# Inferring Comparison Classes of Gradable Adjectives

## The Role of Informational Goals and Sentence Structure

### By

### Polina Tsvilodub

Thesis Supervisor:

Prof. Dr. Michael Franke, Institute of Cognitive Science, Osnabrück University

Thesis Supervisor:

Dr. Michael Henry Tessler, Postdoctoral Associate, Department of Brain and Cognitive Sciences, MIT

# Abstract

Understanding gradable adjectives like "big" requires making reference to a so-called comparison class - a set of objects the referent is implicitly compared to. For example, the utterance "That Great Dane is big" could mean "That Great Dane is big compared to dogs in general" or "That Great Dane is big compared to other Great Danes"; yet the comparison class is rarely stated explicitly. So how do listeners establish the comparison class, given multiple a priori reasonable options? Research on gradable adjectives has focused on the representation and integration of comparison classes into compositional semantics, but little is known about how human listeners decide upon a comparison class. This work takes a functional perspective on comparison class inference, guided by informational goals that speakers pursue when producing an utterance with a gradable adjective, and how listeners expect these goals to be achieved syntactically. For instance, given simple *"Subject Predicate"* sentences listeners expect that the subject aids reference (i.e., identifies the target), whereas the predicate accomplishes predication (i.e., asserts a property of the subject). Therefore, a noun appearing in the predicate is more likely to be intended to constrain the comparison class, whereas a noun in the subject can be explained away as intended for reference, leaving comparison class inference to other pragmatic reasoning. Converging evidence from four behavioural experiments supporting this proposal is presented alongside a novel formalisation of the inferential account in a qualitative computational model within the Rational Speech Act framework. This work contributes to the body of research on gradable adjectives, and provides a case study of context-dependent language, emphasizing the complexity of the relation between form and meaning of linguistic expressions.

# Acknowledgements

# Contents

# Chapter 1

# Introduction

The meaning of natural language expressions heavily depends on the context in which these expressions are used, but speakers rarely explicitly outline which aspects of the context are relevant for their interpretation.

This issue is clearly illustrated by utterances involving gradable adjectives like *big, small, tall, expensive* etc. These adjectives are typically taken to describe a *degree* to which an object possesses some property, e.g., the degree of bigness (i.e., size) for the adjective 'big', but specific degrees a speaker intends to convey vary a lot depending on the particular referent and context. Intuitively, the utterance "That's big!" denotes quite different size degrees, depending on whether it was uttered in reference to a flower or in reference to a house, while both objects could potentially co-occur in the same perceptual context; given this utterance, it is left to the listener to identify the correct referent and size degree. The aspect that goes unsaid and allows for this flexible use of the adjective *big* across referents and contexts is *what the intended referent is big relative to.* Humans easily infer that these two objects might be compared to different things: for instance, it is more likely that the flower is big for this specific kind of flowers or relative to other flowers around it, whereas the house is probably rather being compared to other houses in the neighborhood.

However, speakers rarely explicitly state this comparison class — the set of entities the target is compared against, and it is left to the addressee to establish the relevant comparison set (Solt, 2009). Listeners feature vast general knowledge and experience about the world helping them interpret context-sensitive language (Tessler et al., 2017), but what additional linguistic features do listeners attend to? In particular, how do listeners establish a comparison class in order to interpret a gradable adjective, given infinitely many a priori conceivable options for the comparison class?

This work investigates the role of syntactic structure for sentences containing relative gradable adjectives, suggesting that the syntax provides a cue to contextually

relevant aspects for adjective interpretation, which are integrated with other cues like perceptual context and world knowledge.[1] In particular, we hypothesize that syntactic structure reflects informational goals interlocutors strive to achieve; they reason about these goals pragmatically when inferring the comparison class of gradable adjectives. Focusing on the informational goals of *reference* and *predication*, this work presents a novel **reference-predication trade-off hypothesis** of comparison class inference, contributing to the body of research on gradable adjectives and providing a case study for the relationship between linguistic form and meaning. Evidence from four behavioral experiments is provided in support of this functional hypothesis, as well as a Bayesian model of gradable adjective interpretation, showing that sophisticated pragmatic reasoning about syntactic structure can be captured using the generic probabilistic Rational Speech Act framework (Goodman & Frank, 2016).

---

[1]This thesis summarizes and extends the work by Tessler, Tsvilodub, Snedeker and Levy published in Tessler, Tsvilodub, Snedeker, et al. (2020), that appeared in the *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society.*

# Chapter 2

# Understanding Gradable Adjectives

Gradable adjectives are a particularly interesting case study of context-sensitive language. That is, it depends on the context what exactly counts as *tall, expensive, small* or *full* — a one-meter tall three-year-old counts as tall, but a one-meter tall redwood tree does not; a three-quarter full cup of coffee counts as full, but a three-quarter full spaceship fuel tank does not. While both examples show context-sensitivity of the adjective's meaning, these two adjectives differ in what exactly about their meaning depends on the context: in case of *relative gradable adjectives* like 'tall' the context determines how much of the feature described by the adjective is required to count as 'tall', whereas in case of *absolute gradable adjectives* like 'full' the context determines how much the degree of the described feature may deviate from total fullness (Aparicio et al., 2016; Hofherr & Matushansky, 2010; Kennedy, 2007).

In particular, the meaning of a relative gradable adjective, for instance 'big', can be described in that 'big' refers to the size of an object, and the size of that object described as 'big' must be at least $X$, such that it counts as big, relative to some standard of comparison $\theta$. This means, relative gradable adjectives convey a feature, like size, and the degree $X$ to which the referent possesses this feature must exceed some threshold $\theta$ for the referent to be felicitously described by the respective gradable adjective (e.g., Kennedy, 2007). At the same time this threshold $\theta$ can vary across contexts or categories: the minimal size of a flower that counts as big is quite different from the minimal size of a house that counts as big. Moreover, this threshold can vary within categories: the minimal size of a big sunflower is different from the minimal size of a big daisy, although both belong to the category flowers. Hence, this threshold $\theta$ is strongly influenced by the set relative to which the object is compared — namely the *comparison class*.

In contrast, the meaning of an absolute gradable adjective, for instance 'full', is debated: some researchers argue that it refers to an endpoint on the feature scale described by the adjective, i.e., 'full' refers to the maximum on the scale of volume for the object under discussion, and differences between absolute and relative gradable adjectives arise from structural differences of the scales described by respective adjectives (Aparicio et al., 2016; Kennedy, 2007; Qing & Franke, 2014). Others argue that the meaning of absolute gradable adjectives is also resolved relative to a context-sensitive threshold $\theta$, by mechanisms universal for all gradable adjectives (Lassiter & Goodman, 2017).

Generally, gradable adjectives are *vague* — their meaning is subject to contextual variability, and to other characteristic features of vagueness: there exist so-called borderline cases, and these adjectives give rise to the Sorites paradox (Kennedy, 2007). Specifically, even when a comparison class is set, there are cases where it is unclear whether an object counts as e.g. 'expensive': while a cup of coffee for $1 is clearly cheap, and a cup for $5 is clearly expensive, it might be difficult to say whether a $3.75 coffee is expensive or not — this is a borderline case. Using the same example, the Sorites paradox can be illustrated for gradable adjectives as follows:

P1: A $5 cup of coffee is expensive (for a cup of coffee).

P2: Any cup of coffee that costs 1 cent less than an expensive one is expensive (for a cup of coffee).

C: Therefore, any free cup of coffee is expensive.

It is the vague nature of gradable adjectives that makes it difficult to pinpoint why exactly people accept the premises so easily, and although this argument seems valid, the conclusion is clearly false (see Kennedy, 2007, for more details).

Investigating these important properties in greater detail is outside of the scope of this work: in the remainder, the focus is to investigate the importance of comparison classes, specifically for relative gradable adjectives. Yet characteristics like borderline cases and eliciting the Sorites paradox emphasize that capturing the kind of implicit comparison to a threshold $\theta$ which occurs in the positive form of gradable adjectives, while accounting for the existence of these properties, is rather difficult. The following sections review state-of-art representations of relative gradable adjective semantics and the role of comparison classes therein. Then, prior related theoretical and experimental work on comprison classes is discussed.

## 2.1 Semantic Representation of Gradable Adjectives

Currently standard theories of gradable adjectives converge on representing gradable adjectives as a function mapping their argument — the referent — to a degree on an ordered scale representing some feature (e.g., 'big' and 'small' represent size), utilizing degree morphology (Kennedy, 2007). Degree morphology for the positive form of relative adjectives is informed by their comparative form, where the degree of a feature of the referent is explicitly compared to another degree of the same feature, and this comparison is overtly realised by a degree morpheme -er. For instance, in the comparative sentence "Bob is taller than Alice" Bob's height is explicitly compared to Alice's height, expressed by the morpheme -er appended to tall. By contrast, unmodified positive forms of relative adjectives which are the focus of this work don't have an overt degree morpheme signaling the comparison to some point of reference; in the currently widely accepted approach (reviewed by Kennedy, 2007) a phonologically silent null degree morpheme *pos* is introduced for this purpose. The morpheme *pos* takes the adjective as an argument and returns a standard of comparison — the context-dependent threshold $\theta$. In Kennedy (2007), the comparison class is assumed to be an argument of the adjective, potentially restricting the domain of entities it applies to — an assumption discussed further in Section 2.2. Formally, *pos* denotes the following:

$$\llbracket _{Deg} pos \lambda g \lambda x.g(x) \rrbracket = \lambda g.\lambda x : g(x) > s(\lambda x : g(x)) \tag{2.1}$$

In other words, the degree to which the referent $x$ possesses the property denoted by the adjective $g$ must exceed some threshold, provided by $s(g)$, where $s$ is "a context-sensitive function that chooses a standard of comparison in such a way as to ensure that the objects that the positive form [of the adjective] is true of 'stand out' in the context of utterance, relative to the kind of measurement that the adjective [i.e., $g$] encodes" (Kennedy, 2007, p. 17). The contextually relevant aspects providing the threshold can be summarised as the comparison class of the adjective. For example, the expression 'big dog' is true if the size of a target dog exceeds some size-threshold, set by the comparison class. Depending on the context and the comparison class this threshold might vary: the minimal size the dog has to have in order to be described as 'big' is different if the dog is a toy dog, and the comparison class are other toys, than for a dog that is a Great Dane and the comparison class is other Great Danes.

Alternative to the degree-semantics framework, delineation-based formalizations of gradable adjectives treat them as unary predicates, forming partial functions depending on contextually provided comparison classes (Klein, 1980). Such an ap-

proach removes degree representations from the semantics, although degrees arguably are an indispensable part of the meaning of gradable adjective (Solt, 2009).

The general issue of outlined semantic representations of gradable adjectives is that they assume the relevant comparison class to be supplied contextually, yet omitting to specify what exactly the comparison class is or how it is determined. While this work assumes a degree-based formalisation, it should be noted that alternative approaches also rely on the notion of contextually appropriate comparison classes, making the question addressed in this work as to how exactly comparison classes are determined a relevant one across different semantic representations.

## 2.2 Understanding Comparison Classes

Comparison classes can be understood as sets of entities, or reference frames the object described by the adjective is compared against (Bierwisch, 1989; Klein, 1980; Solt, 2009). In the examples outlined so far comparison classes were assumed to be sets of physical objects like dogs or flowers. But comparison classes need not be comprised of individuals or objects, they can also comprise events or locations: In the utterance "The store is crowded for a Tuesday" the fullness of a particular store is naturally compared to other Tuesdays, rather than to other stores (Solt, 2009). It is crucial that "the comparison class provides statistical information that serves to determine the thresholds [...., and] what is relevant is not only the central value but also some measure of the extend of dispersion of values corresponding to members of the comparison class" (Solt, 2009, p.193). Interestingly, the width of the value distribution might be closely related to the specificity of the comparison class: more general categories serving a comparison classes like *basic-level* categories tend to imply a wider distribution than more specific comparison classes, for instance based on *subordinate* categories (Rosch et al., 1976). From a pragmatic perspective, cooperative speakers should tend to use relatively specific comparison classes appropriate in context, since these are more informative with respect to the underspecified threshold $\theta$ than more general ones. Pragmatic listeners assuming cooperative speakers would then tend to infer maximally specific comparison classes, respectively (Tessler et al., 2017).

This naturally leads to the question of how exactly the standard of comparison — the threshold $\theta$ — is determined by a given comparison class. For instance, Cresswell (1976) suggested that the threshold $\theta$ is the average of the relevant feature over the comparison class, but arguments have been laid against this idea, showing that these thresholds do not seem to comprise a single point on the degree scale, but should rather be represented as a range of values (Kennedy, 2007; Stechow,

1984). One proposal by Solt (2009, p. 194) is that this range is computed as an interval around the median $median_{x \in C}$ provided by the comparison class $C$ (which the target referent $x$ is a member of), where the width of this interval is determined by the degree of variability of the feature in the comparison class, as provided by the measure function $MEAS$ and captured by the median absolute deviation ($MAD$):

$$R_{Std:C} = median_{x \in C} MEAS(x) \pm n \bullet MAD_{x \in C} MEAS(x) \qquad (2.2)$$

However, it is still unclear how the relevant comparison class $C$ is determined. Comparison classes can be expressed overtly using prepositional *for*-phrases, for instance, as in "That Great Dane is *big for a dog*" or in "That shirt is *big for you*". In the first example, additionally to expressing the comparison class, the *for*-phrase acts as a *presupposition trigger*, implying that the Great Dane is also a dog (cf. Bale, 2011; Solt, 2009). Notably, this is not the case for the second example.

There are several proposals with respect to compositional semantic integration of *for*-phrases. Kennedy (2007) suggested that *for*-phrases introduce a domain restriction on the gradable adjective via direct composition, hence being an argument of the adjective. That is, the comparison class restricts the domain of entities the adjective applies to. But this approach has difficulties accounting for cases when it is not the subject of the sentence that combines with the gradable adjective, or when adjectives appear in what has been labeled by Ebeling and Gelman (1994) as *functional uses*, e.g., "That short is big for you" (Solt, 2009).

An alternative is to interpret *for*-phrases in relation to the *pos*-morpheme, as marking its scope, similar to the relation between *than*-phrases and the comparative morpheme *-er*. In order to account for their presupposition-triggering behavior, the *pos*-morpheme is then assumed to take a comparison class $C$ as an argument, which by presupposition the referent is a member of (Solt, 2009). Formally:

$$\llbracket POS \rrbracket = \lambda C_{\langle et \rangle} \lambda P_{\langle d,et \rangle} \lambda x : x \in C. \forall d \in R_{Std:C}[P(x,d)] \qquad (2.3)$$

where $P(x,d)$ denotes the measure function mapping individuals onto respective degrees on the feature scale described by the adjective, and $R_{Std:C}$ is the standard of comparison, e.g., computed as described above. This view follows the proposal by Bartsch and Vennemann (1972), wherein the comparison class is an argument of a function computing the standard of comparison, whatever the nature of this function may be. In cases like "John is tall for a gymnast", overt *for*-phrases may directly provide this argument. However, in cases like "Sara reads difficult books for an 8-year-old" the *for*-phrase does not provide the comparison class directly: it is not "compared to other 8-year-olds", but "compared to other books for 8-year-olds",

which requires more complex compositional mechanisms (e.g., as suggetsed by Solt, 2009).

Finally, another approach to comparison class representation proposes that they "restrict binary relations, and these binary relations form the basis for the construction of [degree] scales [..., which] serve to relativize the calculation of a standard" of comparison (Bale, 2011, p. 170). This proposal is based on deriving scales described by gradable adjectives from quasi-orders, i.e., those binary relations, for instance by creating so-called equivalence classes (sets of objects with equivalent degrees on that scale), which then are ordered based on the original quasi-ordering, and finally by defining a measure function via mapping each element onto its equivalence class in the scale (Bale, 2011). Comparison classes then restrict the quasi-order before formation of the scale, limiting the quasi-orders to "ordered pairs consisting only of members of the comparison class", such that the scale only describes degrees of members of the comparison class (Bale, 2011, p. 178). This structure is then passed as an argument to some function returning the standard of comparison, analogous to approaches described above. One feature of this approach is the possibility to introduce a scale for gradable adjectives which are not inherently connected to some metric scale, e.g., for adjectives like 'beautiful' or 'interesting', because it considers the process of scale generation as the initial step of meaning computation (Bale, 2011).

In cases where no overt *for*-phrase is used, it is assumed that the argument of *pos* is a contextually appropriate implicit comparison class, where one potential option is that it is supplied syntactically by the nominal modified by the adjective. Many assume the modified noun to supply the comparison class universally (cf. e.g. Cresswell, 1976; Heim, 2000; Kamp, 1975), while Solt (2009) restricts this mechanism in terms of the *pos*-morpheme scope, proposing that comparison class saturation is local given a modified nominal, but involves raising in case of *for*-phrases. This leaves open the origin of comparison class arguments in sentences where the adjective appears predicatively without a *for*-phrase - a question focused on in sections 2.3, 2.4.

This work focuses on the determination of relevant comparison classes even before they are integrated compositionally, so no commitment to a specific compositional approach shall be made here.

Additionally to linguistic aspects, gradable adjectives and comparison classes, respectively, have also been addressed from a developmental and psychological perspective, in particular as a case study of children's developing understanding of context. Barner and Snedeker (2008) have shown that by the age of 4 years, children are able to track statistical regularities of a property described by an adjective (e.g., height described by 'tall') in a novel population of toys ('pimwits') and flexibly

adjust their use of the adjective according to changes of the property distribution.

Ebeling and Gelman (1994) distinguish three prominent uses of gradable adjectives children are exposed to, which can be loosely related to distinct linguistic constructions they tend to occur in, and how the comparison class may be supplied: namely, occurrences of adjectives where the comparison class is supplied *normatively, perceptually* or *functionally*. Normative comparison classes are based on a mental representation of the referent, for example it can comprise general world knowledge about the kind of things the referent belongs to. For example, a parent could describe a dog they see outside to their child by saying "That's a big dog!", taking advantage of the kid's knowledge about dogs, even when there are no other dogs around. One could hypothesize that here the relevant knowledge remains implicit and requires interlocutors to infer relevant cues from context, thus making these adjectives cognitively more challenging to interpret. Perceptual comparison classes are based on other objects of the same type as the referent physically co-present at the moment of utterance (Ebeling & Gelman, 1994). For instance, a parent could point at a particular dog when there are several around (e.g., in a park) and say to theit child "That one is big". The notion of perceptual comparison classes could naturally be extended to incorporate perceptually co-present objects of other kinds, in general. These comparison class uses might require less implicit general knowledge, but might still require figuring out which aspects of context are relevant. Finally, functional comparison class uses reference the intended use of the object, as in the aforementioned example "This shirt is *big for you*" (Ebeling & Gelman, 1994; Sera & Smith, 1987). While 'functional' comparison classes may be an exception in that they are very often stated overtly via the prepositional *for*-phrase, both normative and perceptual comparison classes often remain implicit, left to the listener to infer from their world knowledge or relevant contextual aspects. A preliminary study shows that adults might use syntactic structure of the utterance containing the adjective to help children establish the intended comparison class in such underspecified cases, consistent with the reference-predication trade-off hypothesis proposed in this work (Sinelnikova, 2020, discussed in greater detail in Chapter 6).

## 2.3 Syntactic and Semantic Aspects of Gradable Adjective Interpretation

While the notion of relative gradable adjectives as interpreted in reference to a comparison class has a long tradition (e.g., Bartsch & Vennemann, 1972; Bierwisch, 1989), there is little agreement on how exactly relevant comparison classes are identified when not supplied overtly. Prior work reviewed in this section has mainly

focused on how syntactic and semantic properties of adjectives determine them.

One line of work on how comparison classes might be determined approaches this question from a purely compositional perspective. In particular, the noun the adjective combines with is said to be at least a very salient cue towards the comparison class (Kamp, 1975). Simple compositional accounts propose that the nominal syntactically modified by the adjective necessarily stipulates the comparison class, such that 'small watch' resolves to 'the watch is small for a watch' (Cresswell, 1976; Kamp, 1975). More sophisticated ideas involve syntactic aspects of saturating the *pos*-morpheme (see Section 2.2). Yet, a lot of examples have been laid against such a simple mapping of the modified noun to the comparison class: intuitively, it is highly unlikely a priori that "John is a rich Fortune 500 CEO" means that he is *rich for a Fortune 500 CEO*, but more likely means that he is *a Fortune 500 CEO and therefore rich*; and "Kyle's car is an expensive BMW" doesn't mean that his car is *expensive relative to other BMWs* (Kennedy, 2007).

However, such syntactic theories focus on gradable adjectives occuring attributively, not accounting for their flexibility to occur both attributively and predicatively (for example, attributive: "That's a big dog"; or predicative: "That dog is big"; cf., Hofherr and Matushansky (2010), McNally and Kennedy (2008)). Furthermore, attributive adjectives can occur prenominally (e.g., 'visible stars') and post-nominally (e.g., 'stars visible [tonight]') (Hofherr & Matushansky, 2010). In English, the common basic position of attributive adjectives is prenominal, but post-nominal in e.g. Italian (Cinque, 2010); for this work focusing on English, post-nominal cases will be disregarded.

The exact relation between attributive and predicative occurences of adjectives is widely discussed; prior work attempted to derive one kind of syntactic construction from the other (e.g., Cresswell, 1976). For instance, predicative adjectives might be seen as elliptical uses derived from underlying attributive adjectives (e.g., "The dog is big" derived from "The dog is a big dog", cf., Kamp (1975)) or anaphoric constructions (e.g., "The dog is big" derived from "The dog is a big one"; however, the most reasonable resolution of the anaphora would stipulate referring to the subject noun 'dog', reducing this idea to the former one, cf., Goldberg and Michaelis (2017)). Alternative approaches suggested that the reversed might be true: attributive adjectives might be derived from predicative ones via a relative clause transformation ("I bought the table. The table was big. $\rightarrow$ I bought the table that was big. $\rightarrow$ I bought the table big. $\rightarrow$ I bought the big table."; cf. Bolinger, 1967, p. 2). However, one short-coming of such approaches is that they cannot account for why some adjectives appear only attributively (e.g., 'the main reason'), or only predicatively (e.g., 'asleep'), and these kinds of derivations have been argued to cause structural ambiguity rather than resolve it (Bolinger, 1967).

This implies the simplest generalisation of these compositional syntactic accounts to predicative adjectives: one could posit that the noun of the sentence generally sets the comparison class, such that the utterance "That Great Dane is big" would be taken to mean "That Great Dane is big for a Great Dane" (Tessler, Tsvilodub, Snedeker, et al., 2020). Yet, intuitive counter-examples might be put forward here: since Great Danes are generally big kinds of dogs, it seems perfectly reasonable to utter this sentence in context of other breeds of dogs and imply that the Great Dane is *big for a dog* (Tessler et al., 2017). Therefore, although the noun the adjective combines with is arguably a salient cue to the comparison class, the degree to which it restricts the comparison class might vary across different utterances and contexts.

Alternatively, one could imagine syntactic accounts of gradable adjective interpretation wherein the presence of syntactic modification would be the critical signal towards the role of the noun for comparison class restriction. Specifically, in presence of direct syntactic modification (i.e., as for prenominal attributive adjectives) the modified noun would set the comparison class akin to the simple syntactic account outlined above, while absence of modification (i.e., for predicative adjectives) would signal that the noun is *not* the comparison class. However, this alternative would not resolve remarks made against the compositional account regarding examples like "John is a rich Fortune 500 CEO" where the nominal is not necessarily the comparison class. Furthermore, it would remain unclear how comparison classes are determined in absence of modification by any compositional mechanisms different from what has been outlined above. The only viable alternative then seems to involve some kind of pragmatic reasoning (e.g., considering general world knowledge, Tessler et al. (2017)), at least for the predicative cases. Such pragmatic aspects are discussed in the next Section 2.4. Finally, Chapter 4 suggests experimental evidence, ruling out purely compositional accounts of comparison class determination.

From a semantic point of view, one property that is potentially relevant for comparison class determination is the difference between *intersective* and *non-intersective* (or *subsective*) adjective readings (Hofherr & Matushansky, 2010; Kennedy, 2012; Sedivy et al., 1999). A third kind — *non-subsective* adjectives like 'former' — will be disregarded for purposes of this work. Intersective adjective interpretations emerge when the referent is interpreted as a member of the intersection of two sets: the one denoted by the noun and the one denoted by the adjective (Kennedy, 2012). For example, the adjectival phrase of the sentence "Look at the red block" is interpreted as referring to a set of objects resulting from intersecting the set of red entities with the set of blocks — hence, resulting in an intersective reading. In contrast, subsective interpretations emerge when the referent is interpreted as a member of a subset of the set denoted by the noun, returned by the adjective combining with the noun: for example, the sentence "John is a skillful surgeon" implies that he is a

surgeon, but not necessarily that he is generally skillful — it only implies that he is *skillful as a surgeon* (Kennedy, 2012). Many vague gradable adjectives like 'big' and 'small' have been counted towards subsective adjectives, since their meaning does often depend on the noun they combine with (Sedivy et al., 1999). However, many examples show that meaning of such vague adjectives depend on more than just the head noun of the adjectival phrase: 'big snowman' clearly means different things in the sentences "My 2-year-old son built a really tall snowman yesterday" and "The D.U. fraternity brothers built a really tall snowman last weekend" (Sedivy et al., 1999, p. 115). These observations led to comparison-class degree-based approaches described in Section 2.1, and to ambiguity considerations between these two readings in the literature: it is argued that specifically prenomial attributive adjectives give rise to ambiguity between intersective and subsective readings (cf. 'Olga is a beautiful dancer', Hofherr & Matushansky, 2010). Yet it seems more plausible a priori to treat gradable adjectives occuring in either position (attributively or predicatively) as eliciting intersective interpretations, therefore leaving the comparison class underspecified. As described above, positing a subsective reading amounts to the simple syntactic hypothesis wherein the noun sets the compison class, which intuitively does not hold in general (especially given examples like "John is a rich Fortune 500 CEO": positing a subsective reading would translate to the sentence "John is rich *even for* a Fortune 500 CEO", which intuitively is very unlikely to be a priori true). However, positing intersective readings implies the existence of some abstract set of things that count as e.g. generally rich, which then intersects with the set denoted by the modified nominal (e.g., Fortune 500 CEOs) — a stipulation rather difficult to capture. At the same time, considering vague scalar adjectives as subsective seems to require direct modification because subsective readings are only assigned to attributive prenomial adjectives, which would require additional ad hoc mechanisms for interpreting the same adjectives occuring predicatively (cf. Hofherr & Matushansky, 2010). Hence, this distinction turns out to be difficult to apply to context-dependent adjectives.

To sum up, compositional syntactic accounts and semantic properties outlined above stipulate that the meaning of an utterance involving gradable adjectives is fully specified by its words: yet, it was shown that several other pragmatic components like context of the utterance and listeners' world knowledge have a large influence on the meaning of vague gradable adjectives (e.g., Kennedy, 2007; Sedivy et al., 1999; Tessler et al., 2017). Psycholinguistic studies investigating the role of these pragmatic factors for gradable adjectives and comparison class determination are reviewed in the next section.

## 2.4 Pragmatic Aspects of Gradable Adjective Interpretation

Being a prominent example for context-sensitive language, gradable adjectives have been used in many studies addressing various pragmatic and psycholinguistic phenomena. This section discusses some research on the role of visual context, world knowledge, typicality, subjectivity, overmodification and information packaging for adjective interpretation, as well as different prominent uses of adjectives discussed in the literature.

Several eye-tracking studies employing the visual world paradigm addressed the role of context for relative adjective interpretation (e.g., Aparicio et al., 2016; Sedivy et al., 1999). Eye-tracking studies mostly focus on *contrastive*, or *restrictive* uses of these adjectives — that is, helping to isolate an object denoted by the noun it combines with from the context. From a rather pragmatic perspective, contrastive use of adjectives is grounded in the assumption that nominal modifiers in general might convey contrastive information, because their presence is most naturally and rationally explained as *necessary* to contrast an intended referent denoted by the head noun from other members of the same category (Clifton Jr & Ferreira, 1989; Sedivy et al., 1999). Alternatively, contrastivity might be explained as triggered by definiteness of the noun phrase and the relation of the modifier to the discourse model (Sedivy et al., 1999; Steedman & Altmann, 1989). For gradable adjectives specifically, contrastivity might also arise from their inherent lexical properties, i.e., from the requirement for a comparison class which naturally implies a contrast (Bierwisch, 1989; Sedivy et al., 1999).

In their seminal work, Sedivy et al. (1999) looked at the effects of visual context and the head noun on the interpretation of prenominal adjectives. In particular, they hypothesized that local ambiguity of referring expressions involving adjectives is resolved incrementally, making use of context to interpret the meaning of vague utterances, additionally to the head noun. In the first experiment, participants heard instructions of the form "Touch the ADJ N", where the adjective ADJ could encode the shape, color or material of an object described by the noun N, presented in a visual context. The visual contexts were manipulated such that the referrring expression could either be disambiguated upon hearing the modifier, i.e., there was only one out of four objects to which the adjective applied (early disambiguating condition); or, such that there were two different objects with the critical property, and the noun disambiguated the utterance (late disambiguating condition). They found that participants were faster to respond to target objects in the early disambiguating condition compared to the late disambiguating condition, confirming

effects of incremental processing of the utterance. Additionally, they used a condition manipulation wherein the modifiers were either focused intonationally or not, where the utterance referred to either an object sharing the category or the critical property with a previously highlighted object, finding that participants were faster to identify targets when the modifier was used contrastively (i.e., the target shared the category with the previously mentioned referent), but the intonational contrast did not play a role. The authors concluded that participants initially expect a contrastive interpretation of such adjectives. In further experiments, relative gradable adjectives were employed; participants saw contexts displaying either a contrastive condition (two out of four objects of the same category differing with respect to the scalar property) or a non-contrastive condition (only one object belonging to the critical category), hearing instructions like in the first experiment; both conditions included a competitor object of a different category which could be felicitously described by the adjective. Furthermore, the typicality of the target object as described by the modified nominal was manipulated. Shorter reaction times were found in the contrastive condition, and overall for typical targets. Sedivy et al. (1999) concluded that for vague adjectives as well, participants used contextual information along with contrastivity expectations to process the utterance incrementally, even before the onset of the head noun. Finally, they hypothesized that contrastive interpretations might be correlated with the presupposition of existence and accesibility of the target object, as elicited by the definiteness of the noun and the overall task setup.[2] In another experiment where questions involving an indefinite noun were used instead of instructions, no effects of presupposition on contrastive interpretation effects were found. Overall, Sedivy et al. (1999) found that adjectives elicit strong expectations of contrastive meaning with respect to visual context when uttered in ambiguos referential expressions, indicating the importance of perceptual context as a cue to the comparison class for relative gradable adjectives. Furthermore, typicality effects indicated that participants make use of their stored representations — i.e., *world knowledge* — when interpreting relative gradable adjectives, implying that, contrary to simple compositional accounts, the modified noun is not the only cue to the comparison class. These typicality effects are also in line with findings from other studies investigating the propensity of interlocutors to use a modifier, or, to infer a target referent over a competitor (Bergey et al., 2020; Kreiss & Degen, 2020).

Aparicio et al. (2016) conducted a visual world study similar to the one by Sedivy

---

[2]Although it is generally accepted that definite descriptions like definite nouns are referential, there are also exceptions to this tendency like idioms (Reboul, 2001). Furthermore, e.g., Donnellan (1966) distinguishes between attributive and referential uses of definite descriptions, where only the latter is actually referring.

et al. (1999), investigating the effects of context on reference resolution for expressions with absolute and relative gradable adjectives. The study design corresponded to the design used by Sedivy et al. (1999), but employed geometric shapes instead of real-world objects, and used color, absolute and relative gradable adjectives. The critical utterances were of the form "Click on the ADJ N", containing a definite noun N and a prenominal adjective ADJ. They found that the target was identified faster in the contrast condition, and more so for relative than absolute gradable adjectives — the onset of eye-movements was observed before presentation of the head noun for color and relative adjectives, but not for absolute ones. Aparicio et al. (2016) concluded that the effect for relative adjectives is mostly driven by a presence of a perceptual comparison class in the contrast condition, while pragmatically imprecise interpretation of absolute adjectives involved higher processing costs and hence longer reaction times, touching upon an important distinction between vagueness and imprecision (cf. Kennedy, 2007).

To sum up, results of both studies are consistent with the hypothesis that context provides a salient cue to the comparison class that is integrated with the category provided by the modified noun, because critical adjectival utterances were interpreted faster when the visual context supplied a more homogenous comparison class and was more consistent with the head noun.

However, the assumption that adjectives are expected to convey contrastive information might be challenged by the observation that speakers also *overmodify* their referential utterances, i.e., use modifiers even when they are not necessary for reference resolution (Degen et al., 2020). Furthermore, contrastive interpretations imply that the modified nominal is used for *reference*, which, as discussed before, might be partly attributed to the definiteness of the modified noun. But while reference is undoubtedly an important primary communicative goal, there clearly are cases where a noun, e.g., combined with an adjective, is used for other goals, like *predication* (i.e., communicating a property of a referent already established in discourse). These non-referential uses of the adjective might be signalled linguistically for instance via an indefinite noun or a predicative adjective position.[3] This important distinction is discussed in detail in Chapter 3.

Finally, a study by Tessler et al. (2017) addressed empirically and computationally the important role of world knowledge for comparison class inference. The authors showed that listeners flexibly adjust comparison classes of gradable adjectives based on their world knowledge, when encountering simple utterances like "He's tall" said of targets about which listeners typically have strong expectations regarding the

---

[3]This might be a general expectation interlocutors have for utterances like "The dog is big"; however, this approximation is not intended as a strict rule — referential uses involving predicative adjectives can be easily imagined, e.g., in utterance like "The dog that is big just barked"

feature degree described by the adjective. Specifically, they showed that listeners are more likely to infer that an utterance like "He's tall" said of a basketball player means "He's tall for a person" (i.e., tall relative to a general, basic-level category), whereas the utterance "He's short" said of a basketball player rather means "He's short for a basketball player" (i.e., relative to the target's subordinate category). This pattern was clearly shown for targets of those categories which exhibit a rather high or low degree of the feature (e.g., basketball players, whose height is generally quite large; and jockeys, whose height is generally rather low). That is, based on their *prior world knowledge about likely feature degrees* of different categories, participants flexibly shifted the standard of comparison and pragmatically inferred the more likely comparison class of a predicative relative adjective. The studies presented in this work build on this experimental paradigm, making use of listeners' knowledge about features of categories saliently exhibiting high or low degrees of that feature, like basketball players or jockeys regarding height. When adjectives consistent with general expectations are attributed of those categories, the basic-level comparison class is a priori more likely to provide a felicitous expression than the subordinate comparison class, allowing to tease apart effects of world knowledge and linguistic cues towards the comparison class. However, the study by Tessler et al. (2017) only considered simple utterances, appearing without much context or a noun. Experiments presented in Chapter 4 extend this paradigm to accomodate a more realistic set-up.

To sum up, this chapter reviewed relevant theoretical and experimental work on representation and interpretation of gradable adjectives. It was shown that several aspects like the noun the adjective combines with, the perceptual and discourse context of the utterance, as well as other syntactic and semantic features contribute to establishing the correct comparison class and ultimately interpreting relative adjectives. Yet, up to date there are few attempts to unify these information sources in a comprehensive theory of gradable adjective interpretation.

# Chapter 3

# A Functional Perspective on Comparison Class Inference

This section aims to integrate both the role of the noun in the sentence as well as the role of pragmatic cues like perceptual context and world knowledge for relative adjective interpretation, presenting the **reference-predication trade-off hypothesis** of comparison class inference.

Specifically, the issue of comparison class determination is approached from a functional perspective, based on the question what *informational goals* speakers might pursue when producing an utterance containing an adjective, and how these goals might influence listeners' comparison class inferences (Tessler, Tsvilodub, Snedeker, et al., 2020). The proposed approach is an inferential account of comparison class determination, informed by the idea of recursive social reasoning mechanisms, applied to rational language use in a Gricean tradition: Speakers have certain informational goals which guide how they craft their utterance in order to facilitate message interpretation with respect to these particular goals for a listener (Goodman & Frank, 2016). Listeners, in turn, infer the most likely state of the world — that is, in case of gradable adjectives, the most likely comparison class — in light of those speaker goals.

In particular, in contrast to cases considered in eye-tracking studies described in Chapter 2, when using adjectives speakers might also primarily intend to convey a property of a target referent. In order to communicate that property of a referent, speakers must achieve at least two informational goals: *reference* — identifying the right target — and *predication* — attributing a property of the target, which in case of relative gradable adjectives amounts to communicating the specific degree of the feature denoted by the adjective (Kennedy, 2007; Reboul, 2001). For these two informational goals, it is reasonable to posit that listeners generally expect the subject to be sufficient in order to establish reference — independent of the

predicate asserted to hold of the subject (Reboul, 2001; Searle, 1969; Syrett et al., 2010). Cooperative speakers then aim to satisfy this general expectation.

This tendency is particularly strong for sentences with subjects containing referential expressions like definite descriptions, pronouns or deictics (cf. Section 2.4). Furthermore, it might be based on general information structural reasons: In order to predicate a property of a target, this target must be clear (Krifka, 2008; Searle, 1969). Therefore, the subject also tends to convey the *topic* of an utterance — that is, "the entity [...] under which the information from the comment constituent should be stored" (Krifka, 2008, p. 265); while the predicate tends to convey the *comment*, i.e., potentially new information about that entity (Chafe, 1976; Krifka, 2008; Reboul, 2001). A further heuristic distinction associated with the subject-predicate contrast comes from linguistic packaging literature, wherein the predicate is assumed to convey the *main news* (as opposed to *secondary information*), and also potentially *new information*, while the subject might convey secondary information which is already *known* (Kaiser & Wang, 2020).

Note, however, that there are exceptions to many of these tendencies: for instance, for the sentence "The boss fired the worker because he was a convinced communist" the pronominal *he* can be resolved not only after applying the predicate, but also only taking into account the context — *he* can either refer to the boss or to the worker (Reboul, 2001). Krifka (2008) also points out that the topic, and hence the subject, doesn't necessarily convey known information. Yet we posit that these structural expectations are a general enough heuristic holding in many contexts.

These expectations have implications for comparison classes of gradable adjectives insofar as speakers have the liberty to choose from truth-conditionally similar sentence options to communicate the same message. For example, in order to tell a friend on the phone about a huge dog that the speaker saw today, they have the liberty to say "That was big!", "That Great Dane was big!" or "That was a big dog!", among many other options. Consequently, the choice of a particular sentence over other equivalent options might respond to particular informational-communicative needs.

From this perspective, the influence of the noun on the comparison class in a simple *Subject Predicate* sentence depends on its position in the sentence. If the noun appears in the predicate of the sentence (e.g., in "That's a big Great Dane"), it can naturally be explained as produced by a speaker intending to constrain the comparison class, by packaging the noun along with the adjective as the most important information. By contrast, if the noun appears in the subject of the sentence (e.g., in "That Great Dane is big"), it can potentially be *explained away* as produced by a speaker who intends it to support reference (especially via combining it with
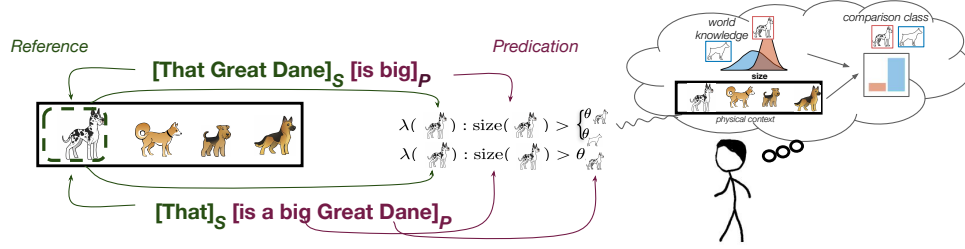
Figure 3.1: Cartoon of the inferential account for comparison class determination. The noun (Great Dane) in a sentence can be employed either for the goal of reference (green) or predication (purple), shown in the case when this distinction is made via the syntactic position of the noun (subject S vs. predicate P). When the noun is used for reference (top), a listener is left with uncertainty about what to use as the comparison class (dogs or Great Danes) and integrates their world knowledge and the physical context to make this inference. When the noun is used for predication (bottom), the listener should have less uncertainty about the comparison class: The comparison class is stipulated by the noun.

the deictic 'that'), and hence the noun is a weaker cue towards the comparison class. The comparison class inference is then guided by other pragmatic cues like world knowledge or perceptual context (Fig. 3.1).

Hence, the utility of the noun as constraining the comparison class is the result of a trade-off between its utility in reference and predication, such that comparison class inference is guided by integrating syntactic with other contextual cues.

## 3.1 Understanding Reference and Predication

This reference-predication trade-off hypothesis focuses on two basic informational goals, reference and predication, which have been discussed in a great deal of work in semantics, pragmatics and philosophy of language (Reboul, 2001; Zalta, 2019).

Searle (1969) conceptualizes both reference and predication as particular kinds of propositional acts, defining conditions to be fulfilled in order to accomplish them. Of particular importance for accomplishing reference is that the expression intended for reference isolates the target referent for the listener (Searle, 1969). Studies have shown that speakers are aware of this requirement, and being sensitive to contextual variability, adjust the informativity of their referential expression correspondingly, such that this requirement is satisfied (e.g., Graf et al., 2016). In particular, definite descriptions which prenominal adjectives might be a part of have been the focus of a

lot of work on reference, converging on the claim that a singular determiner phrase of the form *the $\phi$* triggers two presuppositions: the *existence* presupposition (i.e., that there is an object satisfying the description $\phi$), and the *uniqueness* presupposition (i.e., that such an object is uniquely identifiable) (Syrett et al., 2010; Zalta, 2019). These same presuppositions generally also hold for pronouns and demonstratives, but do not for indefinite descriptions of the form *a $\phi$* (Reboul, 2001; Zalta, 2017). Therefore, our experimental operationalization focusing on predication employs gradable adjectives in indefinite descriptions (s. Section 3.2)

The goal of predication builds upon reference, in that one of the requirements for accomplishing predication is that the same sentence contains a reference to the intended target of predication (Reboul, 2001; Searle, 1969). Specifically for relative adjectives, predication is tantamount to communicating a particular property degree, and therefore supplying a felicitous comparison class, for the referent under discussion. Accomplishment of the goal of predication is often roughly equated with the syntactic predicate, which notably might consist of a bare predicative gradable adjective, introduced with a copula. Therefore, one might hypothesize that the noun cannot be the only cue to the comparison class, since predication might be accomplished by a bare adjective.

This review does not attempt to resolve the debate on how exaclty reference and predication might be accomplished. But of particular importance for this work is the flexibility of nouns with respect to both informational goals: combining with the deictic 'that', the noun can accomplish reference; but being part of a non-referential expression (e.g., an indefinite description), the noun can also contribute to predication (Reboul, 2001).

The focus of this work are these two relatively basic informational goals, but clearly there are other communicative uses of adjectives. For example, Barker (2002) distinguishes between *descriptive* and *meta-linguistic* uses of vague adjectives. The former refers to what so far has been considered *predication* applied to relative adjectives, while the latter refers to giving "guidance concerning what the prevailing relevant standard" of comparison is for the adjective under discussion (Barker, 2002, p. 2). That is, the goal in this case is to teach the appropriate use of the vague adjective, given a particular property value in context. Another related goal of adjective use might be conveying a subjective opinion about a property (Kaiser & Wang, 2020). Interestingly, gradable adjectives have been shown to differ in the degree of subjective content they might convey (Scontras et al., 2017). Further investigation of these communicative goals and their relation to reference and predication is left open to future research.

The discussed properties of reference and predication lead to the particular experimental operationalisation of the reference-predication trade-off hypothesis, de-

scribed in the next section.

## 3.2   Experimental Operationalization

In present studies, the flexibility of nouns to contribute to either informational goal leads to the operationalization of the reference-predication trade-off hypothesis via a syntactic manipulation, wherein the noun (N) which combines with the gradable adjective (ADJ) appears either in the subject or in the predicate of a sentence. Experiments 1-3 employ sentences including only one critical noun N (Tessler, Tsvilodub, Snedeker, et al., 2020):

*Subject N*: That N is ADJ.
*Predicate N*: That's a ADJ N.

Experiment 4 focuses on the critical noun N1 syntactically modified by the adjective, which then appears either in the subject or in the predicate of an utterance (Tessler, Tsvilodub, & Levy, 2020):

*Subject N*: That ADJ N1 is a N2.
*Predicate N*: That N2 is a ADJ N1.

Given the referential presupposition of the deictic 'that', subject nouns should be taken as establishing reference. For the predicate noun condition, reference should be taken as being established by the bare deictic or the second noun N2, respectively. Given the presuppositional nature of definite descriptions, the predicate N conditions were chosen to include an indefinite description, such that the predicate may apply to several members of the context and referential pressure be shifted to the subject of the utterance. Furthermore, in the experimental set-up the referent described by critical sentences was perceptually salient, and the task did not involve direct reference resolution, such that referential pressure was generally lower than in eye-tracking experiments described in Section 2.4.

Referential utility is operationalized not only through the syntactic position of the noun, but also via the category of the noun: both *basic-level* and *subordinate* referent labels were used (e.g., a Great Dane might be described by the subordinate noun 'Great Dane' or by the basic-level noun 'dog'). While referential utility depends on the context, more specific nouns — i.e., subordinate labels — have generally higher referential utility than more general ones (Graf et al., 2016).

The critical question addressed by this manipulation is how speakers and listeners treat these syntactic frames, asserting the ADJ of referents for whom they are felicitous given one comparison class, but not another (e.g., a *normal-sized* Great

Dane can felicitously be described as 'big' given the comparison class 'dogs', but not 'Great Danes').

The reference-predication trade-off hypothesis predicts that nouns that are more likely to establish reference are less likely to constrain the comparison class. Therefore, when the noun appears in the subject of the utterance, it can be explained away as establishing reference, and hence is a weaker cue towards the comparison class, leaving it open to influences of world knowledge and perceptual context.

Conversely, when the noun is taken to contribute to predication, i.e., when it appears in the predicate of the sentence, it is more likely to constrain the comparison class. Therefore, listeners would a priori rather expect this noun to be consistent with the comparison class felicitous in order to describe the target referent: for instance, the basic-level category label would be more appropriate for setting the comparison class when describing a normal-sized Great Dane as 'big' than the subordinate category label. That is, the utterance "That's a big dog" would be more appropriate than "That's a big Great Dane" in order to describe a normal-sized Great Dane, because the subordinate category *Great Danes* is generally a large-subordinate category compared to the basic-level category *dogs*, but normal-sized representatives are not necessarily large compared to their subordinate category.

Note that although the differences in comparison class restriction are approached through the lense of this syntactic manipulation, the underlying communicative goals are the primary driving force in comparison class inference, to which the syntax is just a cue. There might well be other syntactic realisations of these informational goals (Reboul, 2001): The sentence "What is big is that Great Dane" seems appropriate in a context where generally big things are discussed; in this utterance reference is accomplished from the predicate, and because of this referential pressure, under the trade-off hypothesis the noun would not be expected to constrain the comparison class, although it appears in the predicate, supporting the view that the syntactic position of the noun is dissociable from the intended communicative goals.

To show that informational goals are primary for comparison class inference as opposed to specific syntactic properties of the adjectival phrase, Experiment 4 focuses on manipulating the informational goal the noun is a cue to, while it is directly syntactically modified by the adjective. This manipulation allows to disentangle the effect of the noun position from the effect of syntactic modification of the noun, which are confounded in experiments 1–3. For example, critical sentences in Experiment 4 are "That big Great Dane is a prize-winner" (subject-N) or "That prize-winner is a big Great Dane" (predicate-N). The trade-off hypothesis predicts that even directly modified nouns in the subject position contribute to reference, and thus should be less likely to constrain the comparison class, compared to nouns

appearing in the predicate.

The next chapter presents results of these four behavioural experiments exploring the reference-predication trade-off hypothesis, specifically investigating the use of the size adjectives 'big' and 'small'. These two adjectives are chosen for practical reasons: size is a visually accessible feature, allowing for easy presentation and manipulation of the context in web-based experiments. Furthermore, humans usually have strong expectations about typical size distributions of different natural categories, from which the target referents were sampled for the experiments. Three distinct dependent measures were used to assess the influence of various cues on comparison class inference.

# Chapter 4

# Experiments

The reference-predication trade-off hypothesis was tested in four preregistered behavioural web-based experiments employing different dependent measures (Table 4.1). The crucial manipulation in all experiments was the varying position of the critical noun — it appeared either in the subject (e.g., "That N is ADJ" or "That ADJ N1 is N2") or in the predicate ("That's a ADJ N" or "That N2 is a ADJ N1") of the sentences presented in the experiments. These sentences described a depicted object which appeared in visual context.

These objects were sampled from five different *basic-level* categories: dogs, birds, flowers, trees and fish (Rosch et al., 1976). Within each basic-level category, at least two *subordinate* categories were chosen which exhibit a rather high or rather low amount of the feature described by the gradable adjectives under investigation — that is, those subordinate categories which people expect to be rather large or rather small representatives of their basic-level categories (s. Table 4.2). For example, for the *dog*-category, the large-subordinate category *Great Danes* and the small-subordinate category *pugs* were chosen. As shown by Tessler et al. (2017), when encountering representatives of such categories described by the adjective consistent with participants' prior expectations about the degree of the feature-under-discussion, people are a priori more likely to infer the basic-level comparison class than the subordinate comparison class. For example, when encountering the sentence "It's big" said of a Great Dane (a large-subordinate category for the basic-level category dogs), humans are more likely to infer that the Great Dane is big relative to other dogs in general, than big relative to other Great Danes. Following the design of Tessler et al. (2017) in these experiments allows to test the effect of syntactic position of the noun on how strong the noun is taken to constrain the comparison class: The reference-predication trade-off hypothesis predicts that nouns in the predicate position constrain the comparison class more strongly than in the subject position, such that a priori using the basic-level noun in predicate position

Table 4.1: Overview of experiments 1-4.

| | E1: Syntax Rating | E2: Noun Production | E3: Comparison Class Inference | E4: Direct Modification |
|---|---|---|---|---|
| Research Question | Given two sentences with Ns in different positions, do participants prefer one syntactic frame over the other, depending on the N? | Do speakers produce different Ns given different syntactic frames? | Do listeners infer different comparison classes given different syntactic frames, Ns and contexts? | Is there an effect of syntax on directly modified nouns? |
| Participants | 80 | 190 | 200 | 36 |
| Task | Slider rating | Fill-in noun | Paraphrase comparison class | Paraphrase comparison class |

is more felicitous in order to describe a normal-sized large-subordinate object (e.g., a Great Dane) than using a subordinate-label of the object in predicate position. Both nouns would be felicitous in the subject position. Furthermore, encountering a subordinate label in the predicate position, should signal a more extreme feature value than the basic-level label. Therefore, in all experiments, the referents were described by the adjective matching prior feature-degree expectations; for instance, Great Danes and sunflowers were always described as *big*, and pugs or daisies as *small*.

The structure of all experiments was similar. First, participants completed a bot-check trial (Fig. 4.1): Participants read a sentence where a named speaker asked a named listener: "It's a beautiful day, isn't it?". The speaker and listener names were sampled from lists of ten popular male and female English names (s. Appendix A). For example, the sentences read: "James says to Linda: 'It's a beautiful day, isn't?'; Who is James talking to?". Participants were asked to fill-in in lowercase who the listener is talking to. Participants were provided feedback and had maximally three attempts to fill-in the correct name. They were only allowed to proceed, if they successfully completed the bot check. Then, participants read instructions and completed practice trials, before completing main trials. After the main trials, they completed a socio-demographic post-test questionnaire, where they were asked to indicate their native language and optionally provide further information. For all experiments, participants were recruited via the crowd-sourcing platform Amazon's Mechanical Turk; only participants with IP addresses in the United States and work approval rating of at least 95% were permitted to participate. Participants were restrained from taking part in multiple experiments of this series.

The first experiment (E1, Sentence Rating Experiment) was a sentence rating experiment, wherein participants had to rate two sentences which differed in the

Table 4.2: Experimental items: each basic-level context had two potential targets from an either saliently small or saliently big subordinate category within the basic-level class. Items marked with * were used only in Expt. 2, items marked with + were used in all experiments including Expt. 4.

| Basic-level category | Smaller referent | Bigger referent |
| --- | --- | --- |
| Dogs[+] | Pug[+] | Great Dane[+] |
| Dogs[+] | Chihuahua[+] | Doberman[+] |
| Birds[+] | Hummingbird[+] | Eagle[+] |
| Fish | Goldfish | Swordfish |
| Flowers[+] | Dandelion[+] | Sunflower[+] |
| Trees[+] | Bonsai[+] | Redwood[+] |
| Birds* | Sparrow* | Goose* |
| Birds* | Canary* | Swan* |
| Fish* | Clownfish* | Tuna* |
| Flowers* | Daisy* | Peony* |

position of the noun (subject-N vs. predicate N) and the specificity of the noun (basic-level vs. subordinate label), as describing an object in context. In the second experiment (E2, Noun Production Experiment), participants had to fill-in the missing noun of a sentence describing the size of a referent in context. The position of the missing noun was varied. In the third experiment (E3, Comparison Class Inference Experiment), participants provided the inferred comparison classes via a free-production paraphrase, given sentences which varied by the noun category and its position, as describing a referent in different contexts. Finally, the fourth experiment (E4, Direct Modification Experiment) gathered inferred comparison classes in a paradigm akin to E3, but from sentences wherein the critical subordinate noun appearing in subject or predicate position was always syntactically modified by the adjective. All experimental materials and data can be found under `https://github.com/polina-tsvilodub/refpred`. All experiments were realized using the _magpie-framework (Ilieva et al., 2018). All experiments and preregistrations can be viewed under `tinyurl.com/yb5ogj5g`.

## 4.1   Experiment 1: Sentence Rating Experiment

The aim of the sentence rating experiment was to investigate whether participants prefer one syntactic frame over the other, given two truth-conditionally equivalent sentences with different noun positions, depending on the noun category. The type of the noun and its syntactic position differed within-subjects.

First, participants completed two warm-up trials to familiarize themselves with

# Are you a bot?

James says to Linds: It's a beautiful day, isn't it?

Who is James talking to?

Please enter your answer in lower case.

LET'S GO!

Figure 4.1: Example view of the bot check trial: The speaker James addresses the listener Linda.

the slider rating procedure (Fig. 4.2). On one trial, participants read: "Imagine you see this basketball" above a picture of an orange basketball, and read below the question: "How well does each of the sentences describe it? (Please click on the slider to provide a rating)". Two sentences appeared below: "The basketball is orange" and "The basketball is green", to be rated on sliders ranging from "very bad" to "very well". In the background, the ratings were mapped onto a scale ranging from 0 to 100. The slider was light gray, with a round handle appearing upon clicking on the slider track. The same sliders were used in the main trials. On the other warm-up trial, participants read: "Imagine you see this chair" above a picture of a purple chair. The sentences to be rated appearing below were: "The chair is yellow", and "The chair is blue". The order of the warm-up trials was randomized.

Then, participants completed six main trials (Fig. 4.3). Participants read "You and your friend see the following:" above a basic-level context picture (e.g., a group of flowers). In all experiments, the basic-level context pictures consisted of six members of the same basic-level category as the referent of the trial, including two other members of the same subordinate category as the referent, and four other objects. The six members consisted of two members of a large-subordinate, a medium-sized subordinate, and a small-subordinate category within the basic-level category each (e.g., the flower-context consisted of two sunflowers, two roses and two dandelions; s. Fig. 4.3). The context was used to set the overall reference comparison class for the targets. It also set the visual reference frame. Below, they read the sentence "You also see this SUB_N", where SUB_N was the subordinate label of the target referent, which appeared depicted below, such that participants knew the subordinate category of the referent. The pictures depicted referents a little smaller than

Imagine you see this basketball.



How well does each of the sentences describe it? (Click on the
slider to provide a rating)

The basketball is orange.   very bad  [=============●====] very well

The basketball is green.    very bad  [●=================] very well

NEXT

Figure 4.2: Example view of the sentence rating warm-up trial wherein participants rated sentences about the depicted basketball.

members of the same subordinate category in the context, such that the felicitous comparison class was pushed towards the basic-level category of the target. Below, the question about the critical sentences appeared: "How well does each of the sentences describe it? (Click on the slider to provide a rating)". Then, the two critical sentences appeared left of the sliders one below the other. The sliders ranged from "very bad" to "very well". On every trial, in one of the sentences the noun appeared in the subject (e.g. "That N is {big/small}"), in the other in predicate position ("That's a {big/small} N"). The order in which these syntactic conditions appeared was randomized between-subjects. On half of the trials, the noun was the basic-level target label (e.g., dog); on the other half it was the subordinate target label (e.g., Great Danes), balanced within-subjects. Participants saw each of the six possible contexts once, and for each context, one of the two possible targets (large-subordinate vs. small-subordinate category representative) was sampled, balanced within-participants (Table 4.2).

The reference-predication trade-off hypothesis predicts that the effect of syntax on the rating will be more pronounced for sentences containing a subordinate noun than for sentences constaining a basic-level noun because a subordinate noun in predicate position would communicate an infelicitous comparison class for a normal-sized referent, while a basic-level predicate would be felicitous. That is, sentences with a basic-level noun in the predicate position are expected to receive a higher rating than sentences with a subordinate noun in the predicate, but a smaller difference in the ratings is expected for sentences with a noun in the subject position because either noun type can be felicitously used to pick out the referent. Given the specifi-

You and your friend see the following:



You also see this sunflower.



How well does each sentence describe it? (Click on the slider
to provide a rating)

That sunflower is big.  very bad ———————— very well

That's a big sunflower. very bad ———————— very well

Figure 4.3: Example view of a sentence rating main trial: The critical noun is a subordinate target label of a large-subordinate category, appearing in the subject or predicate of the sentence.

cation of the statistical model (s. Section 4.1.2), this prediction would be evidenced by a a negative credible estimate of the noun $\times$ syntax interaction.

## 4.1.1  Participants

113 participants were recruited and 33 were excluded for indicating a native language other than English, failing the practice trials or providing the same responses on every trial (see Appendix A). The experiment took about 5 minutes and participants were compensated $0.80. If partial data was missing from a participant, available data was used for analyses.

## 4.1.2  Results

For all reported experiments, maximal random effects structure licensed by the design was used (Barr et al., 2013). All statistical analyses were performed using the language R, in particular using the package `brms` for computing Bayesian regression models (Bürkner, 2017; Team et al., 2013).

A Bayesian linear mixed-effects regression model was fit for this experiment, predicting the sentence rating from the syntactic condition of the sentence (subject vs. predicate N), the noun type (basic-level vs. subordinate target label), their interaction and by-participant and by-target random intercepts and random slope
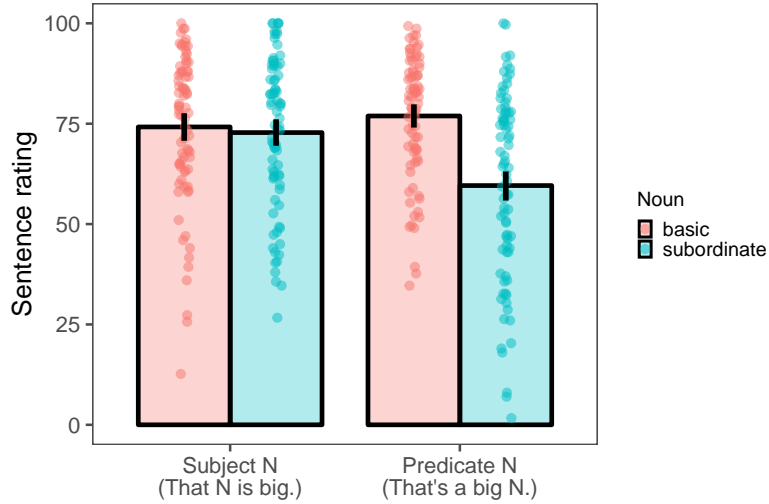
Figure 4.4: Experiment 1 results: Mean ratings for how well sentences which differed in the syntactic position of the noun (x-axis) and the noun-label (color) described a typically-sized referent (e.g., a Great Dane) in basic-level context. Points represent participant means within condition. Error-bars denote bootstrapped 95% confidence intervals (bootstrapping independent of random-effects structure).

effects of syntax, noun type and their interaction.[4] Both predictors were sum-coded, coding both the subject-noun and the basic-level noun conditions as 1 and the other levels as -1, respectively. Default priors were used. An exploratory model including a main effect of syntactic condition order was also fit, revealing no effect of syntactic condition order, so the data was collapsed across the two conditions for further analyses.

Consistent with predictions, participants substantially dispreferred sentences with a subordinate noun in the predicate compared to the subordinate position, but no effect of syntax was found for the basic-level nouns, as indicated by the syntax × noun-type interaction ($\beta = -4.01[-5.84, -2.18]$) (Fig. 4.4).[5] Additionally, an overall preference for basic-level nouns ($\beta = 5.44[2.76, 8.09]$) and the subject-noun syntactic structure ($\beta = 2.69[0.69, 4.77]$) was found. Furthermore, a relatively high by-target variance revealed that some items received overall lower ratings, possibly due to differing namability or typicality of the items (by-target intercept: $\beta = 9.53[5.76, 15.73]$). A relatively high by-participant variation indicated differences in overall rating preferences (by-subject intercept: $\beta = 11.96[9.80, 14.52]$). Finally, an exploratory analysis including a target size predictor (small-subordinate vs. large-subordinate category) did not reveal any size-effects on the rating.

To sum up, the sentence rating experiment showed that participants are sensitive

---

[4]Model in `brm`-style syntax: `rating ~ syntax * NP + (1 + syntax*NP | subject) + (1 + syntax*NP | target)`

[5]All results report the mean and 95-% Bayesian credible interval

**Warm-up trials**

Please label the pictures below.



This is a [____] This is a [____]

These are both [____]

Figure 4.5: Example view of the noun production warm-up trial: Participants have to label a large-subordinate (Great Dane, right) and a small-subordinate target (pug, left) for the dogs-category, and provide the common label below.

to the position and the type of the noun, dispreferring sentences where a noun that provided an a priori less felicitous comparison class appeared predicatively.

## 4.2   Experiment 2: Noun Production Experiment

The goal of the noun production experiment was to investigate whether participants produce nouns of different categories in a free-production setting, given different syntactic frames. The noun slot of the critical sentences in the main trials appeared either in the subject position (i.e., in "That __ is {big/small}") or in the predicate position (i.e., in "That's a {big/small} __"), manipulated between-subjects.

Participants completed two experimental blocks, each consisting of three warm-up trials and three main trials. In the warm-up trials participants familiarized themselves with the subordinate categories used in the main trials. They saw pictures of a member from a large-subordinate and a small-subordinate category within one of the basic-level categories used in the main trials within the same block (e.g., a Great Dane and a pug) (Fig. 4.5). Participants were prompted to provide labels for these pictures. Below they were prompted to provide a common label for both pictures (i.e., dogs), so that they were 'warmed-up' to provide lables of different categories. They were provided feedback for the labels and could proceed upon adjusting their labels to correct responses. The number of attempts participants needed until they filled-in the correct labels was recorded. In this experiment, four additional subordinate categories were used, which can be found in Table 4.2 marked with *. For each participant, six out of ten possible contexts were sampled. Three of these contexts

You see the following:



You also see this one:


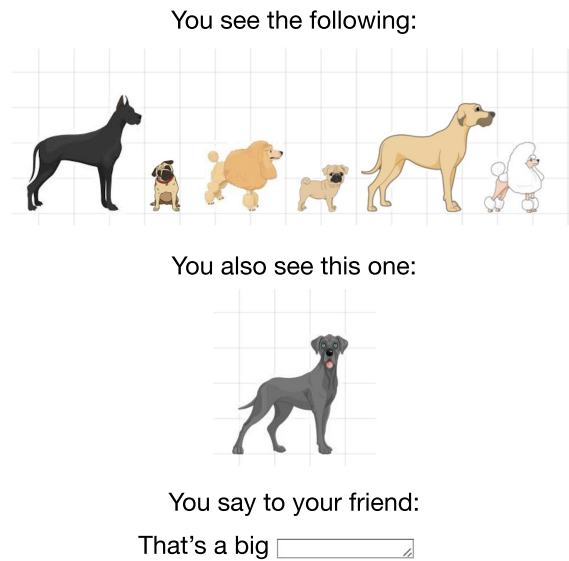
You say to your friend:

That's a big [                ]

Figure 4.6: Example view of the noun production main trial: Participants fill-in the noun in the predicate position of a sentence describing a large-subordinate target in basic-level dogs context.

and their corresponding targets appeared in the first experimental block, and the other three in the second. The trial order within the warm-up block and the main block was randomized.

On the main trials, participants read: "You see the following:" above a basic-level context picture, akin to the contexts used in Experiment 1. Below, they read "You also see this one:" and saw a picture of the target referent. Then they read: "You say to your friend:", prompting them to fill-in the missing noun in the sentence: for the subject-noun condition, the template was "That__ is {big/small}", for the predicate-noun condition, the template to be completed was "That's a {big/small} __ " (Fig. 4.6). The size of the target referent was balanced within-participants: on three trials, participants saw referents from a small-subordinate category, and on three, they saw referents from a large-subordinate category. For each context, participants saw only one of the possible targets (e.g., the large or the small subordinate target).

The reference-predication hypothesis predicts that speakers sensitive to listeners' expectations about accomplishment of communicative goals should be more likely to produce basic-level target labels than subordinate target labels in the predicate compared to the subject position. A positive credible regression coefficient for the effect of syntax would confirm this prediction, indicating a higher proportion of basic-level responses in the predicate compared to the subject position.
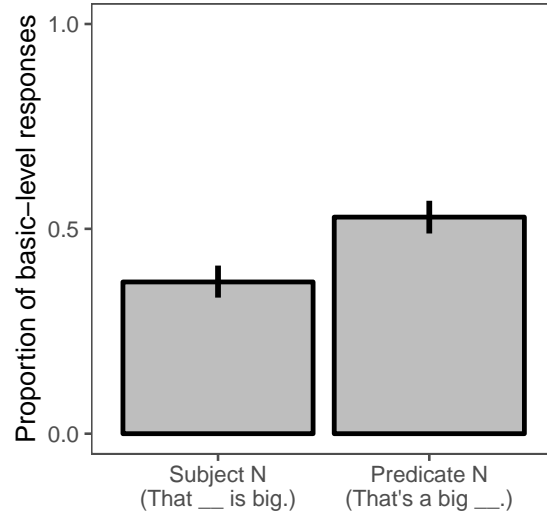
Figure 4.7: Experiment 2 results: Proportions of freely-produced basic-level labels (e.g., *dog*) in different syntactic frames (x-axis) when the referent was a typically-sized member of a subordinate category (e.g., a normal-sized Great Dane). Error-bars denote 95% bootstrapped confidence intervals.

### 4.2.1 Participants

242 participants were recruited, and 52 were excluded for indicating a native language other than English or for failing the warm-up trials. The exclusion criterion was taking more than four attempts on any warm-up trial to provide the expected answer upon correction. The experiment took about 7 minutes and participants were compensated $1.00. If partial data was missing from a participant, available data was used for analyses.

### 4.2.2 Results

The responses provided by participants were categorized manually into basic-level or subordinate-level labels of the targets, disregarding the noun number and spelling mistakes. 5 responses were superordinate referent labels (i.e., more general labels like 'animals') and were collapsed with basic-level labels. 16 (1.4%) uncategorizable responses were excluded from analysis. A logistic generalized mixed-effects Bayesian regression model was fit, regressing the response category (basic-level vs. subordinate target label) against the syntax of the sentence (subject-N vs. predicate-N), random by-participant and by-referent intercepts and random by-referent slope effects of syntax.[6] Default priors were used. The predictor was sum-coded, coding predicate-N syntax as 0.5 and subject-N syntax as -0.5.

---

[6]In brm-style syntax: `response_category ∼ syntax + (1 | subject) + (1 + syntax | target)`

Consistent with predictions, a strong effect of syntactic position of the noun was found, indicating that participants were more likely to use basic-level labels in the predicate position ($\beta = 2.25[0.74, 4.01]$; Fig. 4.7). That is, participants were more likely to provide the noun matching the felicitous comparison class in the predicate position, but more likely to use the noun with higher referential utility in the subject. As expected, an exploratory model including a main effect of referent size (large-subordinate vs. small-subordinate category) did not reveal any differences between target types. By-target variation revealed that participants were generally more likely to produce subordinate labels for some targets than for others (by-target intercept: $\beta = 1.16[0.72, 1.80]$). For example, participants were very likely to produce the subordinate label for the swan-item, possibly due to namability effects.

The noun production experiment showed that speakers are sensitive to the syntactic structure of the sentence and flexibly adjust their noun choices in order to communicate a felicitous comparison class, when presented with a free-production task.

## 4.3 Experiment 3: Comparison Class Inference Experiment

The two previous experiments support the reference-predication trade-off view, by showing that participants disprefer sentences like "That's a big Great Dane" in order to describe a normal-sized Great Dane, but accept either target label in the sentence subject. The goal of this comparison class inference experiment was to measure comparison class inferences more directly, presenting participants with sentences they had to paraphrase. The types of inferred comparison classes were investigated, as influenced by the position of the critical noun in the sentence (subject vs. predicate), the type of noun (basic-level vs. subordinate vs. 'one') and the visual context of the sentence (basic-level vs. subordinate context). All three factors were manipulated within-subjects.

In this experiment, participants first completed a comparison class paraphrase practice trial, akin to the paradigm employed in the main trials. Participants were told that on the main trials they will see a sentence containing a word that is relative, and their task will be to figure out what this word is relative to. They read an example task: "Speaker A: 'The Empire State building is tall.' What do you think speaker A meant?". Below they saw a paraphrase template where they provided the inferred comparison class of the adjective *tall*: "The Empire State building is tall relative to other＿" (blank to be completed with the inferred comparison class). Participants were provided feedback on their response and had to correct it to one of

You and your friend see the following:

Your friend runs far ahead of you, and you see him in the distance:

Your friend says:
**That's a big Great Dane.**

What do you think your friend meant?

It's big relative to other [            ]

Figure 4.8: Example view of a comparison class inference main trial: Participants paraphrased the critical utterance with a subordinate noun in predicate position, which appeared in basic-level context, describing a large-subordinate target.

the possible options among {buildings, skyscrapers, houses, constructions}. Then, participants completed two blocks consisting of labeling warm-up trials and main paraphrase trials. Three of the six basic-level categories used in this experiment were sampled for the first block, with the respective subordinate category members appearing in the warm-up trials, the other three categories appeared in the second block (Table 4.2). These labeling warm-up trials are of the same kind as in Experiment 2 (Fig. 4.5).



Figure 4.9: Example view of a subordinate context: For the category Great Danes, the context depicts six different Great Danes.

For the main trials there were basic-level and subordinate-level contexts for each possible referent. Basic-level contexts were identical to the contexts of respective categories in Experiment 1 and Experiment 2 (Figs. 4.6, 4.3); the subordinate contexts consisted of six other representatives of the same subordinate category as the target referent. For example, the subordinate context for a Great Dane consisted

of a picture of a group of six other Great Danes (Fig. 4.9). Within each main trial block, there were six trials, wherein for each of the three sampled categories, one possible referent appeared in the corresponding basic-level context (e.g., for the category dogs, the Great Dane appeared in basic-level dog context), and the other possible referent appeared in the corresponding subordinate context (i.e., then the pug appeared in subordinate pugs-context). The referent was described by a critical sentence in which the noun could appear in the subject or in the predicate of the sentence. The noun could be either the basic-level (e.g., dog) or the subordinate label of the referent (e.g., Great Dane). Furthermore, a baseline condition with an anaphoric 'one' in the noun position was included, in order to measure the baseline influence of the visual context on comparison class inference: the anaphora is most likely to be resolved contextually, meaning "dog" in the basic-level context and "Great Dane" in subordinate context (Goldberg & Michaelis, 2017). Crossing the visual context (basic vs. subordinate), the syntax (subject-N vs. predicate-N) and the possible nouns (basic vs. subordinate vs. 'one') results in a 2x2x3 design, yielding 12 unique conditions.[7] Each participant saw a total of 12 main trials.

On main trials, participants read "You and your friend see the following:" above a context picture (Fig. 4.8). Below, they read: "Your friend runs far ahead of you, and you see him in the distance:". The illusion of distance was created contextually in order to disguise the perceptual size of the target referent and push participants towards inferring the size of the referent from the sentence, rather than perceptually, while conveying its subordinate category. This illusion was supported by the picture appearing below, wherein the small target referent was depicted next to a small person (as compared to the context, i.e., appearing in distance). Below, participants read: "Your friend says:", followed by the critical sentence. Participants were asked "What do you think your friend meant?", followed by the paraphrase template "It is {big/small} relative to other __", blank to be completed with the inferred comparison class. The order of context, noun and syntax conditions was randomized for each participant.

When participants don't have access to visually assessing the size of a referent and need to infer the comparison class from the sentence, they might be more sensitive to linguistic cues like the sentence structure. According to the reference-predication hypothesis, they would be more likely to take the noun as a cue to the comparison class when the noun appears in the predicate of that sentence, than when it appears in the subject. When the noun appears in the subject, comparison class

---

[7]Due to my coding mistake, the conditions were balanced at the level of individual factors. That is, each participant saw six trials in basic-level and six trials in subordinate context, six trials in the subject and six in the predicate condition, as well as four trials in each noun-condition. However, participants potentially did not see all 12 possible combination of these factors.

inference can be driven by other pragmatic inferences, e.g., by world knowledge and visual context. If this is true, more basic-level comparison classes should be inferred from sentences appearing in basic-level context, compared to subordinate context. A credible positive regression coefficient for the effect of context would support this prediction. Additionally, inferences drawn from the anaphoric 'one' which is the baseline condition for the effects of visual context are then expected to mirror this difference, such that more basic-level comparison class should be inferred in the basic-level context than in the subordinate context, evidenced by a credible positive estimate for the respective contrast (s. Section 4.3.2; cf., Goldberg and Michaelis (2017)).

In particular, nouns with a high referential utility (e.g., subordinate target labels given a basic-level set of distractors, cf., Graf et al. (2016)) should highlight the main reference-predication prediction, by providing strong referential cues in the subject, but potentially signalling a comparison class different from prior world knowledge and perceptual context in the predicate. That is, participants are expected to infer more subordinate comparison classes (i.e., less basic-level ones) from predicate subordinate nouns than from subject subordinate nouns. In contrast, a smaller difference in comparison classes inferred from basic-level nouns in different positions is expected, since this noun in the predicate would signal the basic-level comparison class, and pragmatic inferences driving comparison class inference when this noun appers in the subject would also suggest the basic-level comparison class (e.g., given world knowledge and especially the basic-level context). For this prediction to be supported statistically, a credible positive syntax $\times$ noun interaction regression estimate is expected.

On the contrary, if comparison class inference was completely driven by the noun of the sentence or other purely syntactic or semantic properties discussed in Chapter 2, no interpretative differences should be observed when the same sentences occur in different perceptual contexts. In this case, no credible effect of context will be observed. Taking an opposite point of view, another conceivable mechanism for gradable adjective interpretation wherein the perceptual context only supplies the comparison class would predict identical inferences drawn from sentences involving different nouns or different syntactic structure. In this case, no credible effects of noun type or syntax should be observed.

### 4.3.1   Participants

245 participants were recruited and 45 were excluded for indicating a native language other than English, or failing either the comparison class inference practice trial or the labeling warm-up trials more than four times upon correction. The experiment
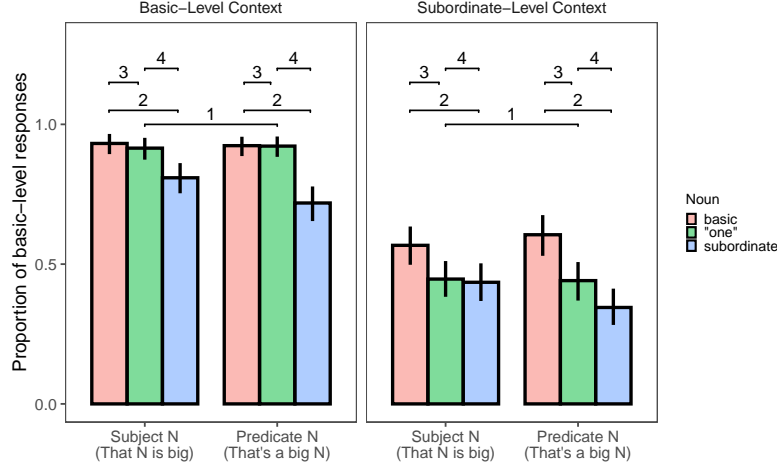
Figure 4.10: Experiment 3 results: Proportions of inferred comparison classes in terms of basic-level responses (e.g., "...big relative to other dogs"), depending on syntactic position of the noun (x-axis), noun-label (color), and context (facets). Context strongly modulated the comparison class (left vs. right panel). The noun additionally provided a cue to the comparison class (red vs. blue) bars, regardless of syntactic position. The effect of noun (red vs. blue) is modulated by syntax. Error-bars denote bootstrapped 95% confidence intervals.

took about 9 minutes and participants were compensated $1.20. If partial data was missing from a participant, available data was used for analyses.

## 4.3.2 Results

Participants' responses were manually classified into basic-level and subordinate comparison classes. 4 superordinate comparison classes were collapsed with the basic-level responses. 39 (1.6%) uncategorizable responses were excluded from the analysis. A Bayesian logistic mixed-effects regression model was used, regressing the response category against the syntactic condition (subject-N vs. predicate-N), the noun category (basic vs. subordinate vs. 'one'), the context (basic vs. subordinate), their two-way and three-way interactions and maximal random effect structure appropriate for this experimental design.[8] Default priors were used. The predictors were sum-coded: predicate-N and basic-level context were coded as 1, subject-N and subordinate context as -1; the basic-level and the subordinate noun-levels were coded against the baseline anaphoric 'one'.

The results indicate that participants flexibly adjust the inferred comparison class according to many factors. First, a large effect of visual context going above

---

[8]In brm-style syntax: `response_category ~ syntax*NP*context + (1 + syntax + NP + context || subject) + (1 + syntax*NP*context || target)`. Correlations of random effects were set to 0 for computational tractability.

and beyond other factors was found ($\beta = 1.88[1.49, 2.31]$; Fig. 4.10, left vs. right facets), supported by the inferences drawn from the baseline condition anaphoric 'one' ($\beta = 0.37[0.10, 0.64]$; Fig. 4.10, comparison 1, left vs. right facet): participants were appreciably more likely to infer basic-level comparison classes given a basic-level context, compared to a subordinate context. Furthermore, an effect of the noun on inferred comparison classes regardless of its position in the sentence was found: participants were more likely to infer basic-level comparison classes from basic-level nouns than from subordinate nouns ($\beta = 2.01[1.37, 2.71]$; Fig. 4.10, comparison 2). The noun-effects can be observed on top of effects of the visual context indicated by the baseline condition 'one': participants inferred more basic-level comparison classes from basic-level nouns than from 'one', and less from subordinate nouns than from 'one' (basic-level vs. 'one': $\beta = 0.60[-0.47, 1.70]$, Fig. 4.10 comparison 3; and subordinate vs. 'one': $\beta = -1.40[-2.17, -0.66]$, Fig. 4.10 comparison 4). Notably, the subordinate comparison class was the minority response even given a subordinate noun in the basic-level context, speaking against a compositional view of adjective comparison classes, wherein the noun would always set the comparison class (Fig. 4.10; blue bars, left facet). Even in the subordinate context inferences drawn from the subordinate noun were not exclusively subordinate comparison classes, indicating a basic-level bias (Fig. 4.10; blue bars, right facet; cf. Graf et al., 2016; Rosch et al., 1976).

Crucially, a credible syntax-by-noun interaction was found, supporting predictions provided by the reference-predication trade-off hypothesis: more subordinate comparison classes were inferred from subordinate nouns appearing in predicate position than in the subject position, compared to basic-level nouns (Fig. 4.10 comparison 2 × x-axis; $\beta = 0.47[0.02, 0.95]$). Examining the interaction by-noun, preliminary evidence was found that the interaction is primarily driven by the subordinate noun: a 90.0% probability was found that the subordinate-N vs. 'one' × syntax interaction term was less than 0 (i.e., more subordinate comparison classes were inferred when the noun was in the predicate; $\beta = -0.36[-0.93, 0.22]$), in contrast to only a 65.5% probability of the basic-N vs. 'one' × syntax interaction being greater than 0 (i.e., more basic-level comparison classes inferred when the noun was in the predicate; $\beta = 0.11[-0.44, 0.68]$). This effect was even more pronounced under an exploratory model, assuming only a two-way syntax-by-noun interaction and a main effect of context: the probability was 95.6% for the subordinate noun vs. 'one' × syntax estimate to be less than 0.[9] This is consistent with the reference-predication trade-off hypothesis, predicting that an effect of syntax is more pronounced for nouns with higher referential utility, which is generally the case for subordinate compared

---

[9]Exploratory model: `response_category ~ syntax*NP + context + (1 + syntax + NP + context || subject) + (1 + syntax*NP + context || target)`

to basic-level nouns, especially in the basic-level context (cf. Graf et al., 2016).

Taking the reference-predication trade-off hypothesis even further, another exploratory analysis suggested that participants might consider informational functions of the noun irrespective of the syntactic position. In particular, when a noun in the subject is referentially uninformative they might reason that this noun too is intended for predication. For instance, the subordinate context yields the basic-level target label referentially uninformative (all members of context can be described by the basic-level noun), such that listeners might reason about the presence of the noun as intended to convey the comparison class when it appears in both subject or predicate position. In line with this idea, in the subordinate context condition a higher rate of basic-level comparison classes was inferred from the basic-level noun compared to 'one' (91.6% of the credible interval of the basic vs. 'one' estimate was greater than 0: $\beta = 0.99[-0.71, 2.59]$; Fig. 4.10; comparison 3, right facet), indicating an effect of the basic-level noun going beyond contextual resolution of this referentially underspecified expression, as in case of 'one'.[10] The data observed in the basic-level context is also consistent with the hypothesis that the referentially uninformative noun in the subject signals the comparison class, but the referentially-uninformative basic-level label condition and the baseline 'one' are both subject to a ceiling effect and hence leave no room for any effects beyond baseline.

These empirical results provide a comprehensive picture of syntactic and pragmatic effects contributing to comparison class inference. In particular, this experiment shows that the same utterances are interpreted differently in distinct contexts, as evidenced by the large effect of context. This speaks against purely syntactic or semantic views arguing that meaning of utterances is fully specified by their words (as discussed in Section 2.3). Furthermore, the influence of the noun in the utterance independent of context and syntactic position confirms that the noun is a salient cue to the comparison class, yet insufficient on its own to account for interpretative differences observed. Finally, evidence consistent with the reference-predication hypothesis is found, suggesting that humans integrate the referential utility with contextual cues when reasoning about the contribution of the noun to the comparison class (as shown by the noun $\times$ syntax interaction). The data suggests that one-dimensional theories of gradable adjective interpretation do not account for flexible comparison class inference; humans integrate both pragmatic and syntactic information to felicitously use gradable adjectives across different contexts.

---

[10]These contrasts were computed on data subsetted by context

## 4.4 Experiment 4: Direct Modification Experiment

However, in order to keep a simple operationalization of the reference-predication distinction, a potential confound was introduced in experiments 1-3. The position of the noun was perfectly confounded with whether the noun was directly syntactically modified by the adjective (predicate-N condition) or not (subject-N condition). Therefore, experiments 1-3 did not allow to rule out that the observed interpretative differences were not due to the differing modification, as discussed in Section 2.3. Yet the reference-predication trade-off view predicts that referential pressure takes off predicational weight from the noun used for reference and therefore decreases its strength in constraining the comparison class, independent of the syntactic modification because informational goals are suggested to be the primary driving force above syntactic phenomena. This prediction was investigated in this direct-modification experiment (Tessler, Tsvilodub, & Levy, 2020).

That is, the main idea of this study was to manipulate the position of the noun while the modification was constant across the positions. Therefore, in critical trials the position of the critical noun in the sentence was varied and the noun was always directly modified by the adjective *big* or *small* (i.e., 'big Great Dane', 'small pug'). The critical nouns were always subordinate referent labels. In order to create maximally symmetric syntactic manipulations of the critical sentences, a second noun was used which described a visually salient feature of the referent. For example, the referents for one of the dog contexts were prize-winners, as indicated by prize-bows depicted on the referents (Table 4.3)). So the critical sentence was either "That prize-winner is a big Great Dane" (predicate-N) or "That big Great Dane is a prize-winner" (subject-N). For the same reason nouns were chosen to describe this feature, as opposed to e.g. adjectives ("That big Great Dane is my favourite" vs. ?¿'That my favourite is a big Great Dane"). The features were chosen such that they were visually accessible, describable by a noun and would be part of the referent. The referents appeared in a basic-level context, which included two other members of the same subordinate category as the referent, and two other individuals with the feature described by the second noun of the sentence: e.g., in the dog-context there were two other prize-winners (Fig. 4.12). Because the reference-predication trade-off is based on explaining away a noun via its potential referential use, through this contextual manipulation the referential utilities of the two nouns of the sentence were equal, such that only the noun's syntactic position and combination with the deictic 'that' could provide a cue towards referential intention. Therefore, the critical subordinate noun is expected to constrain the inferred comparison class more strongly when it

Table 4.3: E4 experimental items: each basic-level context had two potential targets from an either saliently small or saliently big subordinate category within the basic-level class. Each category had a corresponding context cover story which was comleted by "...and you see the following:". The referents had an additional visually salient feature, described by the second noun in critical sentences (N2).

| Basic-level category | Smaller referent | Bigger referent | Context | Visual feature / N2 |
|---|---|---|---|---|
| Dogs | Pug | Great Dane | You and your friend are at a pet show. | prize-winner |
| Dogs | Chihuahua | Doberman | You and your friend are at an animal training ground. | service-animal |
| Birds | Hummingbird | Eagle | You visit your friend who works at an animal shelter. | rescue |
| Flowers | Dandelion | Sunflower | You and your friend are at their garden. | gift |
| Trees | Bonsai | Redwood | You and your friend walk to their cabin in a park for the first time. You want to memorize the path. | landmark |

appears in the predicate of the sentence rather than in the subject.

The experimental set-up was similar to Experiment 3. Five different contexts were used in this experiment: there were two dog contexts, a flower, a bird and a tree context (Table 4.3). Four out of five contexts were randomly sampled for each participant. Participants completed two experimental blocks, each consisting of warm-up and main trials using two of the sampled categories. In the first block, participants first completed three rounds of labeling warm-up trials. A round consisted of a demonstration trial where participants saw two subordinate members of a basic-level category used in this block and read their labels. For example, they saw pictures of a Great Dane and a pug next to each other and read "This is a Great Dane" and "This is a pug", respectively. They could proceed after 3.5 seconds to the next trial where they had to label other instances of the same categories themselves. They also had to provide a common label for the pictures (i.e., dogs; see Fig. 4.5). The order of the pictures was randomized between-participants. They were provided feedback on their labels and could proceed only after correcting their labels. After two labeling warm-up rounds, participants completed two demonstration trials of at least 3.5 seconds each, learning about the additional features of the referents described by the second noun of the critical sentences in main trials (Table 4.3). For example, participants saw a picture depicting the Great Dane and the pug with prize-bows, and read: "These dogs are prize-winners. Notice the bow on them." (Fig. 4.11). Finally, participants completed a comparison class paraphrase practice trial, identical to the one used in Experiment 3. The warm-up trials in

**Warm–up trials**



These dogs are prize-winners. Notice the bow on them.
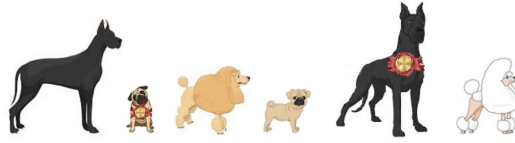
NEXT

Figure 4.11: Experiment 4 feature demonstration trial: Participants learn about additional features of the referents used on one critical main trial. For the dog-context participants learn that the pug and the Great Dane are prize-winners, indicated by bows on them.

the second experimental block were identical, but there was no paraphrase practice trial.

Then, participants completed four main trials — two critical and two filler trials, in randomized order, where a filler trials was always the first trial of the block. In the critical trials, a subordinate referent with an additional feature (e.g., a prize-winner bow) appeared in the corresponding context, as described above (Fig. 4.12). Participants read different context stories for each context (Table 4.3). For example, for a dog context, they read "You and your friend are at a pet show and you see the following:" above the context picture. Below, they read "Your friend runs far ahead of you. You see your friend in the distance:", followed by a depiction of the referent with the additional feature next to a person; to induce the illusion of distance, both were small relative to the context picture. Then they read "Your friend says:", followed by the critical sentence. Finally, they were asked: "What do you think your friend is saying it is big, small relative to?", introducing the paraphrase template, like in Experiment 3. For a given category, one of the possible targets appeared in this critical trial (e.g., the Great Dane). The other possible target (i.e., the pug) then appeared in a filler trial in the same block. Filler trials were identical to main trials with basic-level contexts and subordinate nouns from Experiment 3. The referent-size (i.e., large-subordinate vs. small-subordinate) was counterbalanced across syntactic conditions and trial types within-participant, resulting in 8 unique conditions. Each participant saw each condition once, resulting in eight main trials.[11]

---

[11]The experimental design described here is the final design which will be used in the main study. This design was used in the last pilot where 36 participants were recruited. The preceding pilot

Figure 4.12: Example critical main trial in Experiment 4: Participants see a dog-context and read the corresponding cover story. The small-subordinate referent is described by a predicate-noun sentence.

According to the reference-predication hypothesis, participants should be more likely to take the noun as signalling the comparison class when it appears in the predicate than when it appears in the subject *irrespective* of modification. That is, in the critical trials where the noun is always directly modified by the adjective, a higher proportion of subordinate comparison classes is expected to be inferred in the predicate condition, compared to the subject condition. A positive credible regression coefficient of the respective contrast would support this prediction. Further, since the filler condition replicates the condition of interest from Experiment 3, the same pattern is expected for those trials. Therefore, no difference is expected between the trial types, such that the respective regression coefficient for the trial effect is not expected to be credible.

## 4.4.1 Participants

The results reported here were gathered from 53 participants recruited in two pilot studies.[12] 5 participants were excluded for indicating a native language other than English, or failing either the comparison class inference practice trial or the labeling

---

where 17 participants were recruited only differed from the described design in that there were no feature demonstration trials, participants did not provide common labels in the labeling warm-ups and the conditions were not counter-balanced across referent-sizes.

[12]The number of participants for the main experiment was determined via a Bayesian power analysis and revealed that 300 participants are required for a power > 0.85 (Kruschke, 2014; Kurz, 2019). The main study is in progress.
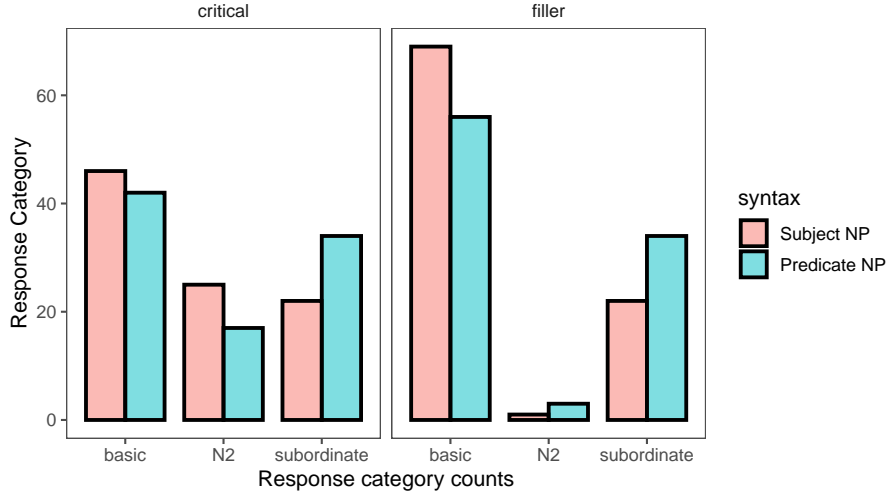
Figure 4.13: Response categories produced in Experiment 4 pilot: Counts of different response types (basic-level target labels, subordinate target labels, N2 denoting the visual feature in critical trials; x-axis) when the subordinate noun occured in different positions (color), by trial type (facets).

warm-up trials more than four times upon correction. The experiment took about 7 minutes and participants were compensated $1.00. If partial data was missing from a participant, available data was used for analyses.

## 4.4.2 Results

Participants' responses were manually classified into responses matching the critical subordinate referent label noun (i.e. subordinate comparison classes) and non-matching responses. That is, non-matching responses included basic-level or superordinate nouns as well as the second supplementary feature-nouns (e.g., "...relative to other prize-winners"). Compound modified nouns were classified according to the head noun: for instance, responses like "prize-winning Great Danes" were classified as matching, and responses like "floral gifts" were classified as non-matching. The overall proportion of responses matching the supplementary feature-nouns was 12.4 %, and more of these nouns were provided when this noun appeared in the predicate than in the subject position (Fig. 4.13). 13 uncategorizable responses (3.4%) were excluded from analysis. A Bayesian logistic mixed-effects regression model was used, predicting the response category from the syntactic condition (subject-N vs. predicate-N), the trial type (critical vs. filler) and their two-way interaction. By-participant and by-item random intercepts and random slope effects of both predictors and their interaction were included.[13] Default priors were used. The

---

[13]In brm-style syntax: `response_category ∼ syntax*trial_type + (1 + syntax*trial_type | subject) + (1 + syntax*trial_type | target)`
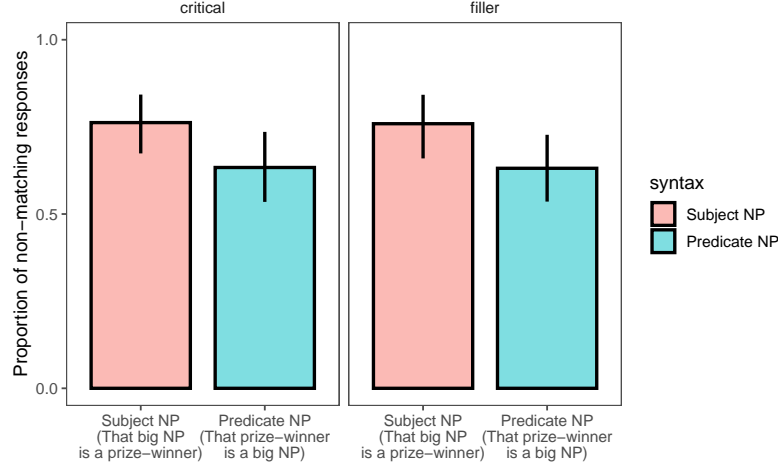
54

Figure 4.14: Experiment 4 pilot results: Proportions of inferred comparison classes in terms of responses not matching the critical subordinate target label (e.g., "...big relative to other dogs/prize-winners/animals"), depending on syntactic position of the noun (x-axis) and trial-type (facets). Error-bars denote bootstrapped 95% confidence intervals.

predictors were sum-coded, coding the subject-N syntax and filler trials as 1, the predicate-N syntax and critical trials as -1.

In line with predictions made by the reference-predication trade-off hypothesis, participants were sensitive to the syntactic position of the noun irrespectively of syntactic modification: even given directly modified nouns, in the critical trials pariticipants were more likely to infer the subordinate comparison class when the subordinate noun phrase appeared in the predicate compared to the subject position ($\beta = 2.14[0.10, 5.00]$; Fig. 4.14, left facet). Furthermore, supporting the findings from Experiment 3, in the filler trials participants also tended to infer more subordinate comparison classes from subordinate nouns in the predicate rather than in the subject position (the probability was 95.35% that the estimate was greater than 0: $\beta = 1.91[-0.34, 4.41]$; Fig. 4.14, right facet). Overall, participants inferred more subordinate comparison classes when the subordinate noun appeared in the predicate than when it appeared in the subject, across trial types ($\beta = 2.02[0.53, 3.90]$). Consistent with the prediction that syntactic modification which was varied between trial types should not play a role for comparison class inference, no credible effects of trial type or difference in syntax-effects on different trial-types were observed (trial effect: $\beta = 0.43[-1.38, 2.44]$, Fig. 4.14, left vs. right facet; trial × syntax interaction estimate: $\beta = -0.06[-0.98, 0.73]$, Fig. 4.14, left vs. right bar in left vs. right facet).

This experiment provided evidence that participants are able to flexibly reason about informational goals the nouns which combining with the adjective are intended for, above and beyond syntactic modification. These results show that the primary

driver behind the cue-strength of the noun towards the comparison class is indeed the informational goal that the noun accomplishes, and not the modification of the noun by the adjective. Therefore, it can be ruled out that nouns appearing in the predicate position of the utterance provided a stronger cue towards the comparison class than nouns in the subject just due to direct modification by the adjective (cf. experiments 1-3).

Together with Experiment 3, these results show that listeners flexibly adjust comparison classes by reasoning about informational goals and sentence structure, integrated with perceptual cues and their general world knowledge.

# Chapter 5

# A Bayesian Reference-Predication Model

Vagueness and context-dependence of gradable adjectives have been promisingly formalized in computational models within the Rational Speech Act framework - a suite of game-theoretically oriented recursive models of pragmatic language understanding (e.g., Goodman & Frank, 2016; Lassiter & Goodman, 2017; Tessler et al., 2017). Introduced by Frank and Goodman (2012), the Rational Speech Act framework is well in line with recent insights in Bayesian cognitive modelling, showing a great deal of flexibility to account for various phenomena studied in pragmatics like scalar implicature, hyperbolic language or generics, among many others (e.g., Scontras et al., 2018; Tenenbaum et al., 2011). This chapter reviews the basics of Rational Speech Act models and specifically prior models of gradable adjectives. While accounting for many aspects of gradable adjectives, many of those models assumed an a priori fixed comparison class. Yet as shown empirically in the previous Chapter 4, listeners flexibly reason about the intended comparison class by taking into account syntactic and contextual cues. Therefore, this chapter attempts to propose a minimal extension to existing RSA-models which will formalize the reference-predication trade-off hypothesis of comparison class inference.

## 5.1   Understanding Rational Speech Act Models

Language is fascinatingly flexible and efficient, and largely so because interlocutors do not have to explicitly encode all information in utterances they produce, but instead rely on each other's ability to infer many aspects of the message from linguistic and situational context. In particular, pragmatic models of communication posit that given these contextual constrains, speakers and listeners can efficiently *reason about each other's intended meaning of utterances* under one important as-

sumption: speakers are approximately *rational* with respect to their communicative goals (Frank & Goodman, 2012). The Rational Speech Act framework (henceforth: RSA) views this process of establishing the intended meaning as recursive reasoning between speaker and listener: in interpretation-oriented models, a pragmatic listener $L_1$ infers a state of the world intended to be conveyed by a rational speaker $S_1$, by using *Bayesian inference* to reason about likely world states given the observed utterance, given that the speaker $S_1$ chooses utterances according to their most likely semantic interpretation by a literal listener $L_0$ (Scontras et al., 2018).

The idea of language as rational action produced by *cooperative* interlocutors was formulated by Grice (1975). The core of his proposal are four conversational maxims that speakers are thought to stick to when producing utterances in order to convey particular messages: the *maxims of relation* (contributions made to the conversation are relevant), *quantity* (the contributions are as informative as required, but not more so), *quality* (the speaker believes their contributions to be true) and *manner* (the way the contributions are expressed is perspicuous). Listeners then reason about intended messages in light of these maxims (Grice, 1975).

Grice's ideas became particularly influential when precise information-theoretic formalisations of such vague concepts like *informativeness, cooperation* and *relevance* were proposed, and, informed by insights from game-theory, gave rise to RSA (Frank & Goodman, 2012). In particular, RSA captures cooperative coordination of intended meaning between interlocutors via recursive application of probabilistic mechanisms; and relevance or informativeness of utterances is captured as the *utility* of the utterance in helping the listener reduce uncertainty in their beliefs about the world, where these beliefs can be represented as a probability distribution over possible states of the world (as advocated by e.g. Tenenbaum et al., 2011). That is, listeners update their beliefs about the world via Bayes' rule upon hearing an utterance produced by an informative speaker (Frank & Goodman, 2012).[14]

The mechanisms of RSA are best illustrated by a simple example from a reference game, described by Frank and Goodman (2012). Consider a simple world consisting of a context $C =$ {blue square, blue circle, green square} (Fig. 5.1). In a reference game scenario, a speaker wants to communicate to a listener a particular referent $s$ in context $C$, e.g., the blue square. To do so, she has a finite set of utterances $U =$ $\{blue, green, square, circle\}$.[15] A listener then tries to recover the intended referent (i.e., the blue square) upon receiving an utterance (e.g., "blue"). As mentioned above, standard RSA models consist of three layers: a pragmatic speaker $S_1$ who

---

[14]Familiarity with basic notions in probability and Bayesian inference is presupposed here. For a review of aspects relevant for RSA, see e.g., Lassiter and Goodman (2017).

[15]The finite fixed set of alternative utterances is a crucial assumption made in RSA. It is an important question for future research how interlocutors actually determine this set of relevant alternatives.

Figure 5.1: A simple reference resolution example scenario: the context $C$ consists of three possible referents (Frank & Goodman, 2012)

chooses an optimal utterance for signalling $s$ (the blue square) to a literal listener $L_0$, who infers all the referents consistent with the literal meaning of the utterance $u$ ('blue'), and a pragmatic listener $L_1$ who reasons about this speaker behaviour given a particular utterance $u$ ('blue'), using Bayes' rule.

So the basis of RSA models is the naïve literal listener agent $L_0$ that $S_1$ reasons about when choosing an optimal utterance to communicate the blue square, who computes a probability distribution over possible referents in context $C$ consistent with the received utterance $u$:

$$P_{L_0}(s \mid u, C) = \frac{[\![u]\!](s)\, P(s)}{\sum_{s' \in C}[\![u]\!](s')\, P(s')} \tag{5.1}$$

The context $C$ is typically assumed to be shared between speaker and listener, so it will be dropped in further derivations for simplicity. Given that the denominator is a constant, it can also be dropped for simplicity, so that the probability of a particular state $s$ given utterance $u$ is proportional to the literal meaning $[\![u]\!](s)$ and the prior probability of $s$:

$$P_{L_0}(s \mid u) \propto [\![u]\!](s)\, P(s) \tag{5.2}$$

The prior $P(s)$ is the prior belief of $L_0$ about which states are likely to be communicated by the speaker. Typically, a uniform prior is used, indicating that a priori any state is as likely as others, but relevant contextual information like perceptual salience or frequency of some referents might be encoded in this prior (Frank & Goodman, 2012).

The second component of the $L_0$ is the *literal meaning* of the observed utterance $u$ (indicated as $[\![u]\!]$). In RSA, literal semantics computation is based on a form of Montague's compositional semantics, typically assuming a mapping from particular states to Boolean truth-values (Montague, 1973) (but see e.g. Degen et al., 2020, for alternative approaches). So, for instance for the context Fig. 5.1, applying the

utterance 'circle' to the blue square would return `false`, but 'blue' would be `true`:

$$[\![circle]\!](blue\ square) = 0$$

$$[\![blue]\!](blue\ square) = 1 \tag{5.3}$$

So for our example utterance 'blue' the literal listener $L_0$ infers a uniform distribution over the blue circle and the blue square, since the utterance equally applies to both objects (Table 5.1):

Table 5.1: The probability distribution over states inferred by $L_0$ when hearing the utterance 'blue' in context $C$.

| State | Probability |
| --- | --- |
| blue circle | 0.5 |
| blue square | 0.5 |

The next RSA layer is the pragmatic speaker $S_1$. $S_1$ is modelled as an agent who chooses an utterance $u$ rationally, i.e., according to its expected utility, in order to communicate a particular state of the world $s$ in context $C$ to $L_0$. This is captured in the speaker-utility function $U_{S_1}(u; s)$, which trades-off the informativity of an utterance for $L_0$ with non-negative cost $C(u)$ of producing the particular utterance (Scontras et al., 2018):

$$U_{S_1}(u; s) = log L_0(s \mid u) - C(u) \tag{5.4}$$

In information-theoretic terms, $L_0$ provides a hook within this utility function to compute the *informativeness* of particular utterances as communicating particular states, where informativeness is quantified by surprisal - a measure of how much uttering a particular $u$ reduces uncertainty about the state of the world, given that $u$ is *true of* $s$ (i.e., $[\![u]\!](s) = 1$) (Frank & Goodman, 2012):

$$I_{\tilde{u}(s)}(s) = -log(\tilde{u}(s)) \tag{5.5}$$

$I_{\tilde{u}(s)}(s)$ measures how much information $L_0$ gains when hearing the utterance $u$, assuming a known distribution $\tilde{u}(s)$ over states that are coveyed by the literal interpretation $[\![u]\!]$ which implies the probability of $s$; i.e., it measures how *surprising* it would be to observe $s$ having observed $u$. Intuitively, assuming a uniform $\tilde{u}(s)$, the less states an utterance applies to, the lower is the surprisal of a particular state, and the higher is its informativeness. Here, informativeness is used in a relatively informal way, meaning that a maximally informative utterance conveys the a message maximally unambiguously. For instance, in the context of Fig. 5.1, the utterance

'circle' is highly informative, because there is only one object it applies to, while the utterance 'blue' is less informative because it applies to two objects. Therefore, the speaker utility is anti-proportional to the surprisal of the utterance (Frank & Goodman, 2012):

$$U_{S_1}(u; s) = -(-log(\tilde{u}(s))) - C(u) = logL_0(s \mid u) - C(u) \tag{5.6}$$

The cost function $C(u)$ is also an important tool for integrating psychologically plausible information about speaker-biases, like frequency or phonological complexity of producing particular utterances compared to others. Now the rational speaker $S_1$ maximizes the probability of conveying the intended state of the world $s$ acting according to Bayesian decision theory, by choosing an utterance $u$ proportionally to its expected utility described by the *softmax* function:

$$P_{S_1}(u \mid s) = \frac{e^{\alpha\, U_{S_1}(u;s)}}{\sum_{u' \in U s.t. u'(s)=true} e^{\alpha\, U_{S_1}(u';s)}} \tag{5.7}$$

The parameter $\alpha$ controls the speaker's *optimality*, assuming $\alpha = 1$ in examples used here; for $\alpha = \infty$ the fully rational decision rule used in game-theory can be recovered (Lassiter & Goodman, 2017; Scontras et al., 2018).

For this example, $S_1$ chooses an utterance $u$ maximizing the probability of the state 'blue square' being recovered by $L_0$. So $S_1$ infers a distribution over utterances applicable to the target 'blue square' (Table 5.2):

Table 5.2: The probability distribution over utterance inferred by the pragmatic speaker $S_1$ in order to communicate the referent 'blue square'

| Utterance | Probability |
|-----------|-------------|
| blue | 0.5 |
| square | 0.5 |

Finally the top-level layer, the pragmatic listener $L_1$, reasons about this utterance-generating speaker behaviour given a particular utterance $u$ ('blue'), using Bayes' rule:[16]

$$P_{L_1}(s \mid u) = \frac{P_{S_1}(u \mid s)\, P(s)}{\sum_{s' \in C} P_{S_1}(u \mid s')\, P(s')} \tag{5.8}$$

That is, the probability of a particular state $s$ (i.e., blue square) given the utterance $u$ ('blue') is equal to the probability that the pragmatic speaker $S_1$ would choose

---

[16]This recursive depth of three model layers is a common assumption in RSA models, and requires a reasonable amount of computational resources (Lassiter & Goodman, 2017). Yet this is just a practical approximation, and some models (e.g., production-oriented models) employ additional levels (Scontras et al., 2018).
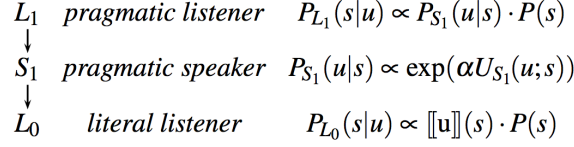
$$
\begin{array}{lll}
L_1 & \textit{pragmatic listener} & P_{L_1}(s|u) \propto P_{S_1}(u|s) \cdot P(s) \\
\downarrow & & \\
S_1 & \textit{pragmatic speaker} & P_{S_1}(u|s) \propto \exp(\alpha U_{S_1}(u;s)) \\
\downarrow & & \\
L_0 & \textit{literal listener} & P_{L_0}(s|u) \propto [\![u]\!](s) \cdot P(s)
\end{array}
$$

Figure 5.2: A schematic depiction of a vanilla RSA model (Scontras et al., 2018)

*blue* in order to convey the *blue square*, multiplied by the prior probability $P(s)$ of occurence of state $s = blue\,square$, normalised by a constant sum of probabilities of all possible speaker behaviors for all possible states $s'$. Equation 5.8 shows that the posterior over states $s$ given an utterance $u$ is proportional to the speaker production probability $P_{S_1}(u \mid s)$ times the state prior $P(s)$ because the denominator is a constant:

$$
P_{L_1}(s \mid u) \propto P_{S_1}(u \mid s)\, P(s) \tag{5.9}
$$

Interestingly, the state prior $P(s)$ might differ between $L_0$ and $L_1$, e.g. incorporating prior world knowledge of the pragmatic agent $L_1$, but being uniform for the naïve agent $L_0$ (Scontras et al., 2018).

In this example, upon hearing 'blue', $L_1$ infers that the speaker is more likely to convey the blue square (Table 5.3):

Table 5.3: The probability distribution over referents inferred by the pragmatic listener $L_1$ upon hearing the utterance 'blue'.

| State | Probability |
|---|---|
| blue square | 0.6 |
| blue circle | 0.4 |

Putting all the elements together results in the vanilla version of an RSA-model (Fig. 5.2). The crucial illustration of the RSA mechanism is the difference between the distributions inferred by $L_0$ and $L_1$ upon observing the same utterance 'blue'. Reasoning about the utterance-generating speaker-model incorporated in $L_1$, while $L_0$ acts according to literal semantics only, is crucial for the pattern of interpretation we observe: $L_1$ infers that the speaker is more likely to mean the blue square, because if she had meant the blue circle, she could have said 'circle', which would have been less ambiguous, and therefore more informative (Table 5.3). That is, $L_1$ does not only consider what the speaker actually said, but also what the speaker *could have said*. In contrast, $L_0$ infers equal probabilities of 'blue' meaning 'blue square' or 'blue circle' because the utterance is literally true of both referents (Table 5.1). Crucially, this pattern predicted by the RSA-model is well in line with rational behaviour of humans gathered empirically in such reference game scenarios, confirming that

human interlocutors make use of more than just the literal semantics of their words when communicating (Frank & Goodman, 2012; Scontras et al., 2018).

## 5.2 Previous RSA Models of Gradable Adjectives

In this section, previous RSA models of gradable adjectives are discussed; they showed that RSA is a flexible tool suitable to integrate more complex pragmatic reasoning required for interpreting vague expressions.

Lassiter and Goodman (2013) first provided a model of gradable adjective interpretation within the RSA-framework, showing that pragmatic reasoning can capture their meaning via inference over the latent standard of comparison variable $\theta$ underlying the vague semantics (s. Chapter 2). The proposed model builds upon the standard RSA model with three levels, adding one crucial extension: the pragmatic listener $L_1$ jointly infers the value of the threshold $\theta$ along with the state of the world $s$ (i.e., the degree to which a referent possesses the property described by the adjective):

$$P_{L_1}(s, \theta \mid u) \propto P_{S_1}(u \mid s, \theta) \, P(s) \, P(\theta) \tag{5.10}$$

The model formalized literal meaning of gradable adjectives in terms of degree-semantics, assuming that the lexical entry of the adjective specifies the underlying scale and its polarity (cf. Chapter 2):

$$[\![u]\!]_\theta(s) = s > \theta \tag{5.11}$$

However, the literal meaning of the adjective by itself is underspecified - it depends on the pragmatically recovered threshold $\theta$, which in turn depends on the comparison class used in the particular context; yet, like vanilla RSA the model is anchored in the literal interpretation of the adjective applied to a referent at the level of $L_0$. So in order to allow specifying the $L_0$ which requires computing the truth-value of an utterance for a given state, the authors proposed $L_1$ consider all possible assignments of the *lifted* latent variable value $\theta$, given a prior over that variable $P(\theta)$. A uniform $P(\theta)$ encoded that a priori any property degree might qualify for being described by the adjective. Crucially, the state prior $P(s)$ encoded prior knowledge about *likely property degrees for a specific comparison class* which the referent might have. For instance, the difference in the meanings of *big* in 'big for a tree' and 'big for a pug' would be encoded in different priors over the likely properties of the respective referents. It is common practice to quantify this prior knowledge empirically via prior elicitation experiments (Scontras et al., 2018).

The assumed $\theta$ values are then iteratively passed down through the model. Given

a particular value, the speaker-model can be specified:

$$P_{S_1}(u \mid s, \theta) \propto exp(\alpha \, log(P_{L_0}(s \mid u, \theta) - C(u))) \qquad (5.12)$$

Then, in contrast to $L_1$ who is uncertain about $\theta$, $L_0$ interprets the utterance literally, so she gets $\theta$ passed down and infers the state distribution consistent with this particular $\theta$ assignment and the observed adjective $u$:

$$P_{L_0}(s \mid u, \theta) = P_{L_0}(s \mid [\![u]\!]^\theta = 1) \propto [\![u]\!]^\theta(s) \, P(s) \qquad (5.13)$$

So putting all the layers together, $L_1$ can compute a joint posterior distribution over all possible combinations of states and values of the latent threshold $\theta$ lifted to the level of the pragmatic listener. Notably, Lassiter and Goodman (2013) assumed the relevant comparison class to be supplied to the listener, such that she was uncertain only about the standard of comparison. Yet as discussed in chapters 2-3 and shown empirically in Chapter 4, listeners flexibly reason about the intended comparison class because there might be multiple a priori plausible options in absence of an explicit *for*-phrase.

Tessler et al. (2017) introduced an RSA-model of gradable adjectives accounting for flexible reasoning about the relevant comparison class via world knowledge (the rationale and behavioural experiments of this study are described in detail in Section 2.4). In particular, in the proposed model the listener is not only uncertain about the standard of comparison $\theta$, but also about the specificity of the relevant comparison class $c$ (superordinate vs. subordinate category of the referent), given prior knowledge about typical property distributions in each category. That is, listeners were assumed to know the a priori probability that an adjective could felicitously apply to a referent given its category, assuming a specific comparison class; they used this knowledge to infer the intended comparison class upon hearing a simple adjectival utterance $u$ of the form '*PRON* is *ADJ*' said of a referent whose *subordinate category was known to the listener*. Similarly to the model proposed by Lassiter and Goodman (2013), the pragmatic listener iterated over all possible value assignments of the lifted $\theta$ variable when reasoning about the utterance-generating process, to jointly infer the property degree $s$, the threshold $\theta$ and the relevant comparison class $c$, given the utterance $u$:

$$P_{L_1}(s, \theta, c \mid u) \propto P_{S_1}(u \mid s, \theta, c) \, P(s \mid c_{sub}) \, P(c) \, P(\theta) \qquad (5.14)$$

That is, the listener reasoned about how a rational speaker $P_{S_1}$ would behave in order to communicate a specific property, given a comparison class and a threshold, together with their prior knowledge about what property degrees are plausible

given the subordinate category $P(s \mid c_{sub})$ the referent belongs to, and their prior beliefs $P(c)$ about which comparison class categories are likely to be used and what properties are likely to qualify for applying the adjective ($P(\theta)$). Accordingly, the speaker-model $P_{S_1}(u \mid s, \theta, c)$ can be specified assuming particular values of the property, the comparison class and the threshold:

$$P_{S_1}(u \mid s, \theta, c) \propto exp(\alpha \, logP_{L_0}(s \mid u, \theta, c)) \qquad (5.15)$$

The $L_0$ again specifies a literal listener who interprets the adjective $u$ according to its literal semantics, assuming a particular comparison class $c$ and $\theta$:

$$P_{L_0}(s \mid u, \theta, c) \propto [\![u]\!]_\theta(s) \, P(s \mid c) \qquad (5.16)$$

Importantly, $L_0$ now samples states according to a prior $P(s|c)$ specified by the received comparison class $c$, which integrates the comparison class into the literal semantics.

Since the predictions of this model strongly depend on prior world knowledge, it is important how the property priors given different comparison classes are specified. Tessler et al. (2017) fixed superordinate property distributions as $\mathcal{N}(0, 1)$ and subordinate property distributions as $\mathcal{N}(\mu_{sub}, \sigma_{sub})$, inferring the paramters of subordinate distributions from experimental data. The comparison class prior $P(c)$ was approxiamted as empirical Google WebGram frequencies of usage of the respective comparison class labels.

Tessler et al. (2017) showed that the model captures human inferences discussed in Section 2.4 very well, confirming the role of world knowledge for comparison class inference. However, this model only considered the effect of world knowledge on comparison class inference, and focused on interpretation of simple underspecified utterances of the form 'He is tall'. In the following section, a further extension of previous RSA models is proposed, which accounts for more complex sentences containing a noun and integrates more sophisticated reasoning about several effects contributing to comparison class inference.

## 5.3 Reference and Predication in RSA

This section outlines a novel RSA model formalizing the reference-predication trade-off hypothesis.

First, let's recall the reference-predication trade-off hypothesis: it posits that speakers pursue particular *informational goals* — *reference* and *predication* — when crafting their utterances. These informational goals are realized by the speaker

syntactically via different positions of the noun: the noun is more likely to contribute to reference when appearing in the sentence subject, especially when combining with the deictic "that", but less likely to do so when appearing in the predicate. Hence, nouns appearing in the subject can be potentially *explained away* by their utility in reference, while nouns appearing in the predicate are less likely to contribute to reference and therefore more likely to contribute to predication — i.e., constrain the comparison class. Listeners, in turn, reason about these speaker goals when interpreting the observed utterance. Therefore, the degree to which the noun of the sentence constrains the comparison class — i.e., serves predication — trades off with its utility in reference.

### 5.3.1   Questions Under Discussion in RSA

Reasoning about particular informational, or conversational, goals is closely related to basic assumptions made in RSA. As described in Section 5.1, the speaker is assumed to be a rational agent producing utterances of optimal utility with respect to some specific *conversational goal* determined by discourse, also called *question under discussion (QUD)* (Lassiter & Goodman, 2017; Roberts, 2012). The communication is structured so as to maximize the probability that the listener infers the intended *answer to this QUD*. In the vanilla example introduced in Section 5.1, the QUD was to determine the intended referent, and the intended answer was 'blue square'. From this perspective, interlocutors' contributions to discourse are *relevant* if and only if they contribute to answering the question under discussion (Roberts, 2012). Note that such questions under discussions might indeed be realized by specific speech acts, but do not have to be - they often remain implicit, and resemble questions rather formally in that they proffer the set of relevant alternatives interlocutors commit to addressing. Yet, interlocutors are fascinatingly good at determining the QUD pragmatically so as to functionally organize their communication around the QUD; they might do so by using different strategies, one prominent strategy in English being to make use of focus (Krifka, 2008; Roberts, 2012). But they might also employ other strategies, like choosing particular syntactic structures over others — e.g., as proposed by the reference-predication trade-off hypothesis.

From a semantic perspective, utterances then might have several dimensions of interpretation, each dimension corresponding to the interpretation of the utterance satisfying a distinct communicative goal, or answering a distinct QUD. This idea was first formlazed within the RSA-framework by Kao et al. (2014), applied to hyperbolic and pragmatic halo effects in interpretation of number words. The crucial novelty they introduced was listener uncertainty about the relevant QUD along with the multi-dimensional literal meaning of utterances, where each dimension satisfied a

distinct potential QUD. The speaker in this model thus chose utterances optimally conveying her particular intended QUD (Kao et al., 2014).

Another approach to formalizing multiple conversational goals was taken by Yoon et al. (2016) who proposed a model of polite language. The authors noted that although politeness phenomena were argued to violate basic cooperative principles of quality, they can be explained as produced by rational interlocutors pursuing not only the goal of informativeness, but simultaneously also a *social goal* (cf. Brown & Levinson, 1987; Yoon et al., 2016). That is, by using polite sentences speakers strive to trade off being informative and saving the listener's self-image, or *face*. Omitting here the details of how exactly this social goal was formalized, the authors proposed a model wherein the speaker tried to balance the utility of an utterance as simulatneously achieving the social goal and the goal of being maximally informative. She did so by maximizing a global utility function, which consisted of a weighted combination of these two utility sub-functions. The pragmatic listener then reasoned about the exact weighting of these two goals (Yoon et al., 2016).

Yet the majority of RSA models usually assume that speakers address one particular conversational goal. The listener might be uncertain about it, as in case of QUD-models, but mostly models might just addressed one pre-defined aspect of communication, i.e., one QUD (see Scontras et al., 2018, for an excellent overview of various RSA models).

This work has argued that interlocutors at least consider multiple informational goals (i.e., QUDs) that might be addressed within the same utterance containing a noun and a gradable adjective, reasoning how different parts of the utterance might probabilistically contribute to achieving each of the goals. Therefore, in the following a model is proposed which attempts to formalize the reference-predication hypothesis. It treats speakers' choice of utterances as *incremental*, where the subject is chosen so as to establish reference, and the predicate is chosen to convey the comparison class — and so one utterance is optimized to fulfil multiple QUDs.

## 5.3.2 Refpred-RSA Model

In the following, the basic structure of the novel reference-predication RSA-model is presented. Then, model predictions are qualitatively compared to the results of the Comparison Class Inference Experiment (s. Section 4.3). It will be shown that minimal extensions to existing models result in accurate predictions about the identity of an unknown referent, its property, and crucially, the intended comparison class. The model is intended to predict comparsion class inferences drawn from sentences differing in the type of the noun (basic vs. subordinate referent label) and its position (subject vs. predicate), investigated empirically in Experiment 3 (s. Section 4.3).

The anaphoric "one" condition is disregared, because it was rather an experimental sanity-check condition. This model is inspired by previous gradable adjective inter-pretation models: The representation of adjectival meaning is represented in terms of threshold semantics, similarly to the proposal by Lassiter and Goodman (2013); the comparison class representation is inspired by the model by Tessler et al. (2017).

In the basic model, upon hearing an utterance of the form "That N is ADJ" or "That's a ADJ N" the pragmatic listener tries to resolve his uncertainty about the specific member of the context the speaker is referring to, its size and the com-parison class intended by the speaker. As a first approximation, driven by present experimental results and previous work by Tessler et al. (2017), it is assumed that the potential comparison class is either the subordinate or the basic-level category of that referent. The potential referent might be one of the members of the percep-tual context (i.e., basic-level or subordinate context), though the listener knows the subordinate category of the intended referent.[17] Therefore, the listener is assumed to know the subordinate categories of all possible referents, which informs their po-tential sizes. Formally, the listener infers the comparison class $cc$, the referent $r$ and its size $s$ given an utterance $u$ and a perceptual context $C$:

$$P_{L_1}(r, s, cc \mid u, C) \propto P_{S_1}(u, cc \mid s, r, C)\, P(r, s \mid C, cc_{r\_sub}) \qquad (5.17)$$

That is, the listener infers the referent, its size and the comparison class by reasoning about how a speaker would behave in order to communicate a particular referent and its size in context, in addition to his prior knowledge about different subordinate categories. All referents in context are treated as equally likely to be discussed, but more sophisticated models might upgrade the prior to incorporate perceptual salience, for instance for referents standing out with respect to their size.

Importantly, rather than representing the comparison class as a variable lifted to the level of the pragamtic listener $L_1$ (as proposed by Tessler et al., 2017), the intended comparison class influencing the resolution of the adjective is represented as a concious *lexical choice, or a local enrichment* made by the speaker (e.g., discussed by Chierchia et al., 2012; Scontras et al., 2018). So while the pragmatic listener in Tessler et al. (2017) assumes that the speaker has a fixed meaning of the adjective (e.g., "big for a dog" or "big for a Great Dane") which the listener is uncertain about, in the current model the pragmatic listener reasons about a speaker who might *variably* choose her intended adjective meaning (cf. Scontras et al., 2018). This was proposed in prior work arguing that pragmatic enrichments might already

---

[17]Note that this representation might not fully correspond to experimental set-up: The intended referents of the sentences were made salient by presenting them in a picture separate from the context. However, if anything, this representation might be seen as more conservative than the empirical set-up, because it puts a generally higher referential pressure on the speaker.

take place at subsentential — i.e., lexical — level, such that the speaker chooses to convey one of grammatically supplied local phrase readings with her utterance (for instance, applied to scalar implicatures by Chierchia et al., 2012). In the current case, the speaker can choose between the subordinate and basic-level comparison class relative to which the adjective will be interpreted, in addition to the option to use a predicative noun as the comparison class.

So the speaker determines an *optimal utterance and comparison class* in order to communicate a referent and its size, given the adjective she wants to use and the context, by optimizing the speaker-utility function $U$:

$$P_{S_1}(u, cc \mid s, r, C) \propto exp(\alpha\, U_{S_1}(u; s; r; cc; C)\; P(cc)\; P(u)) \qquad (5.18)$$

The cost of producing uterances $C(u)$ is assumed to be uniform over all possible utterances and is therefore neglected. The optimality parameter $\alpha$ was arbitrarily set to a generic value of 3 for the simulations presented below. The utterance utility $U$ is combined with the speaker's prior $P(cc)$ over possible comparison classes (a uniform distribution over basic vs. subordinate referent categories) and possible utterances ($P(u)$). As mentioned above, one novel aspect of this model is that the speaker constructs the utterance incrementally: she considers the subject and the predicate separately, choosing uniformly at random where to put the noun. Since two potential comparison classes are considered in this model, the possible noun options are the subordinate or the basic-level label of the intended referent. That is, the speaker may put the noun in the subject, so the sentence would match the experimental subject-N condition. The predicate is then a bare adjective ("big" or "small"). Alternatively, the speaker may put the noun in the predicate, directly modified by the adjective; the subject is then the bare deictic "that", resulting in the predicate-N experimental condition. Furthermore, the speaker may choose the adjective to convey one of the two possible comparison classes, since she infers the optimal comparison class: for instance, she might choose "big" to mean "big for a dog" or "big for a Great Dane", irrespective of the noun position. This results in eight possible utterances the speaker considers (noun position × noun type × comparison class).

In contrast to models including several QUDs reviewed in Section 5.3.1, for this case-study one might posit that speakers a priori pursue both the goal of reference and the foal of predication. Therefore, most intuitively, the speaker utility can be represented jointly with respect to two-dimensional states of the world consisting of a referent and a size dimension she wishes to communicate (but see Appendix A for alternative formalizations, following those discussed in Section 5.3.1). Conceptually, when communicating the referent she addresses the reference-QUD, when

communicating the size — the predication-QUD, respectively. According to the operationalization assumed in this work, the subject part of the utterance is optimized for transmitting reference, the predicate for predication:

$$U_{S_1}(u; r; s; cc; C) = log\, L_0(s, r \mid u, C, cc) \qquad (5.19)$$

In line with basic RSA mechanisms, the speaker utility is grounded in a listeral listener $L_0$ whom the speaker reasons about when choosing the utterance and the comparison class. $L_0$ then interprets the utterance chosen by $S_1$ according to its literal meaning, considering the subject as establishing reference, and the predicate as communicating the size of the referent. That is, $L_0$ returns a joint distribution over referents and over properties from the context for which the utterance subject is true, and which are possible under the comparison class and the adjective:

$$P_{L_0}(r, s \mid u, cc, C) \propto [\![u_{Subj}]\!](r)\, [\![u_{Pred}]\!]^{\theta}(s)\, P(r, s \mid C, cc)\, P(\theta) \qquad (5.20)$$

The state prior $P(r, s \mid C, cc)$ represents the literal listener's prior beliefs about possible referents and sizes; crucially, the possible sizes now depend on the explicit comparison class $cc$ or the utterance $u$ passed to $L_0$ rather than on the subordinate category of a sampled referent (as for $L_1$). Critically, the reference-predication hypothesis posits that listeners might consider nouns appearing in the predicate as the comparison class; therefore, the literal listener probabilistically considers the noun to be the comparison class, when it is observed in the predicate (taking uniformaly at random the noun or the comparison class $cc$). This provides a hook for representing the effect of a comparison class: for example, intuitively, the distribution of sizes that would count as *big for a Great Dane* is a priori different from the distribution of sizes that would count as *big for dogs*. These size priors $P(cc)$ are represented by Gaussian distributions relative to particular comparison classes, as proposed by, Tessler et al. (2017). Since this work is concerned with qualitative predictions only, the priors representing properties of different classes are set to $\mathcal{N}(\mu = 0, \sigma = 1)$ for the basic-level comparison class, $\mathcal{N}(\mu = -1, \sigma = 0.5)$ for small-subordinate, $\mathcal{N}(\mu = 0, \sigma = 0.5)$ for medium-subordinate and $\mathcal{N}(\mu = 1, \sigma = 0.5)$ for large-subordinate comparison classes (Fig. 5.3). Furthermore, the inference over the threshold of comparison $\theta$ is moved to the level of $L_0$ for better computational tractability, different from the original proposal by Lassiter and Goodman (2013).

Since the literal listener infers states consistent with the literal meaning of utterances, the representation of literal semantics of the utterance she observes is a crucial component of the model. This model employs classic Boolean semantics common in RSA models (Montague, 1973). For the meaning computation, the utterance is
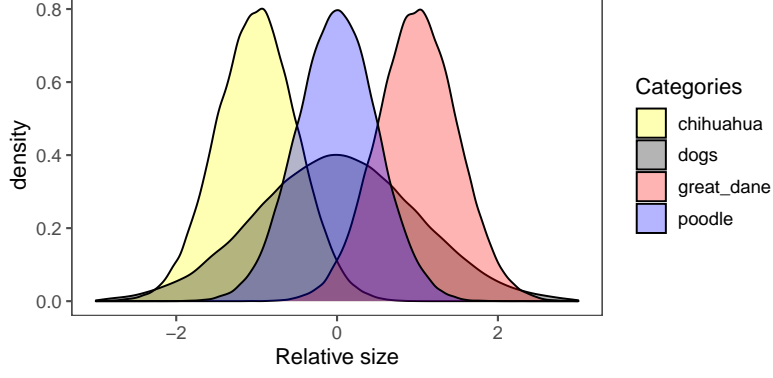
70

Figure 5.3: Hypothetical prior size distributions over a basic-level, a small-subordinate, a medium-subordinate and a large-subordinate category. These distributions were used for qualitative tests of the refpred-RSA model.

split into subject and predicate, and each component is evaluated with respect to one goal (reference or predication, respectively); an utterance is assumed to be true of a state if and only if both components are true. When computing the meaning of the subject, the bare deictic 'that' and the basic-level nouns are assumed to be true of any member of the context, while subordinate labels only apply to respective category members. The referential utility of the subject is therefore formalized as in the vanilla RSA model (s. Section 5.1). The meaning of the predicate is computed in terms of threshold semantics, as first proposed by Lassiter and Goodman (2013):

$$[\![u]\!](s,r) = [\![u_{Subj}]\!](r) \wedge [\![u_{Pred}]\!]_\theta(s) = [\![u_{Subj}]\!](r) \wedge \theta > s \qquad (5.21)$$

So putting all these elements together, the proposed model formalizes a pragmatic listener inferring an intended referent in context, its size and the intended comparison class upon observing a sentence containing a noun and an adjective:

$$L_1(r, s, cc \mid u, C) \propto P_{S_1}(u, cc \mid s, r, C) \, P(r, s \mid C, cc_{r\_sub})$$
$$S_1(u, cc \mid s, r, C) \propto exp(\alpha \, log \, L_0(s, r \mid u, C, cc) \, P(cc) \, P(u))$$
$$L_0(r, s \mid u, cc, C) \propto [\![u_{Subj}]\!](r) \, [\![u_{Pred}]\!]^\theta(s) \, P(r, s \mid C, cc) \, P(\theta)$$

In order to investigate whether the qualitative predictions of this model capture the essence of predictions made by the reference-predication hypothesis, supported by the data observed in Experiment 3, the model was implemented using the probabilistic programming language WebPPL (Goodman & Stuhlmüller, 2014). The pragmatic listener model was then tested on example sentences matching critical conditions from Experiment 3 (s. Section 4.3). Most importantly, the model should capture the contrast in comparison classes inferred from subordinate-noun sentences
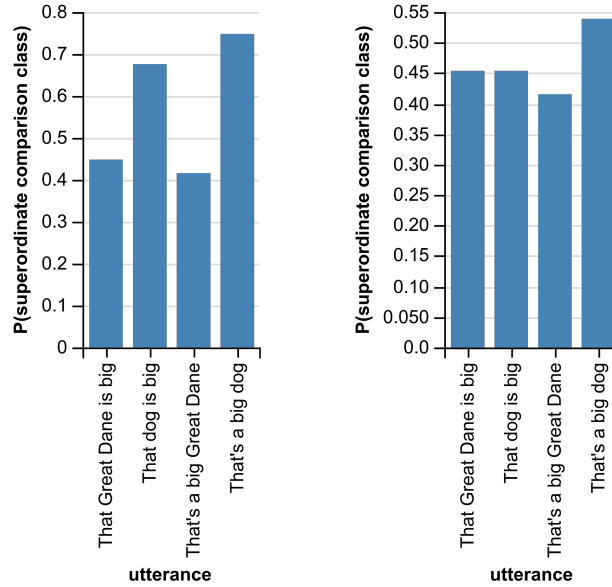
Figure 5.4: Qualitative predictions made by the refpred-RSA model: Pragmatic listener inferences are plotted in terms of the probability to use the basic-level comparison class (i.e., "big for a dog"), given utterances differing in the noun and its position, presented in a basic-level context (left) vs. subordinate context (right). Qualitatively, the crucial noun×syntax interaction can be observed. Note the different y-axis scaling.

differing in the position of the noun, distinct from basic-level noun sentences.

Figure 5.4 (left) shows the proportion of basic-level comparison classes (i.e., "big for a dog") the model predicts upon observing example utterances containing the nouns "dog" vs. "Great Dane", appearing in the subject vs. predicate position, given a basic-level context. Visually, the critical contrasts can be observed: the model predicts that the basic-level comparison class is less likely given the utterance "That's a big Great Dane", relative to "That Great Dane is big". Furthermore, the model captures that the basic-level comparison class is generally less likely if the utterance contains a subordinate noun. Finally, the pragmatic listener infers more basic-level comparison classes given the utterance "That's a big dog" than the utterance "That dog is big", which is generally consistent with reference-predication hypothesis predictions, though was not observed in Experiment 3 due to a ceiling effect in the basic-level context (s. Figure 4.10).

The model performs comparably well given the same utterances in *subordinate* context (Fig. 5.4, right). First, consistent with data observed in Experiment 3, the model captures a grand influence of context: overall, the basic-level comparison class is less likely given the subordinate context than the basic-level context (Fig. 5.4, left vs. right). Second, the model captures reasoning about contextual referential utility, rendering both nouns in the subject position equally referentially
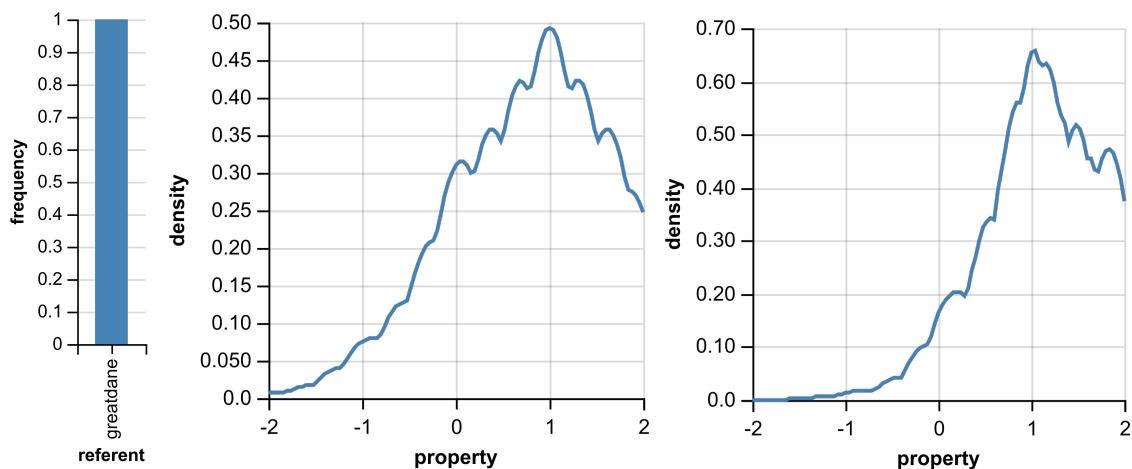
Figure 5.5: Qualitative predictions made by the refpred-RSA model: Given the utterance "That Great Dane is big", the pragmatic listener infers a large-subordinate size distribution (middle) and is certain that the referent is a Great Dane (left). Given the utterance "That's a big Great Dane", the pragmatic listener shifts her size distribution even more towards large size values (right), and is again certain about the referent (left).

uninformative. Finally, both nouns in the predicate position are more likely to signal their respective comparison class than in the subject position.

In both contexts, given the critical utterances containing a subordinate noun, the model infers the expected referent and size distributions (Fig. 5.5). That is, the inferred referent matches the noun; the inferred size distribution is shifted more strongly towards the subordinate category size values when observing an utterance containing a predicate noun relative to an utterance containing a subject noun. When the noun is the basic-level target label, in the basic-level context the pragamtic listener is uncertain about the intended category of the referent (Fig. 5.6). As predicted, the inferred size distributions do not differ qualitatively in the basic-level context across the two sentence frames, but do so in subordinate context: when the noun appears in the predicate, the listener shifts her distribution more towards the basic-level category properties, but infers a more subordinate — contextually informed — distribution when the noun is in the subject (Fig. 5.7). These results provide further support that the model represents the reference-predication hypothesis well.

In sum, this model formalizes the reference-predication trade-off hypothesis via recursive probabilistic reasoning, representing a pragmatic listener as considering an
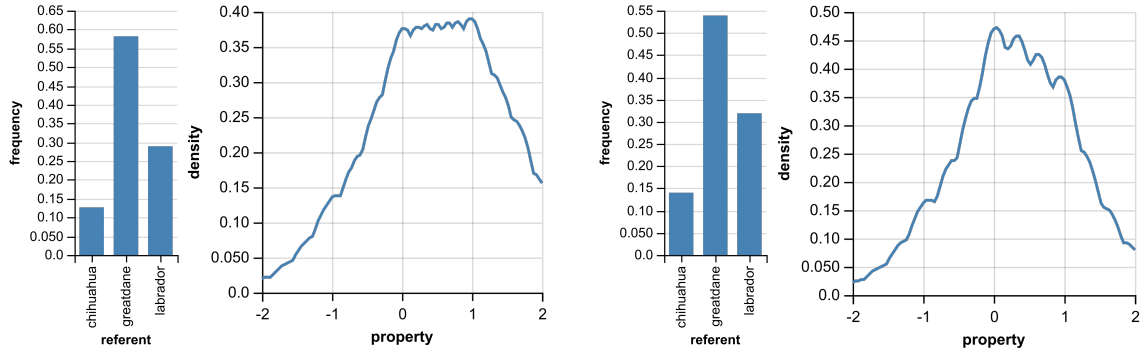
Figure 5.6: Qualitative predictions made by the refpred-RSA model given the utterances with a basic level noun in basic-level context (from left to right): distribution over referents, distribution over sizes inferred from "That dog is big" (subject N); distribution over referents, distribution over sizes inferred from "That's a big dog".

agent choosing the adjective to convey a particular comparison class. It integrates reasoning about both syntactic and semantic aspects of the utterance and generally captures comparison class inferences observed empirically, driven by the trade-off of the noun utility for different informational goals. It was shown that such a complex inferential hypothesis can be qualitatively explained by minimally extending existing generic Rational Speech Act tools. To the best of our knowledge, this is the first attempt to formalize reasoning about both meaning and structure of an utterance within the RSA framework, contributing to both creating more accurate gradable adjective interpretation models and extending the scope of RSA models.
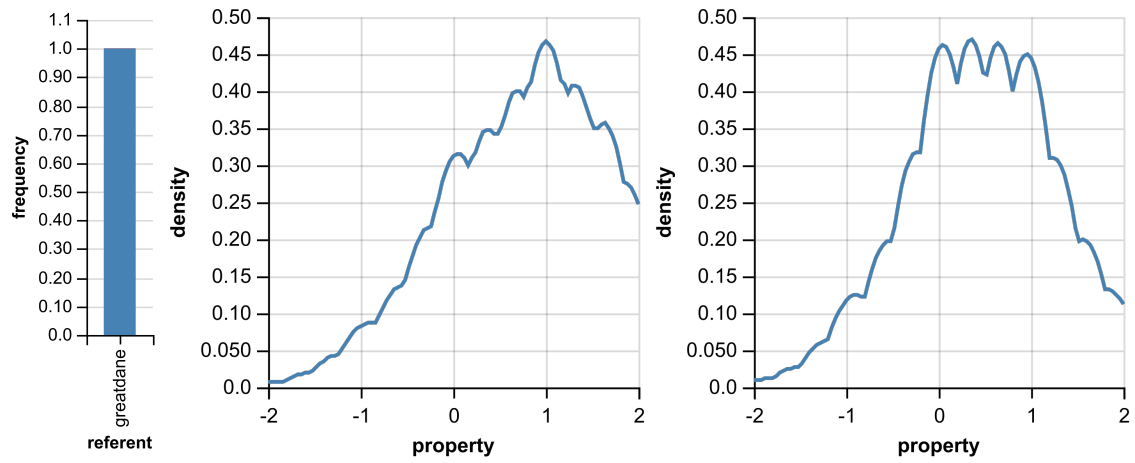
Figure 5.7: Qualitative predictions made by the refpred-RSA model given the utterances with a basic level noun in subordinate context (from left to right): distribution over referents, distribution over sizes inferred from "That dog is big" (subject N); distribution over sizes inferred from "That's a big dog".

# Chapter 6

# Discussion

Humans use language across infinitely many different situations. By the age of two infants already learn that the same words might be used in different contexts and convey different meanings, allowing for fascinatingly versatile and efficient language use (Ebeling & Gelman, 1994; Mintz & Gleitman, 2002). Relative gradable adjectives are one, but not the only, example of natural language expressions whose meaning greatly depends on context (e.g., next to indexicals, Zalta (2017); or anaphoras, Goldberg and Michaelis (2017), inter alia). Yet speakers almost never establish explicitly which aspects of context are relevant and leave it to the listener to pragmatically reconstruct. Interpreting gradable adjectives requires establishing a context-dependent comparison class by inferring to what aspects of (non-)linguistic context the referent is being compared. This task is relevant for the interpretation of a number of other expressions, as comparison classes are also employed when understanding other kinds of context-sensitive language like quantifiers (e.g., "John ate *many* of hot dogs", Schöller & Franke, 2017) or generics (e.g., "Dogs are friendly" *[relative to other animals]*, Tessler & Goodman, 2019).

This thesis investigated the interpretation of relative gradable adjectives as a case-study of context-sensitive language and attempted to contribute to understanding how exactly listeners might establish relevant comparison classes. Standard literature on gradable adjective semantics suggests that the comparison class might be supplied by the noun the adjective combines with; however, this work argued that purely compositional accounts might be too rigid to explain the flexible use of gradable adjectives across contexts. Furthermore, even when allowing that relevant contextual aspects supply the comparison class, much of the research on gradable adjective semantics eschewed the question how exactly a comparison class might be determined contextually, instead focusing on its integration into compositional semantics (Cresswell, 1976; Kamp, 1975; Kennedy, 2007, 2012; Solt, 2009). Therefore, this work outlined a novel functional hypothesis regarding the role a noun

might have for comparison class inference in addition to contextual cues — namely the *reference-predication trade-off hypothesis*. The quintessential prediction derived from this hypothesis is that listeners are less likely to take the noun as a cue towards the comparison class when the noun can be expained away as intended for the informational goal of reference, operationalized via the syntactic position of the noun the adjective combines with. Specifically, nouns appearing in the subject of the sentence can be explained by their utility in reference and are therefore less likely to set the comparison class, while nouns in the predicate are less likely to be referential and are therefore more likely to constrain the comparison class.

## 6.1    Experiments

Four experiments were conducted to investigate this hypothesis. Experiment 1 showed that participants appreciably disprefer sentences where a subordinate noun appeared in the predicate compared to sentences with a basic-level noun in the predicate as describing a normal-sized referent, indicating that participants preferred sentences with a more felicitous comparison class label in the predicate. Experiment 2 revealed that participants flexibly adjust their noun choices given different syntactic frames in a free-production setting, producing more basic-level nouns in the predicate than in the subject position. Experiment 3 indicated that participants are highly sensitive to the perceptual context of the utterance, to the noun of the utterance and to its syntactic position when asked to paraphrase the comparison class, trading-off the noun's utility in reference and predication when reasoning about its cue strength towards the comparison class. Finally, a pilot study for Experiment 4 showed that reasoning about informational goals accomplished by the noun is indeed the driver behind participants' reasoning about comparison classes, as opposed to mere presence of direct syntactic modification of the noun by the adjective. Together, these experiments provide converging evidence that humans use information structure when reasoning about the comparison class, consistent with the reference-predication trade-off hypothesis.

Studies presented here investigated the role of perceptual context, noun type and its syntactic position on comparison class inferences; while covering major cues towards the comparison class, the particular operationalization of the hypothesis via syntactic manipulation between subject and predicate position of the noun invites investigation of further aspects in future work. For one, all experiments presented critical utterances in written form, yet one other factor that is closely related to information structure is *prosody* (Krifka, 2008). Prosody might provide a strategy to both structure information and construct new content: for instance, prosody

realising focus of a sentence might indicate the presence of relevant alternatives for the focused expression, convey the topic of the utterance or emphasize new information (Krifka, 2008; Selkirk, 1995). Applied to sentences used in present studies, uttering the sentence "That Great Dane is BIG" (BIG being prosodically prominent) might convey that the speaker deems the size of this referent particularly noteworthy and therefore shifts the communicative goal towards predication - a perfectly reasonable scenario where the Great Dane is actually big *for* a Great Dane, making the subordinate subject-noun the comparison class. Alternatively, the same utterance is imaginable in a situation where someone said "That Great Dane is small" and the speaker replied "No, that Great Dane is BIG". The latter example goes in the direction of meta-linguistic or predagogical uses of gradable adjectives where the speaker informs the listener about what he considers a viable use of size adjectives given the particular referent. Uttering "That GREAT DANE is big" might signal that a Great Dane is being contrasted against other categories or referents, highlighting the referential goal of the noun. Prosodically different readings also seem possible for the predicate-N utterance in different contexts. For instance, saying "THAT's a big Great Dane" might contrast a particular referent against other Great Danes; or teach the listener about what the speaker considers a good representative of a big Great Dane. The latter example seems plausible in a context where e.g. other big dogs or Great Danes were previously discussed. Uttering "That's a BIG Great Dane" again seems to imply meta-linguistic or contrastive intentions, while "That's a big GREAT DANE" would highlight the category of the referent, relevant in some way for the current communicative goal(s). Generally, prosody as a tool for both information packaging and content construction might affect the question under discussion and therefore the informational goal in focus, so effects of prosody on comparison class inferences and their connection to the refernence-predication hypothesis should be addressed in future work (cf. Krifka, 2008).

For another, it should be noted that this particular experimental approach was chosen in order to keep a maximally simple and controlled design. So for practical reasons of stimulus presentation only size-adjectives 'big' and 'small' were used; however, the reference-prediction trade-off hypothesis is applicable to relative gradable adjectives in general, so further experiments involving other adjectives should be conducted. Furthermore, the experiments used stimuli from natural basic-level categories only; salient taxonomic representation of such categories might play a role for listeners' reasoning about potential comparison classes, by providing salient viable alternative comparison categories (i.e., subordinate and basic-level categories, cf., Rosch et al. (1976), Tenenbaum et al. (2011)); there might be potentially relevant psychological differences between natural and articifical concept representa-

tions though (Kalish, 2002). Additionally, only one specific (schematic) referent picture per subordinate category was used across experiments; future studies should consider varying the pictures in order to average out potential typicality and nameability effects of the pictures as representations of those categories. Relatively high by-target random intercepts across experiments suggested that targets might have varied in those aspects (e.g., resulting in different propensities of the speakers to use subordinate labels). Running a typicality or nameability rating study on the used categories and pictures might also be useful for eliciting priors for fitting the RSA model quantitatively to observed data (Franke et al., 2016).

Furthermore, in order to be able to create maximally symmetric utterances across different contexts and syntactic conditions, the noun phrase in the predicate position was always indefinite, and the subject noun phrase was always definite. If the predicate was definite, the triggered uniqueness presupposition would have been violated, deeming these sentences generally less natural (i.e., saying "That's *the* big Great Dane" would trigger the expectation that there is only one Great Dane, which is not true of contexts used in the experiments) (cf. Syrett et al., 2010). For this reason, the definiteness of the noun is perfectly confounded with its position throughout the experiments. Yet utterances with a definite predicate noun could be tested in an appropriate discourse context where there is only one individual denoted by the noun. The reference-predication hypothesis would predict that definite nouns in the predicate might still be more likely to establish the comparison class than subject nouns irrespective of their higher referential value compared to indefinite ones, as long as reference was already established in the subject. The hypothesis, however, does not seem to be operationalized by alternative sentences where the subject noun would be indefinite, since indefinite nouns generally do not refer; these sentences might rather interface with generic or meta-linguistic uses of adjectives (Barker, 2002; Reboul, 2001; Tessler & Goodman, 2019).

Generally, referential pressure was not very high in current experiments. Referential purposes of the noun were mainly communicated through its combination with the deictic and its position in the subject, which might rely on general information-structural expectations (s. Chapter 3). But the target referent was perceptually highlighted in a separate picture throughout the experiments. The implied contextual referential pressure might be potentially manipulated in further experiments, e.g., by presenting the referent as a non-highlighted member of the context.

Finally, the interesting relation between the results of Experiment 2 and Experiment 3 shall be noted. Specifically, while the rate of inferred basic-level comparison classes in Experiment 3 practically does not decrease below 0.5 (s. Fig. 4.10), the rate of produced basic-level nouns in Experiment 2 stagnates at around 0.5 (s. Fig. 4.7). This indicates a contrast between a potential basic-level bias in the inference task

and speakers' inclination to to use referentially more informative subordinate nouns in the production task. On the one hand, these results are not directly comparable in that Experiment 2 provides rather indirect evidence for the reference-predication hypothesis, compared to the direct inference Experiment 3. On the other hand, the production Experiment 2 might have implied higher referential pressure on the participants by presenting a production task for a target in a basic-level context. Therefore, these results point to more general issues of how different informational goals might interact and differ when considering speaker versus listener-centric scenarios, and which factors contribute to their trade-off or potential synergy (cf. Heller et al., 2008).

## 6.2   RSA Model

Besides experimental evidence, this work also provides a computational model of the reference-predication hypothesis within the Rational Speech Act framework. The proposed model builds upon and advances beyond previous RSA-models of gradable adjectives by incorporating simultaneous reasoning about general knowledge, lexical and syntactic cues towards the comparison class. This pragmatic listener model infers a potential state of the world (a referent and its size) along with the likely comparison class. In particular, the listener reasons about how a speaker would behave in order to communicate a specific state, also referring to their prior knowledge about likely sizes of different categories. Crucially, the listener assumes that the speaker might variably choose the intended adjective meaning, and therefore the comparison class. In addition, the speaker constructs her utterances incrementally by deciding whether to put the noun in the utterance subject or in the predicate, so as to optimally achieve reference with the subject, and predication with the predicate. This representation allows the speaker to achieve two informational goals simulatneously by using one sentence. By using hypothetic fixed priors, qualitative predictions were derived from the model. It was shown that it is a good first formalization of the reference-predication trade-off hypothesis because it captures essential contrasts observed in Experiment 3. Specifically, the pragmatic listener was more likely to take the noun as a cue towards the comparison class when it appeared in the predicate than in the subject, while taking into account referential utility of the noun. In sum, the proposed model provides a first attempt to integrate pragmatic reasoning about both syntax and semantics of an utterance within the Rational Speech Act framework.

However, the predictions derived from the proposed model were only qualitative. Future work should evaluate its potential to fit the data observed in Experiment

3 quantitatively. To do so, experiments eliciting participants' prior knowledge of relative category sizes could be conducted (Franke et al., 2016), or the priors could be reconstructed from indirect experimental results relying on the same prior knowledge (following Tessler et al., 2017). The free speaker optimality paramter should then also be determined more accurately, e.g., via a maximum a posteriori estimate.

Furthermore, additional assumptions were made for this model: for one, inference over the standard of comparison value used for computing literal semantics of the adjective was performed by the literal listener. The cognitive plausibility of this adjective meaning representation remains to be investigated. In addition, it was assumed that the listener and the speaker share identical world knowledge about likely sizes for different categories. Incorporating their reasoning about each other's world knowledge in future work would potentially extend this model to account for meta-linguistic uses of gradable adjectives (s. Chapter 2; Barker (2002)). Lastly, the model utilized uniform priors over potential referents in context, as well as over potential comparison classes. These assumptions should be revised because referents might differ in their perceptual saliency — e.g., referents whose size stands out might be generally more likely to be talked about; and comparison class avaliability might vary across different subordinate categories due to namability or typicality effects. One possibility would be to consider corpus frequencies of respective nouns as an approximation of the prior (following the model by Tessler et al., 2017), or gather speaker production data.

More generally, the relation of this formalization to alternative possible models potentially explaining observed data should be addressed in future work. For instance, models of incremental processing might explain general interpretative differences between the subject and the predicate, because the speakers might build their utterance in a way such that sufficient contrastive information is provided at each point in the utterance. Since the adjective generally appeared in the predicate of sentences used in Experiment 3, the subject might be expected to feature referentially more informative nouns than the predicate, because the adjective can only be used for potential target disambiguation later in subject-N sentences. However, such a model might not account for the results observed in Experiment 4.

In sum, the proposed lexical enrichment-based model provides the first qualitative example of context-sensitive language interpretation based on unified reasoning about several cues to the intended meaning. The hope is to supply a basis for future investigations of such holistic RSA models applied to various cases of vague expressions, in addition to providing a more sophisticated model of gradable adjective understanding.

## 6.3 Developmental Perspective

Cues that contribute to gradable adjective interpretation also have implications for investigating the developmental course of children's understanding of complex context-sensitive expressions. In particular, already by the age of two infants understand adjectives like 'big' and 'small' and appreciate their context-sensitive nature, sometimes even without adults specifically pointing out the relevant features of context (Ebeling & Gelman, 1994; Mintz & Gleitman, 2002). That is, the critical skill to acquire for gradable adjective interpretation is establishing the correct comparison class; while adults often use prepositional *for*-phrases in child-directed speech, such phrases are often indicative of a particular kind of gradable adjective uses - the functional uses (s. Chapter 2, Ebeling and Gelman (1994)). Yet when not producing a *for*-phrase, adults might include other syntactic cues towards the comparison class compensating for more ambiguous adjective uses.

Taking a developmental perspective on the reference-predication hypothesis, one might derive the prediction that adults use nouns in the predicate of sentences more often than in the subject in order to restrict the comparison class more strongly and make the adjective interpretation easier for children. A study was conducted wherein the Providence corpus was annotated with respect to linguistic and environmental cues available to interlocutors using the adjective "big" (Sinelnikova, 2020). It provides preliminary evidence that the majority of recorded uses of the adjective "big" were indeed prenominal cases where the modified nominal appears in the predicate of the sentence. Moreover, it showed that more modified nouns appeared in the predicate when there were less distractors in the context; when there were more perceptual distractors the modified noun was rather placed in the subject, contributing to reference. Furthermore, it was found that the referent was mostly physically copresent when the noun was used in the predicate, suggesting that reference might have been established by means other than the noun (e.g., pointing). Finally, the intended comparison class for the adjective-nominal predicate constructions was overwhelmingly the normative comparison class (i.e., making reference to implicit general knowledge of the abstract basic-level category of the referent, mostly denoted by the noun; s. Chapter 2), indicating that these syntactic frames were used when establishing the comparison class might have been more challenging for the children (Sinelnikova, 2020). These preliminary results are consistent with predictions of the reference-predication trade-off hypothesis — nouns in the predicate seem to communicate the comparison class of a referent, facilitating understanding of the gradable adjective for kids who might lack other means for pragmatic reasoning like substantial world knowledge. Understanding what kinds of cues are available to children in naturalistic settings might provide a starting point

for investigating how they succeed in learning context-sensitive expressions.

## 6.4 Conclusion

To sum up, this thesis took a step towards understanding how interlocutors flexibly use relative gradable adjectives across contexts, by presenting a novel reference-predication trade-off hypothesis of comparison class inference. This hypothesis provides a holistic account of how humans might integrate various cues and reason about referential utility of the noun in context, trading it off with the noun's utility in communicating the comparison class. This inferential account is supported by converging evidence from various experiments. Finally, the hypothesis was formalized computationally within the Rational Speech Act framework, utilizing domain-general Bayesian inference tools widely used in models of cognition. It emphasized the sophisticated reasoning at the interface of syntax, semantics and pragmatics that humans perform in order to make clear use vague language. Through the lens of reference and predication, this work highlighted a topic of broad interest to cognitive science — namely the complex interaction between intentions and informational goals in communication, and the factors shaping this relationship.

# Declaration of Authorship

I hereby certify that the work presented here is, to the best of my knowledge and belief, original and the result of my own investigations, except as acknowledged, and has not been submitted, either in part or whole, for a degree at this or any other university.

Signature: _____

City, date: _____

# Appendix A

# Appendix

## A.1 Experimental Materials

### A.1.1 Bot-Check Trial

The names used in the bot-check trials were:

- Male names: James, John, Robert, Michael, William, David, Richard, Joseph, Thomas, Charles

- Female names: Mary, Patricia, Jennifer, Linda, Elizabeth, Barbara, Susan, Jessica, Sarah, Margaret.

This trial view was developed and provided by Elisa Kreiss.

### A.1.2 E1 Exclusion Criteria

In the Sentence Rating Experiment (E2), data from 33 participants was excluded. 3 participants indicated a native language other than English. Data from 3 subjects was excluded due to failed warm-up trials. This means, participants provided a lower rating of the sentence "The chair is blue" than the sentence "The chair is yellow" on the chair warm-up trial; it was also counted as a fail if participants rated the sentence "The basketball is green" higher than the sentence "The basketball is orange", or if the rating of the sentence "The basketball is orange" received a rating of less than 50 on the basketball trial.

Furthermore, data from 27 participants were excluded who provided the same ratings within 5 points for one syntactic condition on every trial (one of the sentences on every trial), or those who provided the same ratings of the two sentences on every trial. However, choosing exclusion criteria based on participants' performance in the main trials might have been an overly restrictive or biasing criterion. So an exploratory analysis was conducted on the full dataset, where participants were

only excluded based on their performance in the practice trials. This exploratory analysis revealed results qualitatively and quantitatively very similar to results from the main preregistered analysis reported in the main work: participants dispreferred sentences with a subordinate predicate noun, compared to sentences with basic-level subordinate nouns, but did not show any preferences in the subject-noun condition (syntax-by-noun interaction: $\beta = -3.07[-4.46, -1.72]$). They also overall preferred the subject-N syntax ($\beta = 1.85[0.07, 3.67]$), as well as basic-level nouns ($\beta = 5.43[3.13, 7.71]$).

## A.2    Refpred-RSA Model Alternatives

The main proposed reference-predication RSA model assumes a speaker utility function wherein the speaker chooses an utterance such that it optimally communicates a two-dimensional state of the world (i.e., the referent and its size):

$$U_{S_1}(u; r; s; cc; C) = log\ L_0(s, r \mid u, C, cc)$$

However, based on previous models addressing multiple potential questions under discussion, there are other conceivable representations of the speaker utility (cf. Kao et al., 2014; Yoon et al., 2016). For instance, another more complex option is to consider a QUD-based model wherein the speaker chooses utterances specifically communicating reference or predication only, or both goals, inspired by Kao et al. (2014). In that case, the speaker-utility is defined with respect to a certain QUD she has in mind:

$$U_{S_1}(u; r; s; cc; C; QUD) = log\ L_0(QUD(s, r) \mid u, C, cc)$$

where $QUD(.)$ projects the state of the world onto the subspace relevant for the particular question under discussion. That is, if the QUD is reference, the literal listener returns a distribution over referents, marginalizing over possible properties. The opposite happens when the QUD is predication. When the QUD is to communicate both aspects, $L_0$ returns a two-dimensional distribution over referents and properties. The latter possibility is equivalent to the main proposed model (Eq. 5.19). The $L_1$ would then jointly reasons about the intended meaning and the intended QUD:

$$P_{L_1}(r, s, cc \mid u, C) \propto \sum_{QUD} P_{S_1}(u, cc \mid s, r, C, QUD)\ P(r, s \mid C, cc_{r\_sub})\ P(QUD)$$

Yet another possibility is to follow the idea proposed by Yoon et al. (2016),

representing the speaker as trying to simulataneously achieve a combination of the two informational goals (reference and predication) weighted by a free parameter $\phi$ representing referential weight of the sentence:

$$U_{S_1}(u; r; s; cc; C; \phi) = log(\phi\, L_0(r \mid C) + (1 - \phi)\, L_0(s \mid cc))$$

To compute the utility for one goal, the distribution returned by $L_0$ is marginalized over the other aspect of the state of the world (equivalently to the QUD-based model). The pragmatic listener then also reasons about the value of $\phi$:

$$P_{L_1}(r, s, cc, \phi \mid u, C) \propto P_{S_1}(u, cc \mid s, r, C, \phi)\, P(r, s \mid C, cc_{r\_sub})\, P(\phi)$$

These are also conceivable representations of reasoning about informational goals. Although the speaker utility proposed in the main model in Section 5.3.2 posits a priori that both goals of reference and predication are relevant for any situation, it has clear advantages. First, this model has the simplest, most intuitive structure, avoiding the use of additional layers of inference or free parameters. Second, it implements the reference-predication trade-off hypothesis in a more elegant way: It explains the qualitative inference pattern observed empirically simply based on explaining away alternative states of the world and potential utterances they could have been described by, especially by shifting the noun position, avoiding architectures wherein the speaker would explicitly be biased to use the noun in a specific position to convey a particular informational goal. Positing that interlocutors pursue both informational goals also seems to be a reasonable assumption, given the fundamental nature of reference necessarily underlying predication. Therefore, it is argued that the main proposed model is a more suitable reference-predication hypothesis formalization, but further possible speaker utility representations are left to future research.

# Bibliography

Aparicio, H., Xiang, M., & Kennedy, C. (2016). Processing gradable adjectives in context: A visual world study, In *Semantics and linguistic theory*.

Bale, A. C. (2011). Scales and comparison classes. *Natural Language Semantics*, *19*, 169–190.

Barker, C. (2002). The dynamics of vagueness. *Linguistics and philosophy*, 1–36.

Barner, D., & Snedeker, J. (2008). Compositionality and statistics in adjective acquisition: 4-year-olds interpret tall and short based on the size distributions of novel noun referents. *Child development*, *79*(3), 594–608.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3). https://doi.org/10.1016/j.jml.2012.11.001

Bartsch, R., & Vennemann, T. (1972). Semantic structures: A study in the relation between semantics and syntax. *Athenäum-Skripten Linguistik Bd*, *9*.

Bergey, C., Morris, B. C., & Yurovsky, D. (2020). Children hear more about what is atypical than what is typical, In *Proceedings of the 42nd annual meeting of the cognitive science society*.

Bierwisch, M. (1989). The semantics of gradation. *Dimensional adjectives*, *71*(261), 35.

Bolinger, D. (1967). Adjectives in english: Attribution and predication. *Lingua*, *18*, 1–34.

Brown, P., & Levinson, S. C. (1987). *Politeness: Some universals in language usage* (Vol. 4). Cambridge university press.

Bürkner, P.-C. (2017). Advanced bayesian multilevel modeling with the r package brms. *arXiv preprint arXiv:1705.11123*.

Chafe, W. (1976). Givenness, contrastiveness, definiteness, subjects, topics, and point of view. *Subject and topic*.

Chierchia, G., Fox, D., & Spector, B. (2012). The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Semantics: An international handbook of natural language meaning*, *3*, 2297–2332.

Cinque, G. (2010). *The syntax of adjectives: A comparative study* (Vol. 57). MIT press.

Clifton Jr, C., & Ferreira, F. (1989). Ambiguity in context. *Language and cognitive processes*, *4*(3-4), 77–103.

Cresswell, M. J. (1976). The semantics of degree. In *Montague grammar* (pp. 261–292). Elsevier.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A bayesian approach to "overinformative" referring expressions. *Psychological Review*.

Donnellan, K. S. (1966). Reference and definite descriptions. *The philosophical review*, 281–304.

Ebeling, K. S., & Gelman, S. A. (1994). Children's use of context in interpreting "big" and "little". *Child Development*, *65*(4), 1178–1192.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Franke, M., Dablander, F., Schöller, A., Bennett, E., Degen, J., Tessler, M. H., Kao, J. T., & Goodman, N. D. (2016). What does the crowd believe? a hierarchical approach to estimating subjective beliefs from empirical data., In *Proceedings of the 38th annual meeting of the cognitive science society*.

Goldberg, A. E., & Michaelis, L. A. (2017). One among many: Anaphoric one and its relationship with numeral one. *Cognitive Science*, *41*, 233–258.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2014). The Design and Implementation of Probabilistic Programming Languages [Accessed: 2020-9-15].

Graf, C., Degen, J., Hawkins, R. X., & Goodman, N. D. (2016). Animal, dog, or dalmatian? level of abstraction in nominal referring expressions., In *Proceedings of the 38th annual meeting of the cognitive science society*.

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Heim, I. (2000). Degree operators and scope, In *Semantics and linguistic theory*.

Heller, D., Grodner, D., & Tanenhaus, M. K. (2008). The role of perspective in identifying domains of reference. *Cognition*, *108*(3), 831–836.

Hofherr, P., & Matushansky, O. (2010). *Adjectives: Formal analyses in syntax and semantics*. John Benjamins Publishing Company.

Ilieva, S., Ji, X., Rautenstrauch, J., & Franke, M. (2018). Minimal architecture for the generation of portable interactive experiments [Accessed: 2020-09-01]. https://magpie-ea.github.io/magpie-site/

Kaiser, E., & Wang, C. (2020). Distinguishing fact from opinion: Effects of linguistic packaging, In *Proceedings of the 42nd annual meeting of the cognitive science society*.

Kalish, C. (2002). Gold, jade, and emeruby: The value of naturalness for theories of concepts and categories. *Journal of Theoretical and Philosophical Psychology*, *22*(1), 45.

Kamp, J. A. W. (1975). Two theories about adjectives. In E. L. Keenan (Ed.), *Formal semantics of natural language*. Cambridge University Press, Cambridge, England.

Kao, J. T., Wu, J. Y., Bergen, L., & Goodman, N. D. (2014). Nonliteral understanding of number words. *Proceedings of the National Academy of Sciences*, *111*(33), 12002–12007.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and philosophy*, *30*(1), 1–45.

Kennedy, C. (2012). Adjectives. In D. Fara & G. Russell (Eds.), *The routledge companion to philosophy of language*. Routledge. https://books.google.de/books?id=cV9LvekiKKAC

Klein, E. (1980). A semantics for positive and comparative delition. *Linguistics and Philosophy*, *4*(1), 1–46.

Kreiss, E., & Degen, J. (2020). Production expectations modulate contrastive inference, In *Proceedings of the 42nd annual meeting of the cognitive science society*.

Krifka, M. (2008). Basic notions of information structure. *Acta Linguistica Hungarica*, *55*(3-4), 243–276.

Kruschke, J. (2014). *Doing bayesian data analysis: A tutorial with r, jags, and stan*. Academic Press.

Kurz, S. (2019). Bayesian power analysis: Part iii.b. what about 0/1 data? [Accessed: 2020-08-01]. https://solomonkurz.netlify.app/post/bayesian-power-analysis-part-iii-b/

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives, In *Semantics and linguistic theory*.

Lassiter, D., & Goodman, N. D. (2017). Adjectival vagueness in a bayesian model of interpretation. *Synthese*, *194*(10), 3801–3836.

McNally, L., & Kennedy, C. (2008). *Adjectives and adverbs: Syntax, semantics, and discourse*. Oxford University Press.

Mintz, T. H., & Gleitman, L. R. (2002). Adjectives really do modify nouns: The incremental and restricted nature of early adjective acquisition. *Cognition*, *84*(3), 267–293. https://doi.org/10.1016/S0010-0277(02)00047-1

Montague, R. (1973). The proper treatment of quantification in ordinary english. In *Approaches to natural language* (pp. 221–242). Springer.

Qing, C., & Franke, M. (2014). Gradable adjectives, vagueness, and optimal language use: A speaker-oriented model, In *Semantics and linguistic theory*.

Reboul, A. (2001). Foundations of reference and predication. In M. Haspelmath (Ed.), *Language typology and language universals. an international handbook, vol.1*. Walter de Gruyter.

Roberts, C. (2012). Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, *5*, 6–1.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Schöller, A., & Franke, M. (2017). Semantic values as latent parameters: Testing a fixed threshold hypothesis for cardinal readings of few & many. *Linguistics Vanguard*, *3*(1).

Scontras, G., Tessler, M. H., & Franke, M. (2018). Probabilistic language understanding: An introduction to the rational speech act framework. [Accessed: 2020-09-15]. https://www.problang.org

Scontras, G., Degen, J., & Goodman, N. D. (2017). Subjectivity predicts adjective ordering preferences. *Open Mind*, *1*(1), 53–66.

Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge university press.

Sedivy, J. C., Tanenhaus, M. K., Chambers, C. G., & Carlson, G. N. (1999). Achieving incremental semantic interpretation through contextual representation. *Cognition*, *71*(2), 109–147. https://doi.org/10.1016/s0010-0277(99)00025-6

Selkirk, E. (1995). Sentence prosody: Intonation, stress, and phrasing. *The handbook of phonological theory*, *1*, 550–569.

Sera, M., & Smith, L. B. (1987). Big and little:"nominal" and relative uses. *Cognitive Development*, *2*(2), 89–111.

Sinelnikova, A. (2020). *Cues to comparison classes in child-directed language* (M. Eng. Thesis). Massachusetts Institute of Technology.

Solt, S. (2009). Notes on the Comparison Class, In *International workshop on vagueness in communication*.

Stechow, A. v. (1984). Comparing semantic theories of comparison. *Journal of Semantics*, *3*(1-2), 1–77. https://doi.org/10.1093/jos/3.1-2.1

Steedman, M., & Altmann, G. (1989). Ambiguity in context: A reply. *Language and cognitive processes*, *4*(3-4), 105–122.

Syrett, K., Kennedy, C., & Lidz, J. (2010). Meaning and context in children's understanding of gradable adjectives. *Journal of semantics*, *27*(1), 1–35.

Team, R. C. et al. (2013). R: A language and environment for statistical computing. Vienna, Austria.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Tessler, M. H., Tsvilodub, P., & Levy, R. (2020). Inferring Comparison Classes from Sentence Structure and Informational Goals (T. von der Malsburg, S. Vasishth, & I. Wartenburger, Eds.). In T. von der Malsburg, S. Vasishth, & I. Wartenburger (Eds.), *Proceedings of the 26th architectures and mechanisms for language processing conference (AMLaP)*, Potsdam, Germany, Universität Potsdam.

Tessler, M. H., & Goodman, N. D. (2019). The language of generalization. *Psychological Review, 126*(3), 395.

Tessler, M. H., Lopez-Brau, M., & Goodman, N. D. (2017). Warm (for winter): Comparison class understanding in vague language, In *Proceedings of the 39th annual meeting of the cognitive science society*.

Tessler, M. H., Tsvilodub, P., Snedeker, J., & Levy, R. P. (2020). Informational goals, sentence structure, and comparison class inference, In *Proceedings of the 42th annual meeting of the cognitive science society*.

Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2016). Talking with tact: Polite language as a balance between kindness and informativity, In *Proceedings of the 38th annual conference of the cognitive science society*. Cognitive Science Society.

Zalta, E. N. (Ed.). (2017). *Indexicals. the stanford encyclopedia of philosophy (summer 2017 edition)*. https://plato.stanford.edu/archives/sum2017/entries/indexicals/

Zalta, E. N. (Ed.). (2019). *Reference. the stanford encyclopedia of philosophy (spring 2019 edition)*. https://plato.stanford.edu/archives/spr2019/entries/reference/