second sense, I will use a MOTIVATION property similar to the AUTH-ORIGIN property we used for *counterfeit*. Like *counterfeit*, I will assume that this is part of the C-structure. Also like *counterfeit*, this results in this reading of *false* being privative in "normal" worlds in all cases where this property is a k-property (as opposed to a t-property) of the noun.

(65) Lexical entry for *false*$_2$, insincere/unfounded

$[\![\text{false}_2]\!]^w$ =

E-structure: $\lambda N.\lambda x.\ \textsc{false-2}_w(N)(x)$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{fls}_w(N)(x)) +$

$$\sum_{\textsc{p}\in\mathbb{P}(N)} \lambda x.(\neg(\textsc{fls}_w(N)(x) \rightarrow \neg\textsc{p}_w(x)) \rightarrow \textsc{p}_w(x))$$

where $\textsc{fls}_w(N)(x) = \exists Q \in \mathbb{P}(N)[\textsc{motivation}(Q) \wedge \textsc{seem-like}_w(x, Q) \wedge \neg Q(w)(x)]$

In other words, x is a false$_2$ N (in a "normal" world / according to the C-structure) if there is some dimension of N Q which is a motivation of x which seems like it holds of x but does not actually hold.

We expect this to yield a subsective inference for things like *laugh* or *alarm*, where Q will be something like the (genuine) humor that normally motivates a laugh or a genuine cause for alarm, and a privative inference for *friend*, where Q targets the motivation or sincerity that normally underlies the actions of friendship, which our experiment participants conclude are necessary for friendship. This suggests that this motivation/sincerity property, however

exactly it is expressed, is a k-property of *friend* but not of *laugh* or *alarm*, where it is only a t-property. (Performance of laughing or the sounding of the alarm appear to be sufficient, and the relevant k-properties.)

A.1.3 *Knockoff*

For *knockoff*, we can implement a similar lexical entry to *counterfeit*, except that deception is not necessary. While *knockoff* can be used with an identical meaning to *counterfeit*, and is preferred over *counterfeit* by some speakers and in some registers[1], it can also used to be refer to (usually cheap) imitations which are not pretending to be the brand or object they are imitating, but just have some of its qualities or appearance.

(66) Lexical entry for *knockoff*

$[\![\text{knockoff}]\!]^w =$

E-structure: $\lambda N.\lambda x.\ \textsc{knockoff}_w(N)(x)$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{kn}_w(N)(x)) + \sum_{P \in \mathbb{P}(N)} \lambda x.(\neg(\textsc{kn}_w(N)(x) \rightarrow \neg P_w(x)) \rightarrow P_w(x))$

where $\textsc{kn}_w(N)(x) = \exists Q \in \mathbb{P}(N)[\textsc{auth-origin}_w(Q) \wedge \textsc{seem-like}_w(x, Q) \wedge \neg Q(w)(x)]$

A.1.4 *Artificial*

The meaning of *artificial* for natural kinds (and even for *intelligence*) is something like "man-made", but it takes on a more complex meaning with abstract nouns like *scarcity*. While a scarcity of some product due to human error in the supply chain is caused by humans, we prefer to use *artificial scarcity* for examples like the scarcity of hand sanitizer in (67), which involve (malicious or ethically questionable?) intent. This example also shows that *artificial* can be gradable in the right context.

[1]Thanks to our former lab manager Chloe Frey for bringing this to my attention.

(67) The scarcity of hand sanitizer during the COVID-19 pandemic was even more artificial than the scarcity of baking supplies – at least the lack of flour wasn't caused by suppliers hoarding it to sell at high prices.

We will capture this as there being a person who was the agent of the making event of x and that the making event was intentional. (See Guerrini (2024b) for an example of intent semantics using possible worlds; in our case the agent is existentially closed rather than contextually specified.)

(68) Lexical entry for *artificial*

$[\![\text{artificial}]\!]^w =$

E-structure: $\lambda N.\lambda x.\ \textsc{artificial}_w(N)(x)$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{af}_w(N)(x)) + \sum_{P \in \mathbb{P}(N)} \lambda x.(\neg(\textsc{af}_w(N)(x) \rightarrow \neg P_w(x)) \rightarrow P_w(x))$

where $\textsc{af}_w(N)(x) = \exists y \exists e_1[\textsc{person}_w(y) \wedge \textsc{making}_w(e_1) \wedge \textsc{agent}(e_1)(y) \wedge \textsc{theme}(e_1)(x) \wedge \textsc{intentional}(e_1, x)]$

This inclusion of intent may seem problematic for things like *artificial intelligence*, where we are calling a large variety of systems "artificial intelligence" even though it is not clear that their creators intend them to be *intelligent* in the full sense of the word. In fact, in AI contexts specifically, the concept of *intelligence* is probably reduced to something more watered down that better captures its use in this field at the moment. This can be achieved by mod_l (Section 3.6.1).

A.2 Inference derivation for recursive adjective modification

Section 3.3.3 gives the example of *plastic edible flowers* in (43), repeated here, to demonstrate that we need compositional C-structure because privativity can occur over composite noun phrases. *Edible flowers* refers to types of flower like rose or hibiscus, which can safely be eaten by humans and taste pleasant. We can then depict these types of flower as plastic models,

resulting in *plastic edible flowers*. Notably, the final result is no longer edible, since it is made of plastic.

(43) Context: A store sells many kinds of plastic and glass flowers. A chef is looking to decorate his restaurant only with (imitations of) edible flowers, such as roses and hibiscus, which he also uses in his cooking. He prefers plastic ones since they are less likely to break. He asks the store owner where to find them.
"You can find some plastic edible flowers over there."

Suppose that the lexical entry for *plastic* is the same as for *stone* and that *edible* is similarly simple, reducing to contributing a single dimension CAN-BE-(SAFELY)-EATEN-(AND-DIGESTED)-BY-HUMANS. Suppose further that *flower* involves a dimension that it is part of a plant and which entails that it is made of plant-stuff (which may or may not be edible). We derive a privative inference that *plastic edible flowers* are not *edible flowers* (though they are *plastic flowers*)

(69) Derivation of a privative inference for *plastic edible flowers*

a. When evaluating *edible flower*, *edible* contributes an edibility dimension to *flower* and retains all the dimensions of flower, since nothing about flowers is inherently contradictory with being edible.

b. Via the Genericity Constraint applied to *edible flower* and the Subsective Meaning Postulate applied to *flower*, *edible flowers* in "normal" worlds are a subset of *flowers* which are also edible.

c. When evaluating *plastic* [*edible flower*], *plastic* contributes a dimension that it is made of plastic. This contradicts multiple dimensions of *edible flower*, including being *edible* and being made of plant-stuff, so these are excluded by the conflict resolution mechanism from the C-structure of *plastic edible flower*.

d. Via the Genericity Constraint applied to *plastic edible flower*, plastic edible flowers are not edible in normal worlds, and by and the Genericity Constraint applied to *edible flower*, edible flowers are edible in normal worlds. By modus tollens, plastic edible flowers are not edible flowers in normal worlds. Likewise, by application of the Genericity Constraint to *flower*, plastic edible flowers are not flowers in normal worlds.

e. Aside: By application of the Subsective Meaning Postulate to *plastic flower*, plastic edible flowers are plastic flowers in normal worlds.[2]

A.3 The effect of *literally*

Del Pinal notes that *literally*, as used in (70), can cause us to "withhold Non-Vacuity Principle-guided meaning modulations", under Partee's account.

(70) Something unbelievable happened at MIT. Scientists discovered a way of making, literally, stone lions and rubber rabbits.

If the Non-Vacuity Principle does not apply in these cases, how do we get a meaning out of *stone lion* that is not contradictory? One possibility put forward by del Pinal (2015) is that we do get an empty set in the actual world, but *literally* instructs us not to care. Another interpretation is that *literally* adjusts our conception of the evaluation world such that (the extensions of) *stone* and *lion* are no longer contradictory in this world. Thus, the Non-Vacuity Principle need not apply. Our account works nicely with this second view of *literally*: if ("normal") worlds involving *literally* do not yield a contradiction between *stone* and *lion*, then the conflict resolution mechanism keeps *lion* and all its k-properties and so we get a subsective inference

[2] In our derivation with these toy lexical entries, it may look like edible has contributed nothing to *plastic edible flower*, since it only contributes edibility and *plastic* discards this. In reality, the interaction between *edible* and *flower* is going to be more complex such that we likely end up with some t-properties relating to what edible flowers look like, capturing the fact that *edible flowers* are typically roses, hibiscus etc. These t-properties will get retained when *plastic* is applied, meaning that *plastic edible flowers* are a proper subset of *plastic flowers* (in normal worlds).

in this new, modified world of evaluation. Our account also explains why *fake* cannot yield a subsective inference with *literally*, as shown in del Pinal's two examples (71) (recall (50) where *fake gun* could not mean *gun that is a fake toy gun*) and (72), unlike a widening account which should treat *fake* and *stone* the same, and so allow a subsective inference for *fake* with *literally* even though the inference is ordinarily privative.

(71) Listen to this unbelievable story. Some immoral toy store owner was, literally, selling fake guns at his store.

(72) Something amazing happened at MIT. Some engineer managed to make, literally, a fake gun.

By assumption, *fake gun* is privative in typical contexts because *fake* targets some k-property of *gun* that is crucial to gun-hood, usually something like *guns can shoot*. Our account of *fake* differs from our account of *stone* in that the meaning of *fake* necessarily negates some k-property, even if the result is overall subsective for some k-properties. In this case, it negates $\textsc{can-shoot}(x)$. While *literally* can tell us to suspend belief about some contradictions, it is probably not telling us to accept that $\textsc{can-shoot}(x) \wedge \neg\textsc{can-shoot}(x)$ is true; the suspension of belief seems to be more about contingent "contradictions" like whether or not lions can be made of stone. So as long as we assume that the context is still requiring $\textsc{can-shoot}(x)$ as a necessary property for gun-hood (i.e. *literally* does not also modify that), then *fake gun* remains *not a gun* independent of *literally*.

A.4 Adjective-noun modification with decomposed E-structure

In Section 3.2 and Section 3.3 we assumed that the E-structure of both nouns and adjectives was strictly atomic and opaque. Since we are linking C-structure to k-properties and kinds, and k-properties do not actually hold of all instances of the noun in all worlds, some separation between C-structure and the E-structure is clearly necessary. Most theories of semantics assume that extensions/intensions are atomic and opaque.

In fact, we can assume that noun intensions are atomic without needing to assume that they are entirely opaque, and we need not assume that adjectives—as modifiers—are entirely atomic and opaque, even if they refer to some atomic concept as part of their meaning. If we are willing to relax this assumption for adjectives, then we can get transparent adjective lexical entries that allow us to derive inferences in the actual world directly by the same mechanism that we use in the C-structure, instead of via the Genericity Constraint. In this section, I will show a way of expressing the adjective's effect on the E-structure via (k-)properties in the same way as in the C-structure, at the expense of a little more machinery.

A.4.1 K-properties in the E-structure

We will continue to assume that for every noun N, there is some property N that describes its extension in w, such as $\lambda x.\, \textsc{lion}_w(x)$ for *lion*. Some k-properties of N will hold in w of all $x \in P(w)$. (If w is reasonable, there should be many of these, but not all: *lions have manes* will not hold for $\textsc{lion}_w$ in any w with female lions; it only holds for those (male) lions that meet the felicity conditions of the generic.) Let $\mathbb{P}_E^w(N)$ denote the set of such k-properties. It would be nice to assume that we can express N as the conjunction of this subset of k-properties that hold in w. However, (a) we don't know whether k-properties are fine-grained enough to achieve this and (b) we do not want to end up with a definitional theory. Prasada and Dillingham (2006) specifically assert that a conceptual representation cannot be reduced to combinations of features/property representations; rather, the type must be represented as having connections to these properties. It is unclear at this stage whether invoking the world variable, such that we would use a different set of k-properties and t-properties to express the extension at each world, would save us from the issues that classical, definitional theories of concepts have (Murphy, 2002; Smith and Medin, 1981). Instead, we will conservatively assume that the extension of N at w can be expressed as a conjunction of the k-properties and t-properties in $\mathbb{P}_E^w(N)$ plus possibly some other properties (of type $\langle s, \langle e, t\rangle\rangle$). Trivially, this can be achieved if one of those other properties is just the original atomic extension in w, as shown in (73).

(73) $\lambda x.N(w)(x) = \lambda x.N(w)(x) \wedge \bigwedge_{\text{P} \in \mathbb{P}_E^w(N)} \text{P}_w(x)$

where N is the original extension / E-structure of N in w and $\bigwedge$ represents a conjunction over all the properties $\text{P} \in \mathbb{P}_E^w(N)$ that hold of the extension of N in w. By definition of $\mathbb{P}_E^w$, the second half of this conjunction is redundant and so the equality holds.

The second half of this conjunction will be what adjective modification acts on, and will allow us to define subsets of N if the $N(w)(x) \wedge ...$ component of the noun meaning is preserved, and nonsubsective meanings if it is not. Suppose that we can retrieve this list of properties that identifies the E-structure ($\mathbb{P}_E^w(N)$ plus the other properties that determine the extension in w) using a new operator $\mathbb{E}_E^w(N)$.

Now, we can modify the adjective lexical entry schema from (31) to split the E-structure into an adjective meaning component and a conflict resolution component, just as in the C-structure.

(74) Revised schema for adjective lexical entries

$[\![\text{A}]\!]^{w,c} =$

E-structure: $\lambda N.\lambda x.f'_E(N)(c)(w)(x) \wedge$

$$\bigwedge_{\text{P} \in \mathbb{E}_E^w(N)} (\neg(f'_E(N)(c)(w)(x) \rightarrow \neg\text{P}_w(x)) \rightarrow \text{P}_w(x))$$

C-structure: $\lambda N.\ (\lambda x.f_C(N)(c)(w)(x)) +$

$$\sum_{\text{P} \in \mathbb{P}(N)} \lambda x.(\neg(f_C(N)(c)(w)(x) \rightarrow \neg\text{P}_w(x)) \rightarrow \text{P}_w(x))$$

where w is the current world, c is the context, and f'_E and f_C are adjective-specific functions. Unlike f_E in (31), f'_E now only represents the adjective's meaning contribution, rather than the whole effect of the adjective on the noun's E-structure.

The formulation of the E-structure looks rather intersective: some adjective meaning is being conjoined with some noun meaning. However, the adjective meaning can depend on N, in which case it would be at best subsective, and moreover we need not conserve all of the meaning of N if properties in $\bigwedge_{\text{P}\in\mathbb{E}_E^w(N)}$ get kicked out due to being contradictory, in which case the overall result would be nonsubsective. In particular, if we take $\mathbb{E}_E^w(N)$ to just be the conjunction of the original extension predicate (e.g. $\lambda x.\textsc{lion}(x)$) and some subset of k-properties, then the extension predicate will get removed due to contradiction whenever the combination is privative. In that case, we will be left with only the adjective meaning part conjoined with some smaller subset of k-properties. We will see this in action in Section A.4.2. As discussed in Section 3.2, this schema will not work as is for modal nonsubsective adjectives like *alleged* which do not want to make any commitments about the status of noun-hood in w_0, but do not yield any contradictions in w_0 that remove the relevant k-properties under this schema. Again, more research is needed here, but for now we can assume that nonsubsective adjectives simply don't use this schema and just use $\lambda N.\lambda x.f'E(N)(c)(w)(x)$ in their E-structure.

A.4.2 Derivation of inferences in the current world

We can now re-derive *stone bowl* and *stone lion* in this system, except that unlike in Section 3.3.2 we also derive the relevant inference in the actual world directly, without need for the Genericity Constraint and its felicity conditions, once we make assumptions about what k-properties hold of the extension in the actual world.

(75) Revised lexical entry for *stone*

$[\![\text{stone}]\!]^w$ =

E-structure: $\lambda N.\lambda x.\ \textsc{made-of-stone}_w(x) \wedge$

$$\bigwedge_{\text{P}\in\mathbb{E}_E^w(N)} (\neg(f'_E(N)(c)(w)(x) \rightarrow \neg\text{P}_w(x)) \rightarrow \text{P}_w(x))$$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{made-of-stone}_w(x)) +$

$$\sum_{\text{P}\in\mathbb{P}(N)} \lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\text{P}_w(x)) \rightarrow \text{P}_w(x))$$

As before we will assume a toy lexical entry for *bowl* that just includes BOWL-SHAPED and HOLDS-SHAPE (when poked/used) as its sole k-property dimensions. Let us assume for the purposes of illustration that BOWL-SHAPED holds of the extension (perhaps to some contextually supplied degree d which is a lower bar than for the C-structure) but that HOLDS-SHAPE does not, to accommodate fragile bowls, paper bowls and so forth.

(76) Derivation of a subsective inference for *stone bowl*

a. $[\![\text{bowl}]\!]^w$ =

E-structure: $\lambda x.\ \textsc{bowl}_w(x) \wedge \textsc{bowl-shaped}_w(x)$

C-structure: $(\lambda x.\ \textsc{bowl-shaped}_w(x)) + (\lambda x.\ \textsc{holds-shape}_w(x))$

b. $[\![\text{stone bowl}]\!]^w = [\![\text{stone}]\!]([\![\text{bowl}]\!])$ =

E-structure:

$\lambda x.\ \textsc{made-of-stone}_w(x) \wedge$

$(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg(\lambda x.\ \textsc{bowl}_w(x))) \rightarrow \textsc{bowl}_w(x)) \wedge$

$(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{bowl-shaped}_w(x)) \rightarrow \textsc{bowl-shaped}_w(x))$

$=\lambda x.\ \textsc{made-of-stone}_w(x) \wedge \textsc{bowl}_w(x) \wedge \textsc{bowl-shaped}_w(x)$

C-structure:

$(\lambda x.\ \textsc{made-of-stone}_w(x)) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{bowl-shaped}_w(x)) \rightarrow \textsc{bowl-shaped}_w(x))) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg(\textsc{holds-shape}_w(x))) \rightarrow \textsc{holds-shape}_w(x)))$

$=(\lambda x.\ \textsc{made-of-stone}_w(x)) + (\lambda x.\ \textsc{bowl-shaped}_w(x)) + (\lambda x.\ \textsc{holds-shape}_w(x))$

c. Subsective inference:

$\lambda x.\ \textsc{made-of-stone}_w(x) \wedge \textsc{bowl}_w(x) \wedge \textsc{bowl-shaped}_w(x) \subseteq \lambda x.\ \textsc{bowl}(x)$

For *stone lion*, let us assume that being born of a lion, being lion-shaped (to some sufficient contextually specified degree, again probably lower than the bar for being a lion in the C-structure) and being a mammal are k-properties that hold of the E-structure. Others may also hold, but this is enough for our derivation. I will assume that $\textsc{made-of-stone}(x) \rightarrow \neg\textsc{lion}_w(x)$ since I want to use $\textsc{lion}_w(x)$ as part of the E-structure and since we do empirically get a

privative inference. In this case, it is less that we "derive" the privative inference and more that we derive the meaning for *stone lion* in the case where it is privative. However, there is no need to assume this if the "other properties" that make up the E-structure are more fine-grained. I will show that assuming that MADE-OF-STONE does not entail a biological lion birth is sufficient to derive the privative inference.

(77) Derivation of a privative inference for *stone lion*

a. $[\![\text{lion}]\!]^w =$

E-structure:

$\lambda x.\ \textsc{lion}_w(x) \wedge \textsc{born-from-lion}(x) \wedge \textsc{lion-shaped}_w(x) \wedge \textsc{mammal}_w(x)$

C-structure:

$(\lambda x.\ \textsc{born-from-lion}(x)) + (\lambda x.\ \textsc{furry}_w(x)) + (\lambda x.\ \textsc{has-mane}_w(x)) +$

$(\lambda x.\ \textsc{lion-shaped}_w(x)) + (\lambda x.\ \textsc{mammal}_w(x))$

b. $[\![\text{stone lion}]\!]^w = [\![\text{stone}]\!]([\![\text{lion}]\!]) =$

E-structure:

$\lambda x.\ \textsc{made-of-stone}_w(x)) \wedge$

$(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{lion}_w(x)) \rightarrow \textsc{lion}_w(x)) \wedge$

$(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{born-from-lion}(x)) \rightarrow \textsc{born-from-lion}(x)) \wedge$

$(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{lion-shaped}_w(x)) \rightarrow \textsc{lion-shaped}_w(x)) \wedge$

$(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{mammal}_w(x)) \rightarrow \textsc{mammal}_w(x))$

$=\lambda x.\ \textsc{made-of-stone}_w(x)) \wedge \textsc{lion-shaped}_w(x)$

(C-structure continued overleaf)

C-structure:

$(\lambda x.\,\textsc{made-of-stone}_w(x)) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{b-f-l}(x)) \rightarrow \textsc{b-f-l}(x))) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{furry}_w(x)) \rightarrow \textsc{furry}_w(x))) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{has-mane}_w(x)) \rightarrow \textsc{has-mane}_w(x))) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{lion-shaped}_w(x)) \rightarrow \textsc{lion-shaped}_w(x))) +$

$(\lambda x.(\neg(\textsc{made-of-stone}_w(x) \rightarrow \neg\textsc{mammal}_w(x)) \rightarrow \textsc{mammal}_w(x)))$

$=(\lambda x.\,\textsc{made-of-stone}_w(x)) + (\lambda x.\,\textsc{lion-shaped}_w(x))$

c. Privative inference:

By assumption in the derivation, $\textsc{made-of-stone}(x) \rightarrow \neg\textsc{born-from-lion}(x)$ since stone lions are made from stone by sculptors and not born live from lions. By assumptions about what k-properties hold of the extension of *lion* and thus the construction of the E-structure, $\lambda x.\textsc{lion}(x) \rightarrow \textsc{born-from-lion}(x)$. By modus tollens, since the extension of *stone lion* contains $\textsc{made-of-stone}(x)$ as a conjunct, *stone lions* are not lions.

If the "other property" that holds of the extension besides the subset of k-properties is just $\lambda x.\textsc{lion}(x)$, then we predict that *stone lion* loses whatever distinguished the extension of lions from its "definition" via this subset of k-properties, since it is privative and so $\lambda x.\textsc{lion}(x)$ must get dropped. It is unclear whether this prediction is good—it depends on how fine-grained the k-properties are, what exactly it is that makes the extension special, and whether that matters for *stone lions*. If we need to preserve some of these extension-only properties, then we need $\mathbb{E}_E^w(N)$ to pick up some longer list of smaller lion-extension-specific properties. In principle, we could take $\mathbb{E}_E^w(P)$ to include all properties in w that happen to be true of the x in the extension. Determining this, however, would require checking all possible properties in w, a power set-strength operation (over countably infinitely many entities of type e in w) that might significantly increase the complexity of our theory. More research is needed here

to determine if there is any empirical evidence or theory-internal motivation that could shed light on this.

In both of these derivations, the arguments are essentially the same as in Section 3.3.2, except that by making appropriate assumptions about what k-properties hold of the extension, we can derive the inference in the actual world, not just in a "normal" world. The derivations for *counterfeit money*, *counterfeit watch* and *plastic edible flower* given in Section 3.3.2 and Section 3.3.3 go through in the same way.

A.4.3 Subsective inferences with *fake*

One advantage of this move is that we can now see how subsective inferences with *fake* are derived in the actual world. In Section 3.4, we observed that since *fake* targets k-properties and the Genericity Constraint requires that k-properties hold of nouns in sufficiently normal worlds (when their felicity conditions are met), *fake* is always privative in sufficiently normal worlds, unless the felicity conditions for the k-property are not met in some cases or *fake* applies to a narrowed meaning of the noun. In the actual world w_0, if we are not restricted by the felicity conditions, not all k-properties need hold. Section 2.7.4 in Chapter 2 observed a distinction between k-properties that always seem to result in privativity and other ones that *fake* can target subsectively. Only the former type of k-properties hold of the extension in w_0.

The context in (21), reprinted here, gives an example of a subsective inference using a k-property from Prasada and Dillingham (2006).

(21) Context: Adam wants to be a lifeguard at his local pool because he wants to make friends with a girl who often swims there. Adam can't actually swim, but he thinks he can still be a good lifeguard as long as he can throw the life ring to people. He fakes a swim test certificate and is able to get the job.
∴ Adam is a fake lifeguard.
∴ Adam is not a lifeguard.

Let us assume the following toy lexical entry for *lifeguard*, which has two k-properties CAN-SWIM and *has-lifeguard-role*, with only the latter holding in the E-structure. Let us further assume that *fake*'s contextually sensitive dimension selecting operator CHP targets CAN-SWIM in the context of (21). Finally, since we are using $\textsc{lifeguard}(x)$ in the E-structure, assume $\lambda x.\ \neg\textsc{can-swim}_w(x)$ does not entail $\lambda x.\ \neg\textsc{lifeguard}_w(x)$ (because the inference should be subsective).

(78) Lexical entry for *lifeguard* with toy dimensions:

$[\![\text{lifeguard}]\!]^{w,c} =$

E-structure: $\lambda x.\ \textsc{lifeguard}_w(x) \wedge \textsc{has-lifeguard-role}_w(x)$

C-structure: $(\lambda x.\ \textsc{has-lifeguard-role}_w(x)) + (\lambda x.\ \textsc{can-swim}_w(x))$

(79) Revised lexical entry for *fake*

$[\![\text{fake}]\!]^{w,c} =$

E-structure: $\lambda N.\ (\lambda x.\ \textsc{fk}_w(N)(c)(x)) \wedge$

$$\bigwedge_{\textsc{p}\in\mathbb{P}(N)} (\neg(\textsc{fk}_w(N)(c)(x) \rightarrow \neg\textsc{p}_w(x)) \rightarrow \textsc{p}_w(x))$$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{fk}_w(N)(c)(x)) +$

$$\sum_{\textsc{p}\in\mathbb{P}(N)} \lambda x.(\neg(\textsc{fk}_w(N)(c)(x) \rightarrow \neg\textsc{p}_w(x)) \rightarrow \textsc{p}_w(x))$$

where $\textsc{fk}_w(N)(x) = \neg(\textsc{ChP}(\mathbb{P}_k(N))(c))(x) \wedge \textsc{deception}_w(x, (\textsc{ChP}(\mathbb{P}_k(N))(c)))$

(80) Derivation of *fake lifeguard* with subsective inference in the context of (21)

$[\![\text{fake lifeguard}]\!]^{w,c} = [\![\text{fake}]\!]^{w,c}([\![\text{lifeguard}]\!]^{w,c})$

E-structure:

$$\neg\textsc{fk}(N)(c)(w)(x) \wedge \bigwedge_{\textsc{p}\in\mathbb{E}^w_E(\text{lifeguard})} (\neg(\textsc{fk}(N)(c)(w)(x) \rightarrow \neg\textsc{p}_w(x)) \rightarrow \textsc{p}_w(x))$$

where

$$\textsc{fk}_w(N)(x) = \neg(\textsc{ChP}(\mathbb{P}_k(N))(c))(x) \wedge \textsc{deception}_w(x, (\textsc{ChP}(\mathbb{P}_k(N))(c)))$$

$$= \lambda x.\, \neg\textsc{can-swim}_w(x) \wedge \textsc{deception}_w(x, \textsc{can-swim}_w) \wedge$$
$$(\neg((\neg\textsc{can-swim}_w(x) \wedge \textsc{deception}_w(x, \textsc{can-swim}_w)) \rightarrow$$
$$\neg\textsc{lifeguard}_w(x)) \rightarrow \textsc{lifeguard}_w(x)) \wedge$$
$$(\neg((\neg\textsc{can-swim}_w(x) \wedge \textsc{deception}_w(x, \textsc{can-swim}_w)) \rightarrow$$
$$\neg\textsc{has-lifeguard-role}_w(x)) \rightarrow \textsc{has-lifeguard-role}_w(x)) \wedge$$
$$(\neg((\neg\textsc{can-swim}_w(x) \wedge \textsc{deception}_w(x, \textsc{can-swim}_w)) \rightarrow$$
$$\neg\textsc{can-swim}_w(x)) \rightarrow \textsc{can-swim}_w(x))$$
$$= \lambda x.\, \neg\textsc{can-swim}_w(x) \wedge \textsc{deception}_w(x, \textsc{can-swim}_w) \wedge$$
$$\textsc{lifeguard}_w(x) \wedge \textsc{has-lifeguard-role}_w(x)$$
$$\subset \lambda x.\, \textsc{lifeguard}_w(x)$$

C-structure: $(\lambda x.\, \neg\textsc{can-swim}_w(x)) + (\lambda x.\, \textsc{deception}_w(x, \textsc{can-swim}_w)) +$
$(\lambda x.\, \textsc{has-lifeguard-role}_w(x))$ (*similarly*)

This derives the subsective inference since it preserves $\lambda x.\, \textsc{lifeguard}_w(x)$.

As before, give that we assume $\lambda x.\, \neg\textsc{can-swim}_w(x)$ does not entail $\lambda x.\, \neg\textsc{lifeguard}_w(x)$, this is perhaps less of a "derivation" and more of an illustration of the meaning; the point is that we can get a transparent meaning of *fake lifeguard* that is compatible with being subsective.

A.4.4 Subsective inferences when *fake* means *counterfeit*

This move does not solve the issue that we raised at the end of Section 3.4.2 that *fake* cannot mean *counterfeit* unless brand / authentic origin is a k-property of the noun. However, it does allow us to derive subsective inferences in those cases. This is still different from the mechanism we used to derive subsective inferences for *counterfeit*, where we argued that the targeted property was a t-property. If we want to be able to apply *fake* in those cases, we still need to contextually strengthen the meaning of the noun to have this as a (non-extension-determining) k-property, as discussed in Section 3.4.2.

Suppose that *watch*^{c} refers to a contextually restricted meaning of *watch* such as *designer watch* such that some authenticity property AUTH-WATCH is a k-property in c. Suppose further that ChP targets this property in c. Then we can derive a subsective inference as follows:

(81) Derivation of a subsective inference when *fake watch* means *counterfeit watch*

a. By assumption, AUTH-WATCH and TELLS-TIME are k-properties of *watch*^{c}, but only TELLS-TIME is also part of the E-structure.

b. By assumption about c, ChP($\mathbb{P}_k(\text{watch}^c)$)($c$) targets AUTH-WATCH in c, and so *fake watch*^{c} contributes $\neg$AUTH-WATCH(x) as part of its meaning.

c. By assumption about watches (equivalently, that the inference is supposed to be subsective), it is not the case that $\neg$AUTH-WATCH(x) entails $\neg$TELLS-TIME(x), and it is not the case that $\neg$AUTH-WATCH(x) entails $\neg$WATCH^c(x) (if using that in the E-structure).

d. Thus, the E-structure and C-structure of *fake watch* both contain TELLS-TIME(x). The E-structure also still contains WATCH^c(x) and we derive that a *fake watch*^{c} is still a *watch*^{c}.

A.4.5 *Real* enforces k-properties in the extension

Copying the effect of the C-structure to the E-structure also gives *real* a non-vacuous meaning on the E-structure without needing to resort to the Generic Felicity Meaning Postulate: it now enforces that the chosen k-property actually holds of the extension by conjoining it to the extension, in addition to vacuously adding it to the C-structure. As del Pinal (2018) notes, this is the same effect narrowing effect on the noun as when mod_I selects an enrichment operator that targets one dimension.

(82) Revised lexical entry for *real*

$[\![\text{fake}]\!]^{w,c} =$

E-structure: $\lambda N.\ (\lambda x.\ \textsc{rl}_w(N)(c)(x)) \wedge$

$$\bigwedge_{P \in \mathbb{P}(N)} (\neg(\textsc{rl}_w(N)(c)(x) \rightarrow \neg P_w(x)) \rightarrow P_w(x))$$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{rl}_w(N)(c)(x)) +$

$$\sum_{P \in \mathbb{P}(N)} \lambda x.(\neg(\textsc{rl}_w(N)(c)(x) \rightarrow \neg P_w(x)) \rightarrow P_w(x))$$

where $\textsc{rl}_w(N)(x) = (\textsc{ChP}(\mathbb{P}_k(N))(c))(x)$

A.5 Additional examples of unconstrained context-sensitivity

The examples in Section 3.5.2 for *fake* show that contextual variable accounts for *fake* (Section 3.5.1) overgenerate. It is also possible to construct similar examples for *counterfeit* and *artificial*, which also display both subsective and privative inferences and would presumably be handled similarly under these theories, since these adjectives are usually taken, like *fake*, to be privative by default.

In (83), we cannot use *counterfeit Canadian dollar* to mean [*counterfeit ~~US dollar~~*] ∩ *Canadian dollar*, as would be predicted by contextual variable accounts. (This example also applies to *fake* in its reading of *counterfeit*.)

(83) Context: A forger takes a genuine Canadian dollar and, through an intricate process, gives it all the watermarks and other features necessary to imitate a US dollar. Since only additional markings have been added, it is technically still valid as a Canadian dollar (just as Canadian dollars that have been scribbled on are valid).

a. #This is a counterfeit Canadian dollar.

b. This is a counterfeit US dollar.

c. This is a Canadian dollar.

In (84), we cannot use *artificial rock formation* to mean [*artificial ~~reef~~*] ∩ *rock formation*, as would be predicted by contextual variable accounts.

(84) Context: Marine biologists are looking to create more habitats for corals and endangered fish species. They find an extremely unusually shaped rock formation in the ocean which has the perfect shape to be converted into an artificial reef, by populating it with corals and other species. The project is a great success and the corals thrive in their new location.

a. #This is an artificial rock formation.

b. This is an artificial reef.

c. ?This is (among other things) a rock formation. (✓: Underneath the reef, there is a rock formation.)

These accounts presumably handle adjectives like *stone* using noun widening alone (since *stone* is typically taken to contribute a subsective, not a privative, meaning by default). Thus, these accounts do not need to have material adjectives (or any of the other adjectives that we showed can yield privative inferences, like color adjectives) take a contextually supplied variable, and so they do not overgenerate meanings for those adjectives in the way discussed here.

A.6 Constraining noun widening

In this section, I will briefly sketch how we can use the constrained contextual modulation operator mod_l from del Pinal (2018) to constrain an account of privativity via noun widening modeled after Partee (2010). This captures all our data as well as no longer overgenerating (much), although it does not do the best job of capturing the generalizations we drew from the data, as discussed in Section 3.1.

A.6.1 Constrained noun widening using mod_l

Recall from Section 3.6.1 that mod_l can widen or narrow the extension (E-structure) of a noun by reference to dimensions of the C-structure.

1. **Contextual widening**: mod_l can return a *dimension operator* that selects one or more k-property dimensions from the C-structure. The dimension operator replaces the E-structure of the noun with the conjunction of these dimensions, which will typically result in a wider meaning than before.

2. **Contextual narrowing:** mod_l can return an *enrichment operator* that selects one or more k-property dimensions from the C-structure. The enrichment operator conjoins these dimensions with existing the E-structure of the noun and sets this conjunction as the new E-structure. This results in a narrower meaning whenever the enrichment operator selects non-crucial properties.

Using the original four qualia-based dimensions from del Pinal (2018), one particular type of widening we can get is if mod_l chooses the perceptual dimension P via the dimension operator Q_P. In this case, we reduce the noun N to "looks (/ sounds / feels) like an N". This is exactly the reading we want for *lion* in *stone lion* and *gun* in *fake gun*, in the case where the gun just looks like a gun and does not actually function as one. Alternatively, if using a k-property C-structure as in Section 3.2, which we need if we want to capture the meaning of *fake* in that way, then mod_l will simply return a conjunction of one or likely multiple perception-related k-properties.

Partee (2010) argues that all adjectives considered here return a subset of the noun. Thus we get subsective inferences for free. Privative inferences are achieved by applying mod_l to widen the noun and then taking a subset of the widened noun. This is the same mechanism we used to derive context-dependent privative inferences for typically-subsective adjective-noun combinations in Section 3.6.2.

In this account, mod_l is the only thing in the E-structure domain that needs access to the C-structure; adjectives themselves do not use dimension operators that access the C-structure to build their extensional meaning the way they do in del Pinal (2018) (see Section A.6.2). However, we still need the C-structure to be compositional, since privativity also occurs in recursive adjective-noun modification. The *plastic edible flower* example given in (43), repeated here, illustrates this.

(43) Context: A store sells many kinds of plastic and glass flowers. A chef is looking to decorate his restaurant only with (imitations of) edible flowers, such as roses and hibiscus, which he also uses in his cooking. He prefers plastic ones since they are less likely to break. He asks the store owner where to find them.
You can find some plastic edible flowers over there.

Plastic edible flower is privative not just with regards to *flower* but also with regards to *edible flower*. mod_l must be applied to *edible flower* before *plastic applies* to get the reading "is plastic and looks like an edible flower". Applying mod_l just to the head noun would yield "is plastic and is edible and looks like a flower", which is not the reading we want (being plastic and being edible is generally a contradiction). Thus, *edible flower* must have a C-structure that mod_l can apply to and retrieve the perceptual dimensions from.

A.6.2 Adjective lexical entries and a subsective meaning for *fake*

For all the adjectives discussed in this dissertation, we can use very similar lexical entries to the ones discussed for our main account. Moreover, there is no need to posit an atomic, opaque E-structure under such an account since we are assuming that these adjectives are

subsective in all worlds. For the E-structure, we can simply posit that each adjective is of the form $f'_E(P)(c)(w)(x) \wedge P(c)(w)(x)$, as shown in schema (85). The C-structure is likewise assumed to be "subsective" and just conjoins the new meaning to the list of dimensions.

(85) Subsective schema for adjective lexical entries for a widening account

$[\![\text{A}]\!]^{w,c} =$

E-structure: $\lambda N.\lambda x.f'_E(N)(c)(w)(x) \wedge N(c)(w)(x)$

C-structure: $\lambda N.\ (\lambda x.f_C(N)(c)(w)(x)) + \sum_{\text{P} \in \mathbb{P}(N)} \lambda x.\text{P}_w(x)$

where w is the current world, c is the context, and f'_E and f_C are adjective-specific functions. Unlike f_E in (31), f'_E now only represents the adjective's meaning contribution, rather than the whole effect of the adjective on the noun's E-structure.

Using a subsective C-structure means that we need to extend the notion of widening to the C-structure to avoid contradictions. However, we need to make sure that dimensions "dropped" during widening are still available for *fake* to access and negate, so we cannot literally drop them before we pass the widened noun to the adjective. In this sketch of a constrained widening account, I will simply assume that when we apply widening to the C-structure, we "cross out" dimensions that are not selected by the dimension operator chosen by mod_l but do not delete them, meaning that operators over C-structure like ChP can still access them. I further stipulate that mod_l cannot choose crossed out dimensions as dimension operators for further iterations of widening applied to the adjective-noun phrase.

(86) **Contextual modulation for dual content semantics**

a. mod_l applies to tuples of the form $[\![\alpha]\!]^c_M = \langle [\![\alpha]\!]^c_{M_E}, [\![\alpha]\!]^c_{M_C} \rangle$. $mod_l(e, c)$ returns either a dimension operator that selects one or more k-property dimensions from the C-structure or an enrichment operator that selects one or more k-property dimensions from the C-structure.

b. A dimension operator has the following effects: it replaces the E-structure with the selected k-properties and it "crosses out" the k-properties in the C-structure that are not selected, making them unavailable to future applications of mod_l. They are still available to operators like ChP.

c. An enrichment operator has the following effect: it conjoins the E-structure with the selected k-properties. It has no effect on the C-structure.

d. mod_l may be applied to any node in the semantic tree at the time of composition, as in del Pinal (2018).[3]

The schema in (85) works even for the lexical entry we gave for *fake* in Section 3.4. This gives us a lexical entry in line with Partee's suggestion that *fake* is subsective, though contra Guerrini (2024b) who also gives a noun widening account but assumes that *fake* is inherently privative. Privative inferences with *fake* are derived by noun widening; this does not affect ChP's ability to select k-properties from the C-structure since noun widening via mod_l only applies to the extension (E-structure).

(87) Subsective lexical entry for *fake* for widening accounts

$[\![\text{fake}]\!]^{w,c} =$

E-structure: $\lambda N.\lambda x.\ \textsc{fake}_w(N)(c)(x)$

C-structure: $\lambda N.\ (\lambda x.\ \textsc{fk}_w(N)(c)(x)) \sum_{P \in \mathbb{P}(N)} P_w(x)$

where $\textsc{fk}_w(P)(x) = \neg(\textsc{ChP}(\mathbb{P}_k(N))(c))(x) \wedge \textsc{deception}_w(x, (\textsc{ChP}(\mathbb{P}_k(N))(c)))$

and ChP is a choice operator that selects one or multiple dimensions from its argument $\mathbb{P}_k(N)$ (the k-properties of N) given a context c.

In other words, x is *fake* N in a context c (in "normal" worlds / according to the C-structure) if the k-property of N returned by the choice operator in this context, $Q = \textsc{ChP}(\mathbb{P}_k(N))(c)$, does not hold of x and there is an intent to deceive people that Q holds of x.

[3] Equivalently, it can be freely inserted at LF similar to exhaustification operators (Chierchia, Fox, and Spector, 2012).

A.6.3 Avoiding overgeneration

Restricting mod_l to C-structure dimensions immediately explains the first concern in del Pinal (2018) and Martin (2022) that we should not be able to widen *gun* to include any non-gun, including *giraffes* which are a kind of non-gun. It also offers an explanation for the *fake* (*toy*) *gun* problem raised by del Pinal (2018) in (50), repeated here:

(50) Context (adapted from del Pinal, 2018): Some terrorists planned a mass shooting at a Wild West-themed costume party. They expected a lot of people to dress up as cowboys and carry toy guns as part of their costume. The terrorists also dressed up as cowboys, but they disguised the gun they wanted to use for the shooting to look like a toy gun, so that they could sneak it into the party.

a. #I heard some disturbing news. Some terrorists constructed fake guns and planned to use them to attack a costume party.

b. I heard some disturbing news. Some terrorists constructed fake toy guns and planned to use them to attack a costume party.

We avoid this issue because *toy-ness* is not part of the C-structure of *gun*. This has two effects: firstly, and most importantly, *fake* cannot target toy-ness to negate as a k-property. It must target some actual k-property of guns; what is most salient here is still that they can or cannot shoot, which ultimately yields a privative inference. Secondly, mod_l can also not widen *gun* to mean *toy gun*; the closest it can get is the perceptual dimensions of *gun* to yield "looks like a gun". And in fact, since *fake gun* is privative and some widening must occur, this is exactly what it does. But all that happens to this widened meaning of the noun in our lexical entry for *fake* is that it gets conjoined with $\neg$CAN-SHOOT(x); it does not itself get negated. Thus (50a) is correctly predicted to not yield the target reading.

Appendix B

Appendix to Chapter 4

B.1 Generated contexts

To illustrate the qualitative analysis of Experiment 7 in Section 4.5.4, we provide a representative context for *homemade cat* and two for *fake leg*. We further show 9 randomly sampled contexts generated by Llama 3 70B Instruct, one for each of the 6 typically-privative adjectives, and one context generated for the typically-subsective adjectives *illegal*, *useful* and *multicolored*. (Contexts for the typically-subsective adjectives tend to follow an uninteresting, predictable pattern; we stumbled across this interesting supposedly-privative example involving *multicolored*).

1. **Context:** Ben's mom is a talented seamstress, and she helps him make a **homemade cat** out of old t-shirts and fabric remnants. They stuff the cat with polyester fiberfill and add a sweet face made from felt. Ben names the cat Stitch and takes it to school for show-and-tell. His classmates ooh and ahh over Stitch's colorful patches and softness. **LLM Rating:** Probably not (privative)

2. **Context:** In the museum of medical history, a **fake leg** was on display as part of an exhibit on the evolution of prosthetics. The fake leg was made in the 19th century, and was crafted from wood and leather. It was a crude but functional device, with a simple hinge at the knee and a leather strap to hold it in place. Visitors marveled at the

primitive technology, and were amazed by how far prosthetic design had come since then. The fake leg was a testament to the ingenuity and perseverance of people with disabilities throughout history.

LLM Rating: Probably not (privative)

3. **Context:** Rachel, a young amputee, was thrilled to receive her new **fake leg** from the hospital. It was her first prosthetic, and she couldn't wait to start learning how to walk again. The fake leg was custom-made to fit her perfectly, with a soft, cushioned socket and a sturdy metal limb. With the help of her physical therapist, Rachel quickly learned to navigate the fake leg, and soon she was running and playing with her friends like she used to.

   LLM Rating: Probably yes (subsective)

4. **Context:** At the sci-fi convention, attendees are amazed by an **artificial painting** that appears to come to life. The painting, created by a team of digital artists, depicts a fantastical landscape with towering mountains and glowing forests. As people walk by, the painting responds by changing color, shape, and even emitting sounds and scents. The artificial painting is a huge hit, and people line up to experience its immersive, interactive magic.

   LLM Rating: Definitely yes (subsective)

5. **Context:** The detective was called to investigate a **counterfeit attack** on a popular online shopping website. The hackers had created fake websites that looked identical to the real one, tricking customers into entering their credit card information. The detective worked tirelessly to track down the culprits, following a trail of digital clues that led him from server to server. Finally, he identified the IP address of the hackers and worked with the authorities to shut down the operation, saving thousands of people from financial loss.

   LLM Rating: Definitely yes (subsective)

   Authors' note: The LLM interpreted *counterfeit attack* as a noun-noun compound here,

not as an adjective-noun combiniation (i.e. it is an attack involving counterfeits, not an attack that is counterfeit).

6. **Context:** Rachel is a set designer for a film production company. She needs a **fake stick** that looks like a real magic wand for an upcoming fantasy movie. She searches high and low and finally finds the perfect fake stick, complete with intricate carvings and a sparkling crystal at the tip. The fake stick becomes a key prop in the movie, and the actors love using it during filming. The movie goes on to become a huge hit, and Rachel's fake stick is praised as one of the most iconic props in the film.
   LLM Rating: Definitely yes (subsective)

7. **Context:** The antique shop owner was thrilled to find an old chest in the attic of his newly inherited store. As he rummaged through it, he found a stack of old bills that looked like they dated back to the 1800s. But as he examined them closer, he realized they were actually **false currency**, likely created by a counterfeiter from that era. Despite being worthless, the bills were still a fascinating piece of history, and the owner decided to display them in his shop as a curiosity.
   LLM Rating: Probably not (privative)

8. **Context:** Jack's **former reason** for going to the beach every summer was a faded, beach towel that his family had used for years. The towel had a fun, colorful design and a soft, worn-out texture, but it was more than just a towel - it was a symbol of family vacations and summer fun. Jack would wrap himself in it after a long day of swimming, but after his family stopped going to the beach, the towel was forgotten in the back of a closet, a reminder of his former reason for soaking up the sun.
   LLM Rating: Definitely not (privative)

9. **Context:** Lily's friend Emma lends her a **knockoff jewel** to wear to a formal dance. The knockoff jewel is a beautiful, long necklace with a big, fake pearl pendant. Lily loves the necklace and feels like a princess when she wears it. However, when she tries to return it to Emma, she realizes that she lost the clasp and can't find it anywhere. Emma

is upset, but Lily promises to buy her a new one to replace it.
LLM Rating: Definitely yes (subsective)

10. **Context:** In the small village of Ravenswood, a mysterious **illegal currency** began to circulate. The coins were made of a heavy, dark metal and featured an image of a raven on one side and a crescent moon on the other. They were stamped with a strange symbol that no one could decipher. At first, the villagers were wary of the illegal currency, but as more and more merchants began to accept it, it became a popular alternative to the official coins. However, the village elder warned that the authorities would soon crack down on the illegal currency, and those caught using it would face severe penalties.
LLM Rating: Definitely yes (subsective)

11. **Context:** In a small village, there lived a respected elder named Nalani who used a **useful fruit** - a sweet and nutritious mango with a yellow skin and creamy flesh. Nalani would dry mango slices to create a natural sweetener for her traditional recipes, and she would also use mango oil to nourish her skin and hair. The villagers would come to Nalani for beauty tips, and she would share her secrets of using the mango to promote healthy skin and hair. Nalani's wisdom and knowledge of the useful mango were cherished by the community, and she was revered as a guardian of traditional knowledge.
LLM Rating: Definitely yes (subsective)

12. **Context:** On the beach, Jack builds an intricate sandcastle, topped with a **multicolored image** of a dragon. The dragon's scales glint in the sunlight, made from tiny shells and pebbles in shades of green, blue, and purple. As the tide comes in, Jack watches as the dragon begins to dissolve, its multicolored image slowly disappearing into the waves. He smiles, knowing that the memory of his creation will stay with him long after the castle is gone.
LLM Rating: Probably not (privative)

B.2 Additional experiment results

B.2.1 Tabular numerical results

Detailed numerical results for the figures in the main body of the paper are shown in Table B.1 (percentage of privative vs. subsective inferences for bigrams in Experiment 5a), Table B.3 (accuracy on Experiment 5b), Table B.4 (accuracy within 1 SD on Experiment 6) and Table B.5 (JS divergences for Experiment 6 with log-probabilities). All but Table B.1 are located in the sections below which discuss results for base models, since they also contain the results for the base models.

Table B.1: *Percentage of privative vs. subsective inferences for bigrams in Experiment 5a (no context, 5-shot).*

Model	priv.	subs.	unsure
Human	29.7%	57.4%	12.8%
Qwen 2 72B Instruct	30.8%	69.2%	0.0%
Llama 3 70B Instruct	46.2%	53.8%	0.0%
Llama 2 70B Chat	42.3%	34.6%	23.1%
Mixtral 7x8B Instruct	80.8%	15.4%	3.9%
Llama 3 8B Instruct	57.7%	42.3%	0.0%
Llama 2 13B Chat	84.6%	15.4%	0.0%
Llama 2 7B Chat	69.2%	0.0%	30.8%
Llama 3 70B	69.2%	30.8%	0.0%
Llama 3 8B	38.5%	53.8%	7.7%
Llama 2 70B	19.2%	19.2%	61.5%
Llama 2 13B	0.0%	0.0%	100%
Llama 2 7B	88.5%	11.5%	0.0%

B.2.2 Accuracy within 1 SD for Experiment 5b

We also evaluate Experiment 5b, which predicts the inference given the context, using the more lenient accuracy within 1 SD of the human mean metric introduced in Section 4.5.1. Figure B.1 shows instruction-tuned models and Figure B.2 shows base models. Table B.2 shows the detailed numerical results. This more lenient metric does not penalise models which use "Unsure", provided that that is within human ratings. With this metric, performance scales uniformly with size for instruction-tuned models.

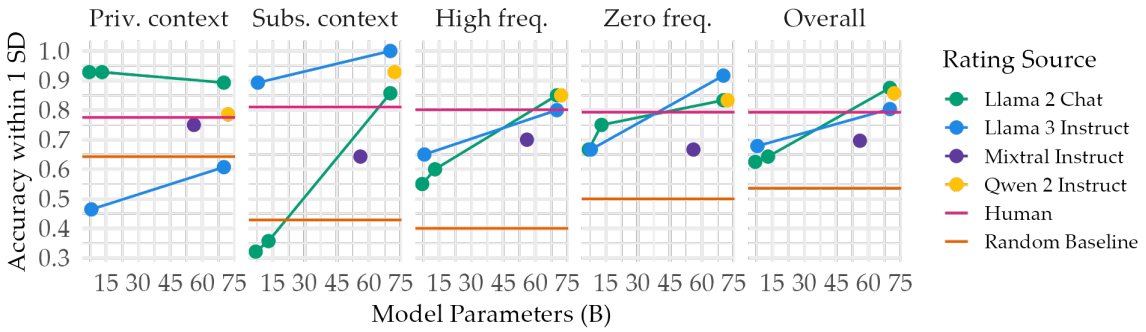


Figure B.1: *Accuracy within 1 SD of the human mean on the context-based inference task for instruction-tuned models (5-shot). Under this more lenient metric, accuracy increases with model size for all models.*

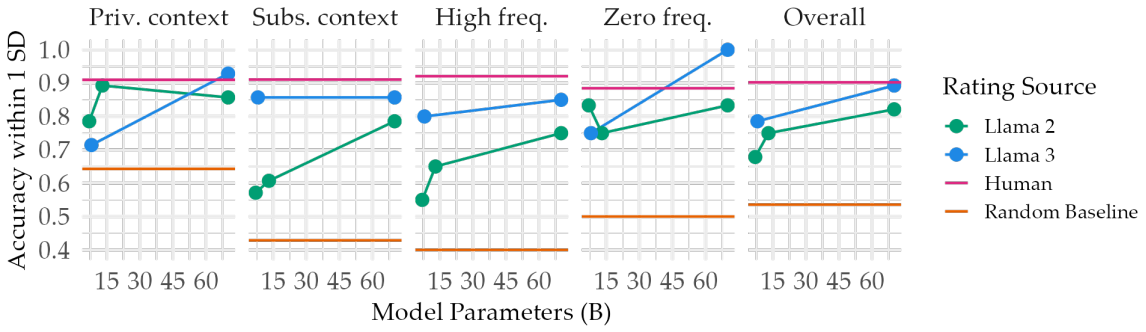


Figure B.2: *Accuracy within 1 SD of the human mean on the context-based inference task for base models (5-shot). Under this more lenient metric, overall accuracy increases with model size, though still not in every category for Llama 2.*

Table B.2: *Accuracy within 1 SD of human mean on the (5-shot) context-based inference task (Experiment 5b) overall, by bigram frequency and by context bias.*

	Accuracy				
Model	Priv. context	Subs. context	High freq.	Zero freq.	Total
Human	0.91	0.91	0.92	0.89	0.90
Llama 3 70B	0.93	0.86	0.85	1.00	0.89
Llama 2 70B Chat	0.89	0.86	0.85	0.83	0.88
Qwen 2 72B Instruct	0.79	0.93	0.85	0.83	0.86
Llama 2 70B	0.86	0.79	0.75	0.83	0.82
Llama 3 70B Instruct	0.61	1.00	0.80	0.92	0.80
Llama 3 8B	0.71	0.86	0.80	0.75	0.79
Llama 2 13B	0.89	0.61	0.65	0.75	0.75
Mixtral 7x8B Instruct	0.75	0.64	0.70	0.67	0.70
Llama 3 8B Instruct	0.46	0.89	0.65	0.67	0.68
Llama 2 7B Chat	0.79	0.57	0.55	0.83	0.68
Llama 2 7B	0.79	0.57	0.55	0.83	0.68
Llama 2 13B Chat	0.93	0.36	0.60	0.75	0.64
Random baseline	0.64	0.43	0.40	0.50	0.54

B.2.3 Experiment 5b: Base models

For the inference task where the context is provided, Experiment 5b, we see in Figure B.3 and Table B.3 that Llama 3 70B actually out-performs its Instruct model overall by 7 points of accuracy. This is largely due to its high accuracy on privative contexts (0.93 instead of 0.61), which compensates for its lower accuracy on subsective contexts (0.75 instead of 0.93). We see the same pattern for Llama 3 8B, which is better at privative contexts and worse at subsective contexts than its instruction-tuned counterpart. One possibility is that Llama 3 Instruct is generally more biased to affirmative ratings (i.e. subsective ratings on this scale) after instruction/helpfulness tuning, thus doing better in the subsective context simply because a subsective rating is correct there. We can see in Table B.1 that when no context is provided (Experiment 5a), Llama 3 Instruct assigns subsective and privative ratings roughly evenly across bigrams, while Llama 3 has a bias (69.2%) towards giving these bigrams with "privative" adjectives privative ratings. For Llama 3, we see scaling with size for both the accuracy metric and the accuracy within 1 SD metric, as shown in Figure B.3 and Figure B.2.

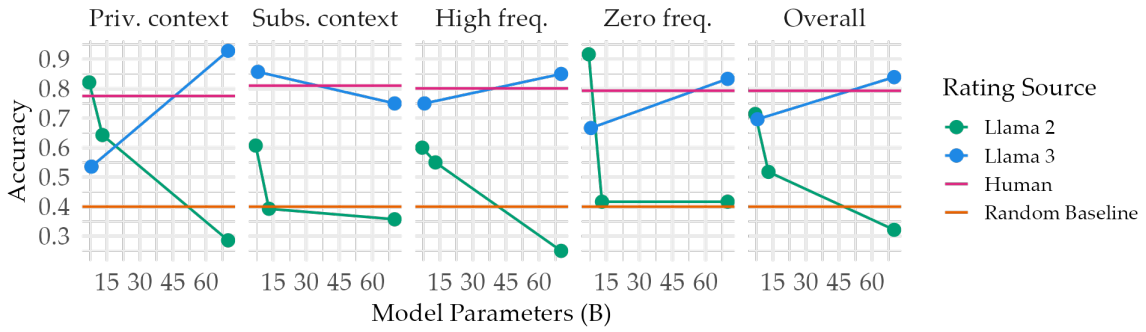


Figure B.3: *Accuracy on the context-based inference task for base models (5-shot). Accuracy increases with parameters for Llama 3, but drops for Llama 2. However, accuracy within 1 SD on this task increases with model parameters – see Figure B.2.*

Table B.3: *Accuracy on the (5-shot) context-based inference task (Experiment 5b) overall, by bigram frequency and by context bias.*

	Accuracy				
Model	Priv. context	Subs. context	High freq.	Zero freq.	Total
Human	0.78	0.81	0.80	0.79	0.79
Llama 3 70B	0.93	0.75	0.85	0.83	0.84
Qwen 2 72B Instruct	0.68	0.93	0.80	0.83	0.80
Llama 3 70B Instruct	0.61	0.93	0.80	0.75	0.77
Mixtral 7x8B Instruct	0.79	0.68	0.80	0.75	0.73
Llama 2 7B	0.82	0.61	0.60	0.92	0.71
Llama 3 8B	0.54	0.86	0.75	0.67	0.70
Llama 3 8B Instruct	0.29	0.96	0.60	0.58	0.63
Llama 2 13B Chat	0.89	0.36	0.60	0.67	0.63
Llama 2 70B Chat	0.50	0.68	0.70	0.42	0.59
Llama 2 13B	0.64	0.39	0.55	0.42	0.52
Llama 2 7B Chat	0.75	0.25	0.50	0.50	0.50
Llama 2 70B	0.29	0.36	0.25	0.42	0.32
Random baseline	0.4	0.4	0.4	0.4	0.4

This pattern does not occur in Llama 2 across the board - we see a marked improvement for Llama 2 7B over Llama 2 7B Chat in both context types, but Llama 2 13B Chat is better at privative contexts than Llama 2 13B and only slightly worse at subsective contexts, suggesting no overall pattern for Llama 2, and no effect of its type of instruction tuning. Likewise in Table B.1, we see that instruction-tuning Llama 2 70B does not result in a large shift in the ratio of privative to subsective ratings, but rather mostly reduces the proportion of "Unsure" ratings.

Notably, we actually see inverse scaling with size for Llama 2 on all splits of this task (see Figure B.3) when using the accuracy metric, but regular scaling with size using the accuracy within 1 SD metric, which does not punish the "Unsure" rating so harshly (Figure B.2).

B.2.4 Experiment 6: Base models

On the inference task where no context is provided, Experiment 6, we see in Figure B.4 and Table B.4 that Llama 3 70B, Llama 3 8B and Llama 2 70B perform comparably to their instruction-tuned counterparts using the within 1 SD metric. Performance is less predictable for the smaller Llama 2 models, with the ranking Llama 2 13B Chat > Llama 2 7B > Llama 2 13B > Llama 2 7B Chat. As discussed in the main body of the paper, this metric is quite lenient – the fact that Llama 2 7B Chat actually *underperforms* the random baseline, and that the smaller Llama 2 models underperform the "majority" baseline by at least 20 points of accuracy), is itself striking, suggesting a distinct lack of comprehension of the task when presented out of the blue with no context. While some of this difficulty may be attributed to the Likert scale, even these small models perform at at least 60% accuracy using the same metric and Likert scale (Table B.2) when a context is provided, so the difficulty must lie at least partially with the out-of-the-blue setting.

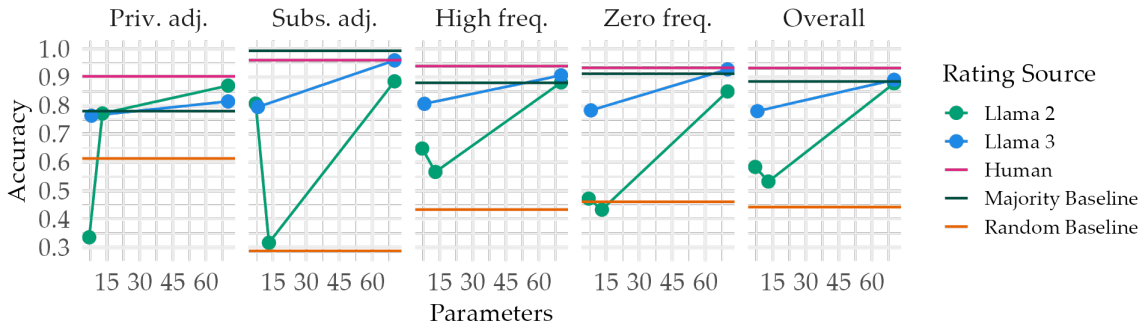


Figure B.4: *Accuracy within 1 SD of the human mean on the no-context inference task (Experiment 6) by bigram frequency and by adjective type (typically-privative or typically-subsective) for base models (5-shot).*

Perhaps these older, smaller models do not capture enough information about how the world typically is from their pretraining (do not have sufficiently human-like "priors" or

Table B.4: *Accuracy within 1 SD of the human mean on the (5-shot) no-context inference task (Experiment 6) by bigram frequency and by adjective type (typically-privative or typically-subsective).*

	Accuracy ($\mu \pm 1\sigma$)				
Model	Privative	Subsective	High freq.	Zero freq.	Total
Human	0.903	0.960	0.939	0.933	0.933
Qwen 2 72B Instruct	0.886	0.995	0.946	0.967	0.944
Llama 3 70B Instruct	0.778	0.995	0.892	0.911	0.892
Llama 3 70B	0.815	0.960	0.907	0.928	0.891
"Majority" baseline	0.781	0.993	0.881	0.912	0.885
Llama 2 70B	0.870	0.886	0.882	0.850	0.878
Llama 2 70B Chat	0.831	0.767	0.824	0.722	0.797
Mixtral 7x8B Instruct	0.653	0.914	0.803	0.756	0.791
Llama 3 8B	0.765	0.795	0.806	0.783	0.781
Llama 3 8B Instruct	0.833	0.679	0.735	0.789	0.752
Llama 2 13B Chat	0.455	0.776	0.706	0.517	0.624
Llama 2 7B	0.336	0.807	0.649	0.472	0.584
Analogy baseline	0.648	0.431	0.527	0.567	0.534
Llama 2 13B	0.772	0.317	0.566	0.433	0.533
Random baseline	0.610	0.325	0.464	0.456	0.460
Llama 2 7B Chat	0.447	0.252	0.394	0.267	0.345

world knowledge). Looking at the distributions they produce, Figure B.5 shows that they distribute their probability mass relatively evenly across the scale for subsective adjectives, resulting in a poor fit. Table B.5 shows the Jensen-Shannon divergences, which are relatively low for subsective adjectives for these models. Interestingly, however, models of all sizes are competitive for fitting the distribution of typically-privative adjectives overall, with Qwen 2 72B Instruct, Mixtral 7x8B Instruct and Llama 2 7B Chat scoring the lowest at around 0.3 JS divergence. Fitting the human distribution of privative adjectives is not a function of model size at all.

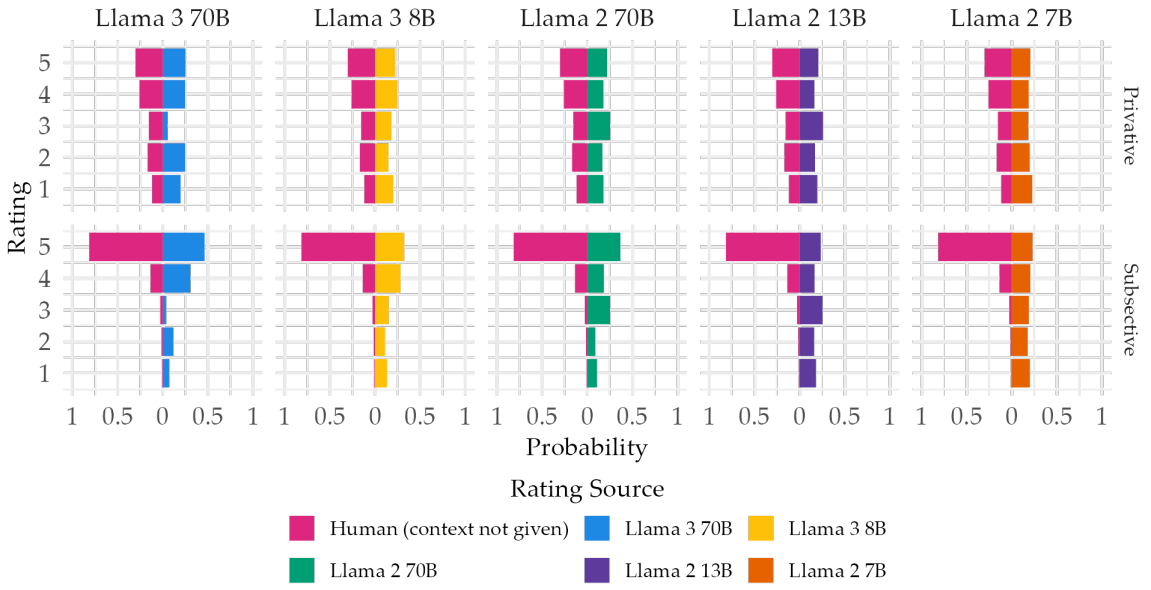


Figure B.5: *Average log-probability distribution for (typically) subsective vs. privative adjectives for base LLMs, compared to the average human distribution.*

Table B.5: *Jensen-Shannon divergence between per-bigram rating distributions for humans and LLMs when sourced from log-probabilities, for privative vs. subsective adjectives, including base models.*

	JS Divergence		
Model	Priv.	Subs.	Total
Human	0	0	0
Llama 3 70B Instruct	0.26	0.08	0.17
Qwen 2 72B Instruct	0.33	0.08	0.19
Llama 3 70B	0.16	0.21	0.19
Llama 2 70B Chat	0.18	0.25	0.22
Mixtral 7x8B Instruct	0.32	0.13	0.22
Llama 2 70B	0.17	0.30	0.24
Llama 3 8B	0.18	0.32	0.26
Llama 3 8B Instruct	0.18	0.34	0.26
Llama 2 13B Chat	0.25	0.35	0.30
Llama 2 7B	0.20	0.43	0.32
Llama 2 13B	0.21	0.43	0.32
Uniform baseline	0.20	0.46	0.34
Llama 2 7B Chat	0.29	0.46	0.38
“Majority” baseline	0.71	0.12	0.40

B.2.5 Experiment 5c: Zero-shot inferences with context

Experiment 5c performs an ablation study on Experiment 5b where we run the same experiment of determining the inference given a biasing context, but 0-shot instead of with 5-shot examples of "is-a" inferences on the Likert scale. Table B.6 shows the results of Experiment 5c.

Table B.6: *Accuracy on the zero-shot context-based inference task (Experiment 5c) overall, by bigram frequency and by context bias.*

	Accuracy				
Model	Priv. context	Subs. context	High freq.	Zero freq.	Total
Human	0.78	0.81	0.80	0.79	0.79
Qwen 2 72B Instruct	0.79	0.93	0.85	0.83	0.86
Llama 3 70B Instruct	0.86	0.82	0.80	0.83	0.84
Llama 3 70B	0.86	0.57	0.70	0.58	0.71
Llama 3 8B Instruct	0.64	0.75	0.70	0.67	0.70
Llama 2 70B	0.68	0.71	0.65	0.67	0.70
Mixtral 7x8B Instruct	0.75	0.64	0.70	0.67	0.70
Llama 2 70B Chat	0.82	0.53	0.55	0.83	0.68
Llama 2 7B Chat	0.89	0.39	0.55	0.75	0.64
Llama 2 7B	0.57	0.64	0.45	0.67	0.61
Llama 3 8B	0.89	0.25	0.40	0.75	0.57
Llama 2 13B Chat	0.93	0.21	0.50	0.58	0.57
Llama 2 13B	0.89	0.07	0.45	0.50	0.48
Random baseline	0.4	0.4	0.4	0.4	0.4

Whether the 5-shot examples help or hinder depend on the individual model. For Llama 3 70B Instruct, Llama 2 13B Chat, Llama 2 7B Chat and Mixtral 7x8B Instruct, we see a drop of 3-14 points in accuracy when we prompt the model 0-shot. For the other three instruct models, however, Llama 2 70B Chat, Llama 3 8B Instruct and Qwen 2 72B Instruct, we see a 6-9 point increase in accuracy when we prompt 0-shot, suggesting that these models found the 5 examples (which were examples of using the scale with "is-a" inferences, but were not exactly the target task which always repeated the noun) misleading rather than helpful.

Turning to the base models, we see a 4-13 point drop for all models (Llama 2 70B, Llama 3 8B, Llama 2 13B, Llama 2 7B) except Llama 2 70B when prompted 0-shot instead of 5-shot. Llama 2 70B shows a 38 point *increase* in accuracy on this task when prompted 0-shot. This is

because its very low score on the 5-shot task is largely caused by it answering "Unsure" for many bigrams, which is always considered incorrect under this metric. Without an example where "Unsure" is used, even though "Unsure" is mentioned as a scale item, Llama 2 70B uses this rating far less often and is able to score much higher on this task. This is in part an artefact of how this metric is scored – recall from Figure B.2 that Llama 2 70B still scores well on the more lenient accuracy within 1 SD metric, which includes "Unsure" for many bigrams.

B.2.6 Regression details for Section 4.5.4

Table B.7 shows the coefficients for the two regressions in Section 4.5.4.

Table B.7: *Coefficients for the regressions* `JSDivergence ~ AdjectiveType * HumanMean + HumanSD + BigramFrequency` *for each method in Section 4.5.4. For the adjective type factor, subsective is the first level (intercept), for frequency bins, 90th-99th percentile is the first level (intercept); dummy coding is used throughout.*

Coefficient	$\hat{\beta}$	p
Log-probability		
Intercept	0.61	$p < 0.01$
Privative	-0.40	$p < 0.01$
Human mean	-0.13	$p < 0.01$
Human SD	0.14	$p < 0.01$
75^{th}-90^{th} percentile	0.01	$p = 0.56$
50^{th}-75^{th} percentile	0.02	$p = 0.22$
25^{th}-50^{th} percentile	0.02	$p = 0.25$
Zero frequency	0.01	$p = 0.36$
Privative:Human mean	0.10	$p < 0.01$
Context generation		
Intercept	0.93	$p < 0.01$
Privative	-0.61	$p < 0.01$
Human mean	-0.19	$p < 0.01$
Human SD	0.15	$p < 0.01$
75^{th}-90^{th} percentile	0.02	$p = 0.38$
50^{th}-75^{th} percentile	0.03	$p = 0.15$
25^{th}-50^{th} percentile	0.03	$p = 0.16$
Zero frequency	0.03	$p = 0.09$
Privative:Human mean	0.16	$p < 0.01$

Appendix C

Appendix to Chapter 5

C.1 Analogy model details

C.1.1 Model configuration

As discussed in Section 5.4.1, the model has three configurable parameters: whether to do analogy over just nouns or also to include up to one adjective ("Noun only" vs."Noun + Adjective"), how many nearby bigrams to retain (k), and whether to return the memorized distributions from the training set when asked about a bigram in the training set, or to apply the algorithm as if that particular bigram were not known.

We consider only up to 1 adjective since a hyperparameter search over up to 10 adjectives showed that 1-2 adjectives were always optimal; moreover, we only have 12 candidate adjectives to begin with, and manual inspection suggests that at most 1-2 of them ought to be relevant.

We consider 100 nearby nouns since we do not want to artificially constrain our model and prevent it from finding enough bigrams that it actually knows. Having separate steps for adjective/noun retrieval, assembling candidate bigrams, and then checking which bigrams are known is an artificial implementation choice that we make for our algorithm; humans could well be retrieving similar nouns and checking whether the resulting bigram is known in tandem. Thus, we always retrieve 100 nearby nouns "just in case" and instead rely on the

number of bigrams k to constrain the model. As discussed in Section 5.4.1, we set $k \leq 5$ to impose constraints akin to human working memory (Adam, Vogel, and Awh, 2017; Cowan, 2001). We allow the model to do a grid search over the exact value of $1 \leq k \leq 5$ by evaluating the model on the training set with memorization disabled. The optimal k typically ranges between 3-5 bigrams. In Table C.1, we also report the special configuration $k = 1$, where the model only considers the most similar bigram it can come up with. This mimics humans going with the "first bigram they can come up with", assuming that their retrieval process chooses a good candidate as its first choice.

The final configuration choice, which we did not discuss in Section 5.4.1, is the training data – what should be considered as bigrams that humans have previously encountered. Option 1 is to include all bigrams classed as "high frequency" by Chapter 4, i.e. all bigrams in the top quartile of their dataset. This results in sparse data for some adjectives. Notably, this only includes a single bigram involving the adjective *knockoff* and no bigrams including *unimportant*, meaning the model will be at a disadvantage for bigrams with these adjectives. In the N+A setting, it will have to rely primarily on bigrams involving e.g. *counterfeit*; in the noun only setting, it will often return no distribution. It is unclear whether this sparsity is precisely realistic, because these adjectives and their bigrams are low-frequency, or not. Options 2a and 2b are to train on the top x most frequent bigrams for each adjective, where we can consider (a) $x = 5$ (akin to the $k \leq 5$ setting for nearby bigrams), or (b) $x = 23$, which results in a nearly identical size training set (276 bigrams) to taking the top quartile (279 bigrams). We report all three settings in Table C.1.

Finally, in the case where no similar bigrams have known ratings, we opt to return a null distribution, which is always incorrect. We could alternatively return a fallback distribution which concentrates all its probability mass on "Unsure", but this will also be very unlike the human distributions under the Jensen-Shannon metric (which tend to have high SD when not concentrated at the ends of the scale), so this makes little difference. In practice, this only occurs in the "Noun only" setting for some bigrams involving *knockoff* and *unimportant* when

we use the top quartile of bigrams as the training set, since these adjectives have few or no high-frequency bigrams (1 for *knockoff*, 0 for *unimportant*).

C.1.2 Detailed results

Table C.1 shows the results for the analogy models built with GloVe embeddings, comparing the noun only setting with the N+A (noun + adjective) setting, and the single bigram setting ($k = 1$) with $k \leq 5$. We report the exact value for k chosen by the hyperparameter search. We also compare training on the top quartile of bigrams vs. training on the top 5 or 23 per adjective. Note that for the top 5 case, the set of novel bigrams (column 2, "Novel bigr.") is larger than in the other cases. We find that the simplest setting, analogy to a single noun (N only, $k = 1$) does not outperform a uniform distribution baseline overall. However, if we allow multiple adjectives, analogy to a single bigram ($k = 1$) is sometimes the best (selected even when we tune on $k \leq 5$). We also achieve similarly good results if we use nouns only but allow averaging over $k \leq 5$ bigrams. In the noun + adjective case, results are also similar whether we train on the top quartile of bigrams or the top 23 bigrams per adjective – training set size appears to be the driving factor, not how it is balanced. However, in the noun only case, which includes all the WordNet models, we unsurprisingly see a performance boost from including more bigrams for each adjective. (When training on the top quartile, the noun only setting necessarily fails for all bigrams involving *unimportant*, since there is no bigram with *unimportant* in the training data, and does poorly for *knockoff* as well, since there is only one bigram with *knockoff* in the training set.) Memorization of the training set boosts overall performance, as expected, though not so much when the training set is very small (top 5 bigrams per adjective).

Further, we observe that performance is generally lower on privative adjectives than overall, which makes sense because many bigrams with subsective adjectives have distributions almost entirely consolidated around "Definitely yes", and can be predicted from other bigrams.

Table C.1: *Average JS divergence (**best** / second) between various configurations of analogy models and human rating distributions, with & without training data memorization, for 'N only' vs. 'N+A' (1 nearby adjective) and $k = 1$ vs. $k \leq 5$ nearby bigrams (exact value of k tuned on training data). 'Novel B' = bigrams held out from each analogy model – for humans and LLMs, we can only be sure that zero-frequency bigrams ('Zero-freq. B') are novel. 'Priv. A' = bigrams with "privative" adjectives. * = set $k \leq 5$ but tuning chose $k = 1$. Llama 3 results and baseline from Chapter 4.*

	JS Divergence (lower is better)				
Model	Novel B	Zero-freq. B	Priv. A	Total	Total (+mem)
Human (resampled)	N/A	0.04	0.05	0.04	N/A
Human (analogy exp.)	N/A	0.14	0.21	0.16	N/A
Llama 3 70B Instruct	N/A	0.17	0.26	0.17	N/A
Uniform distr. baseline	N/A	0.33	0.20	0.34	N/A
Analogy models: GloVe					
N only, $k = 1$, top qt.	0.44	0.57	0.45	0.39	0.29
N only, $k = 1$, top 5/A	0.32	0.34	0.44	0.32	0.30
N only, $k = 5$, top qt.	0.41	0.55	0.39	0.36	0.27
N only, $k = 3$, top 5/A	0.28	0.28	0.36	0.28	0.25
N only, $k = 4$, top 23/A	0.26	0.25	0.33	0.26	0.17
N+A, $k = 1$, top qt.	0.29	0.31	0.39	0.29	0.19
N+A, $k = 4$, top qt.	0.26	0.26	0.34	0.26	0.17
N+A, $k = 3$, top 5/A	0.27	0.27	0.36	0.27	0.25
N+A, $k = 3$, top 23/A	**0.25**	0.25	**0.32**	0.26	0.17
Analogy models: WordNet					
N only, $k = 1$*, top qt.	0.41	0.54	0.36	0.36	0.26
N only, $k = 1$*, top 23/A	**0.25**	**0.24**	**0.32**	**0.25**	**0.16**
Analogy models: Llama 3 70B embeddings (final layer)					
N only, $k = 1$, top qt.	0.44	0.53	0.44	0.40	0.28
N only, $k = 4$, top qt.	0.40	0.50	0.37	0.35	0.26
N only, $k = 5$, top 23/A	0.26	0.26	0.34	0.26	0.17
N+A, $k = 1$, top qt.	0.33	0.33	0.44	0.34	0.22
N+A, $k = 4$, top qt.	0.28	0.27	0.35	0.28	0.18
N+A, $k = 5$, top 23/A	0.27	0.26	0.34	0.28	0.18
Analogy models: Llama 3 70B embeddings (initial layer)					
N+A, $k = 5$, top qt.	0.28	0.30	0.35	0.27	0.18

C.2 Estimate of analogy difficulty

C.2.1 Overview

For our analogical reasoning experiment, we attempt to estimate which bigrams might be difficult to find analogies for and balance evenly for this.We suppose that analogy could be difficult for bigrams with one or more following qualities:

- the noun has no high-frequency neighbors (below median among the nouns in the dataset)
- there are multiple convergent nearby bigrams with ratings that conflict
- there are non-convergent nearby bigrams (i.e. bigrams for which the conclusion is uncertain)

We use WordNet (Miller, 1995) rather than word embeddings to find neighboring nouns, since WordNet is manually annotated by human experts, and the British National Corpus for noun frequencies (Leech, Rayson, and Wilson, 2014). We manually define adjective similarity, since WordNet only provides a hierarchical taxonomy – and thus, a similarity metric – over nouns, described in Section C.2.3.

C.2.2 Results

In fact, we find that these criteria do not predict how often participants were able to come up with an analogy.

We fit a logistic mixed effects model in R (Bates et al., 2015) that predicts whether participants could find an analogy or not. As fixed effects, we include the three factors described in Section 5.5.1, as well as adjective class (typically privative or subjective) and specificity of the noun (measured by depth in the Wordnet taxonomy). We include adjective and participant as random effects. We also fit a second model where we replace specificity of noun with bigram frequency (the two are too correlated to include in the same model). In fact, we find that none of these factors are significant ($p < 0.05$) except for the presence of nearby divergent bigrams. This feature, however, only applies to 6 bigrams in the experiment, so

this may just be spurious. This non-significance may be the result of many false negatives in our labeling of these factors, since we can only test for nearby bigrams among the bigrams that Chapter 2 studied, not among the totality of nearby bigrams. It may also result from our participants construing analogy much more broadly than we did, as discussed in Section 5.5.2.

C.2.3 Adjective similarity details

We use the following (asymmetric) similarities, which are approximately scaled to match the Wu-Palmer similarity metric (which is 0.5 for siblings).

1. *artificial* → *fake*, *false*: 0.75
   → *counterfeit*, *knockoff*: 0.5
2. *counterfeit* → *knockoff*: 0.9
   → *fake*, *false*: 0.75
   → *artificial*: 0.5
3. *fake* → *artificial*, *counterfeit*, *false*, *knockoff*: 0.75
4. *false* → *fake*: 0.9
   → *counterfeit*, *knockoff*, *artificial*: 0.75
5. *knockoff* → *counterfeit*: 0.9
   → *fake*: 0.75
6. *former* → *artificial*, *counterfeit*, *fake*, *false*, *knockoff*: 0.5
7. *homemade* → *artificial*, *fake*, *false*: 0.8
   → *tiny*, *multicolored*: 0.75
   → *useful*, *illegal*, *unimportant*: 0.5
8. The remaining 5 subsective adjectives, *useful*, *tiny*, *illegal*, *unimportant* and *multicolored* are all assigned a similarity of 0.5 to each other and to *homemade*.

Note that we provide an unusually privative-looking set of similarities for *homemade* since the examples with *homemade* in the experiment are disproportionately chosen to be less subsective and thus challenging for analogy. Moreover, these similarities are adjusted for the fact that these are the only 12 adjectives available – of course they would be scaled differently if there were more options. We do not expect small changes to these similarities to have a noticeable difference on the selected bigrams.

C.3 Using human analogy bigrams in the analogy model

One bottleneck of our analogy model appears to be its lack of available bigrams with which to draw an analogy, i.e. which it has ratings for, compared to humans. We can try to ameliorate this by additionally giving it all the analogies found in the human analogy experiment, by assuming that the rating that they provide for the target bigram is the same as the rating they would assign to the analogical bigram. (This should be true if they are using the analogy as intended.) We filter the provided analogy phrases through WordNet to retain only two-word phrases whose first word is an adjective and the second a noun. This adds 340 bigrams involving 91 adjectives and 260 nouns. (The original dataset contained only 12 adjectives and 102 nouns.)

Unfortunately, we do not have full distributions for these bigrams; only 68 of the 340 bigrams so found have more than one rating, and only 11 have more than three. For target bigrams with privative adjectives, whose distributions are often spread out, analogy to these new bigrams will thus yield a high JS divergence simply because the distribution is too sparse. In line with this, the results in Table C.2 show that adding these additional bigrams worsens or does not improve the two best-performing GloVe models from Table C.1, though it does result in different hyperparameter choices during the grid search ($k \leq 5$).

To compensate for only having single ratings, we can instead evaluate the analogy models with the more lenient "accuracy within 1 SD of the human mean" metric proposed for single ratings in Chapter 4, which lets the model predict a mean rating instead of a full distribution. It is then judged "accurate" (enough) if this rating falls within 1 SD of the mean of the

Table C.2: *Average JS divergence (**best**) between analogy models and human rating distributions for the best GloVe models in Table C.1 and their counterparts trained on the additional bigrams from the human analogy experiment. This additional training data does not improve model performance as measured by JS divergence, because we do not have full distributions for many of the additional bigrams. 'Novel B' = bigrams held out from each analogy model – for humans and LLMs, we can only be sure that zero-frequency bigrams ('Zero-freq. B') are novel. 'Priv. A' = bigrams with "privative" adjectives.*

	JS Divergence (lower is better)				
Model	Novel B	Zero-freq. B	Priv. A	Total	Total (+mem)
N+A, $k = 4$, top qt.	0.26	0.26	0.34	**0.26**	**0.17**
N+A, $k = 4$, top qt. + exp.	0.45	0.62	0.41	0.39	0.29
N+A, $k = 3$, top 23/A	**0.25**	**0.25**	**0.32**	**0.26**	**0.17**
N+A, $k = 4$, top 23/A + exp.	0.26	0.26	0.33	**0.26**	**0.17**

human rating distribution that bigram (rounded to the nearest integer), incorrect otherwise. The problem with this metric, besides being ad-hoc, is that the simple "majority" baseline described in Chapter 4, which simply guesses "Unsure" for all bigrams with privative adjectives and "Definitely yes" for all those with subsective adjectives, achieves an accuracy of 0.89 using this metric. Bigrams with privative adjectives generally have such a high SD that this is a large and easy target to hit. Nonetheless, a random guessing baseline scores only 0.46 on this metric, so the metric is still somewhat informative.

If we add the new bigrams provided by the analogy prompting experiment to the training set and evaluate with this Within 1 SD metric, we do see a significant performance increase compared to using just the original training set, as shown in Table C.3. Note that optimizing over this metric yields new values for the parameter k, within the constraint $k \leq 5$. $k = 1$ is uniformly chosen during tuning even when we set $k \leq 5$. In contrast to the JS divergence, where we generally saw lower (better) values for subsective adjectives and higher (worse) values for privative ones, this metric yields the opposite, since the SDs for subsective-adjective bigrams are much smaller: we see lower (worse) accuracies for subsective adjectives.

This suggests that if we had full distributions for these bigrams, adding more training data might indeed significantly improve the model. What amount of training data is appropriate for modeling humans remains an open question.

Table C.3: *Results for the best GloVe models in Table C.1 and their counterparts trained on the additional bigrams from the human analogy experiment using the more lenient "accuracy within 1 SD of human mean" metric proposed in Chapter 4. All models use $k = 1$ even when tuned with $k \leq 5$; this makes sense as averaging is less likely to improve this metric. Unlike for the JS divergence shown in Table C.2, results do improve. However, results must be interpreted relative to the "majority" baseline provided in Chapter 4, which highlight the difficulty with this metric. 'Novel B' = bigrams held out from each analogy model – for humans and LLMs, we can only be sure that zero-frequency bigrams ('Zero-freq. B') are novel. 'Priv. A' = bigrams with "privative" adjectives.*

	Accuracy within 1 SD of human mean				
Model	Novel B	Zero-freq. B	Priv. A	Total	Total (+mem)
"Majority" baseline	N/A	0.91	0.78	0.89	N/A
Random guessing baseline	N/A	0.46	0.61	0.46	N/A
N+A, top qt.	0.71	0.77	**0.72**	0.69	0.78
N+A, top qt. + exp.	**0.76**	0.76	0.69	**0.74**	**0.81**
N+A, top 23/A	0.70	0.76	0.71	0.68	0.76
N+A, top 23/A + exp.	0.75	**0.79**	**0.72**	**0.74**	0.80