Inferring Comparison Classes of Gradable Adjectives

The Role of Informational Goals and Sentence Structure

By

Polina Tsvilodub

Submitted in partial fulfillment of the requirements for the degree of
Bachelor of Science in Cognitive Science
to the
Institute of Cognitive Science at the Osnabrück University
September, X 2020

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

[pt: The abstract will be updated when all results are in]

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

I want to thank...

Contents

List of Figures

List of Tables

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 plausible 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 is 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 *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 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 θ. This means, relative gradable adjectives convey a feature, like size, and the degree to which the referent possesses this feature must exceed some threshold θ for the referent to be felicitously described by the respective gradable adjective (e.g., Kennedy, 2007). At the same time this threshold θ 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 θ 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 θ, 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 the 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 θ 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 presented.

2.1 Semantic Representation of Gradable Adjectives

Currently standard theories of gradable adjectives converge in 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 positive forms 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 specifying 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 θ. 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 **??**. Formally, *pos* denotes the following:

$$[\![_{Deg} pos \lambda g \lambda x.g(x)]\!] = \lambda g.\lambda x : g(x) > s(\lambda x : g(x))$$

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 approach 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 outlined examples 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 θ 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 is how exactly the standard of comparison - the threshold θ - is determined by a given comparison class. For instance, Cresswell (1976) suggested that the threshold θ 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 comprising 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):

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

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 (Bale, 2011; Solt, 2009, cf.). 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:

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

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. However, in cases like "John is tall for a gymnast", overt *for*-phrases may provide this argument, but for cases like "Sara reads difficult books for an 8-year-old" Solt (2009) assumed 'books' to be the basis of the comparison class rather incidentally, focusing on the representation of the presupposition triggered by the *for*-phrase.

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 restricts the quasi-order before formation of the scale, restricting 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' (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, e.g., 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 *norma-*

tively, 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. 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). 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 Semantic and Syntactic 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 contextual 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, 'John is a rich Fortune 500 CEO' doesn't mean that he is *rich for a Fortune 500 CEO*; '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', cf., Goldberg and Michaelis (2017); however, the most reasonable resolution of the anaphora would stipulate referring to the subject noun 'dog', reducing this idea to the former one). 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 et al., 2020). Yet, similar intuitive counter-examples might be put forward here. 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.

Discuss Bolinger 1967

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 syntactic modification (i.e., in prenominal adjectives) the modified noun would set the comparison class akin to the simple syntactic account outlined above, while absence of modification (i.e., in predicative adjectives) would signal that the noun is *not* the comparison class. However, this alternative would not resolve remarks made against the compositional account, and 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 target 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 triangles - 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 prenominal attributive adjectives give rise to ambiguity between the two 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 for a Fortune 500 CEO, but not rich in general", which intuitively isn't correct). However, positing intersective

readings implies the existence of some set of things that count as e.g. generally rich - a stipulation rather difficult to capture. Considering vague scalar adjectives as subsective seems to require direct modification, which would require additional ad hoc mechanisms for interpreting the same adjectives occuring predicatively. 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 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, conrastive 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 comaprison 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 object 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); both conditions included a competitor object of a different category which could be felicitously described by the adjective, hearing instructions like in the first experiment. 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 set-up. [1] In another experiment where questions involving an indefinite noun were used

[1] 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

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 similar visual world study, 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 fo 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

the latter is actually referrring.

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. [2] 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 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. 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' expectation about such categories highly salient with respect to some feature. 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

[2] 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'

a comprehensive theory of gradable adjective interpretation.

Chapter 3

A Fucntional Perspective on Comprison 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 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 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.X); 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 to speakers have the liberty to choose from truth-conditionally similar sentence options to communicate the same message. For example, in order to tell some 4-year-old kids on a playground in winter that they built a big snowman, a speaker has the liberty to say "That's big!" pointing at the snowman, "That snowman you built is big!" or "You built a big snowman!", 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 the deictic 'that'), and hence the noun is a weaker cue towards the comparison class.

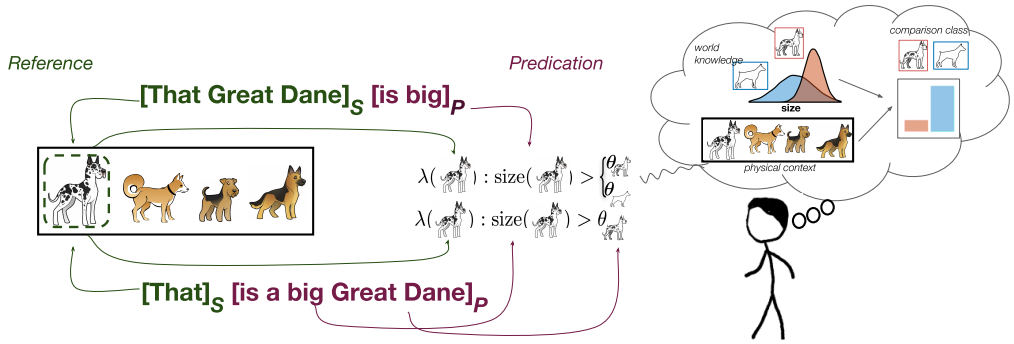

Figure 3.1: Cartoon of the inferential account for comparison class determination. The noun (Great Dane) in a sentence can be employed either for the goal of reference (green) or predication (purple), shown in the case when this distinction is made via the syntactic position of the noun (subject S vs. 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 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* ϕ triggers two presuppositions: the *existence* presupposition (i.e.,

that there is an object satisfying the description ϕ), 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* ϕ (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 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, described 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 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:

> *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 in 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 experiments described in section 2.4. [pt: discuss in chapter 6 that deconfounding definiteness from syntactic manipulation should be addresses in future research; keep it maximally symmetric in E1-3; tentative predictions for E4: same distinction for "A prize-winner is a big great dane" vs "A big great dane is a prize-winner"; infelicitous presuppositions for both parts being definite; also discuss connection to plural / generics / predagogical language;]

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, this noun is rather expected to be consistent with the comparison class felicitous in order to describe a target: 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.

Operationalisation of referential utility following reference-RSA.

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 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 rea-

sons: 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. This experimental data provides a comprehensive overview of pragmatic and syntactic effects on comparsion class inference.

Chapter 4

Experiments

The reference-predication trade-off hypothesis was tested in four preregistered behavioural web-based experiments employing different dependent measures. 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 N is N2") or in the predicate ("That's a ADJ N" or "That N2 is a ADJ N") 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 4.1). 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 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 is more felicitous in

Table 4.1: 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*

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 most 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

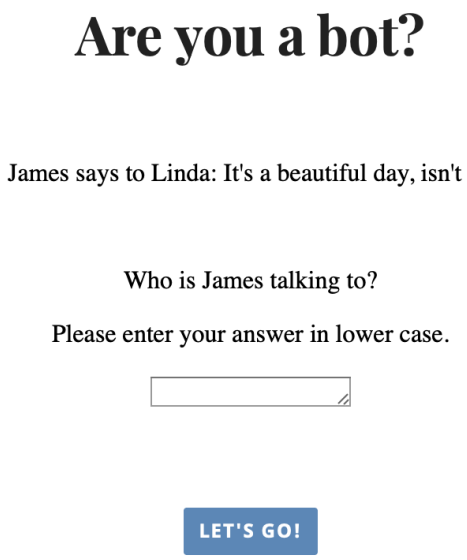

Figure 4.1: Example view of the bot check trial: The speaker James addresses the listener Linda.

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., n.d.). All experiments and preregistrations can be viewed under `tinyurl.com/yb5ogj5g`. [pt: preliminary link]

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, depending on the noun category. The type of the noun and its syntactic position differed within-subjects.

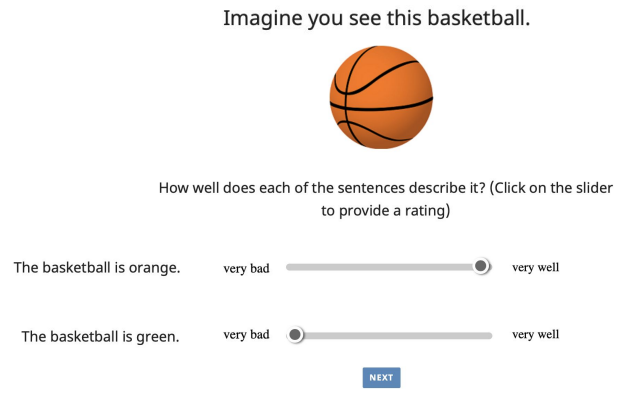

Figure 4.2: Example view of the sentence rating warm-up trial wherein participants rated sentences about a depicted basketball. [pt: Screenshots will be made more readable]

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

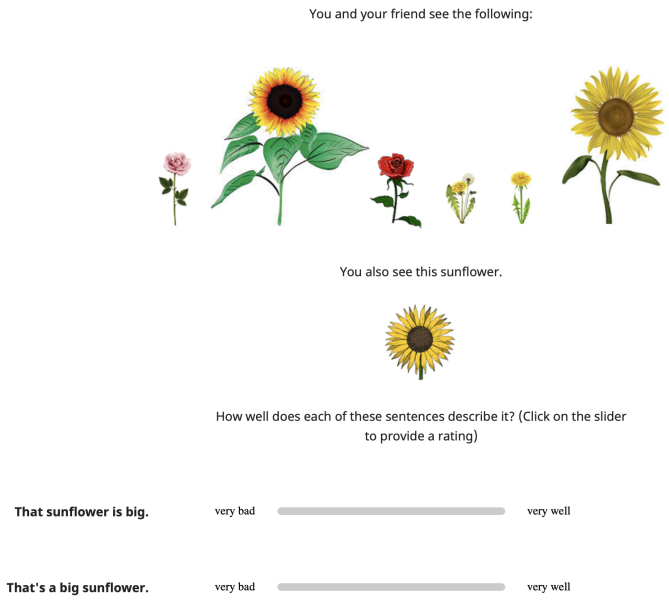

Figure 4.3: Example view of a sentence rating main trial: The critical noun is a subordinate target label of a large-subordinate category, appearing in the subject or predicate of the sentence.

nate category of the referent. The pictures depicted referents a little smaller than 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 4.1).

The reference-predication trade-off hypothesis predicts that sentences with a basic-level noun in the predicate position should receive a higher rating than sentences with a subordinate noun in predicate position, but there should be no difference in the ratings of sentences with a noun in the subject position.

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, the 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 `brms`-pakcage for Bayesian 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 effects of syntax, noun type and their interaction. [1] Both predictors were deviation coded, coding both the subject-noun and the basic-level noun as 1 and the other levels as -1, respectively. 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 X noun-type interaction ($\beta = -4.01[-5.84, -2.18]$) (Fig. 4.4). [2] 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: $\beta = 9.53[5.76, 15.73]$). As expected, an exploratory analysis including a predictor of target size (small-subordinate vs. large-subordinate category) did not reveal any size-effects. Finally, a relatively high by-participant variation indicated differences in overall rating preferences (by-subject intercept: $\beta = 11.96[9.80, 14.52]$).

To sum up, the sentence rating experiment showed that participants are sensitive to the position and the type of the noun, dispreferring sentences where a noun that

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

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

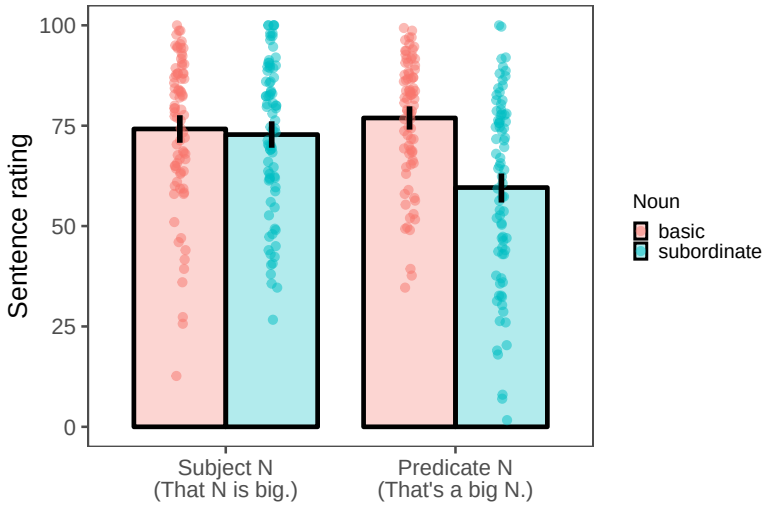

Figure 4.4: Experiment 1 results: Mean ratings for how well sentences which differed in the syntactic position of the noun (x-axis) and the noun-label (color) described a typically-sized referent (e.g., a Great Dane) in basic-level context. Points represent participant means within condition. Error-bars denote bootstrapped 95% confidence intervals (bootstrapping independent of random-effects structure)

provided an infelicitous comparison 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 (e.g., in "That _ is {big,small}") or in the predicate position (e.g. 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 4.1 marked with *. For each

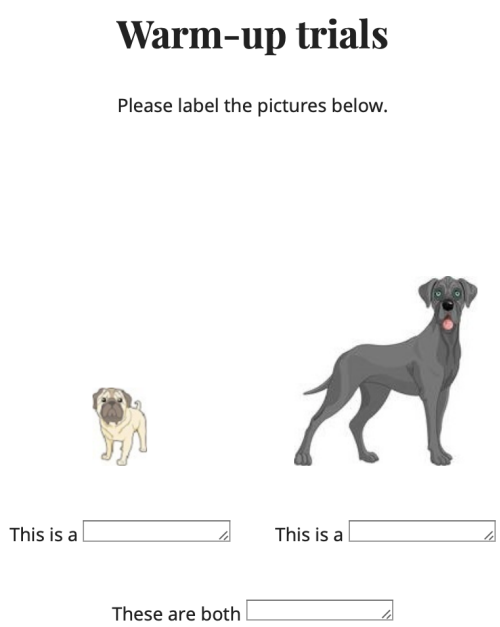

Figure 4.5: Example view of the noun production warm-up trial: Participants have to label a large-subordinate (Great Dane, right) and a small-subordinate target (pug, left) for the dogs-category.

participant, six out of ten possible contexts were sampled. Three of these contexts 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 acomplishment 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.

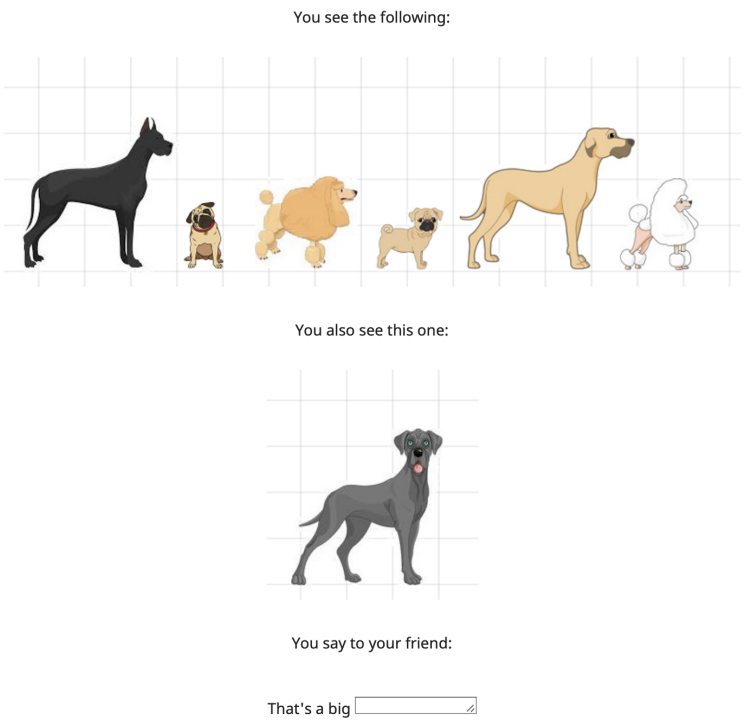

Figure 4.6: Example view of the noun production main trial: Participants fill-in the noun in the predicate position of a sentence describing a large-subordinate target.

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.

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 (see Appendix A for all responses). 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

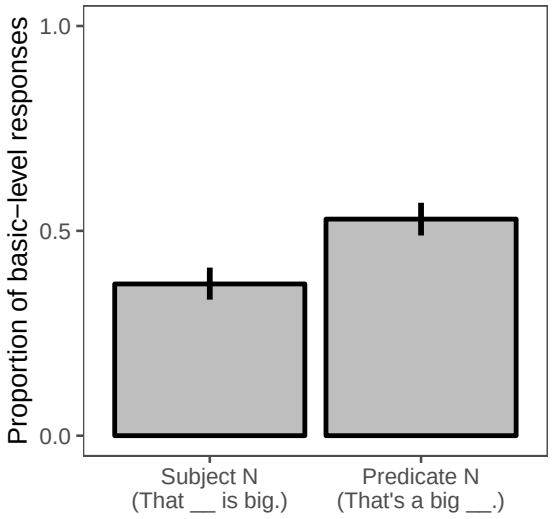

Figure 4.7: Experiment 2 results: Proportions of freely-produced basic-level labels (e.g., *dog*) in different syntactic frames (x-axis) when the referent was a typically-sized member of a subordinate category (e.g., a normal-sized Great Dane). Error-bars denote 95% bootstrapped confidence intervals.

random by-referent slope effects of syntax. [3] The predictor was deviation-coded, coding predicate-N syntax as 0.5 and subject-N syntax as -0.5.

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 predicative 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 random effects 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 in order to communicate a felicitous comparison class, when presented with a free-production task.

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

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. subordinte vs. 'one') and the visual context of the sentence (basic-level vs. subordinate context). All three factors were manipulated within-subjects.

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 sensitive to linguistic cues like the sentence structure. According to the outlined 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 inference can be driven by other pragmatic inference, 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, comapred to subordinate context. In contrast, if comparison class inference was completely driven by the noun of the sentence or other syntactic or semantic properties discussed in chapter 2, no interpretative differences should be observed when the same sentences occur in different perceptual contexts. Another potential mechanism might be that perceptual context only supplies the comparison class, in which case one would expect identical inferences drawn from sentences involving different nouns.

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 the possible options among {buildings, skyscrapers, houses, constructions}. Then,

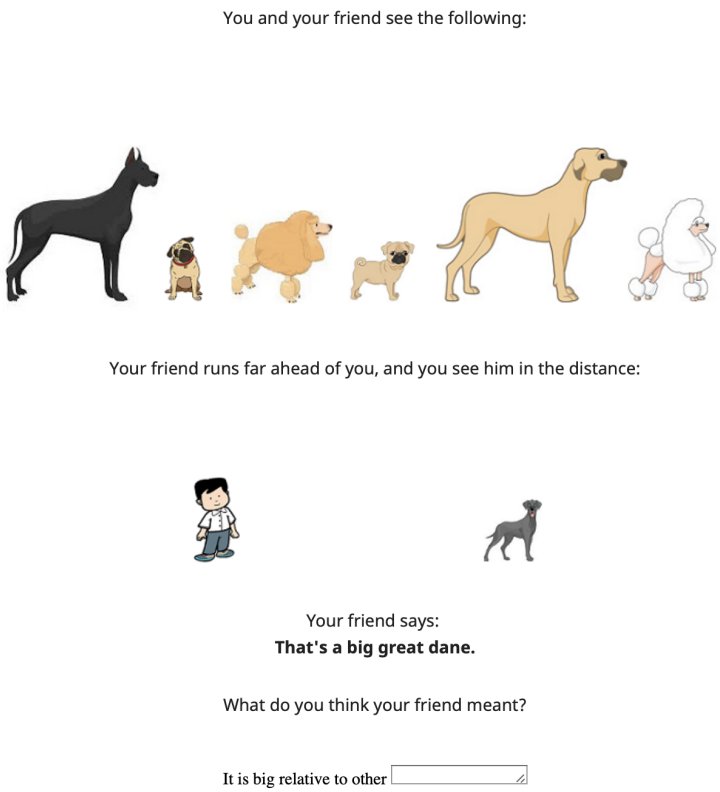

Figure 4.8: Example view of a comparison class inference main trial: Participants paraphrased the critical utterance with a subordinate noun in predicate position, which appeared in basic-level context, describing a large-subordinate target.

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 4.1). These labeling warm-up trials are of the same kind as in Experiment 2 (Fig. 4.5).

In this experiment, 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. [pt: add example picture] 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 flowers, the sunflower appeared in basic-level flower context), and the other possible referent appeared in the corresponding subordinate context (i.e., then the daisy appeared in subordinate daisies-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. [4] 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. 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.

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 took about 9 minutes and participants were compensated $1.20.

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 (s. Appendix A). A Bayesian logistic mixed-effects regression model was used, regressing the response category against the syntactic condition (subject-N vs.

[4] Due to my coding mistake, the conditions were balanced at the level of individual factors.

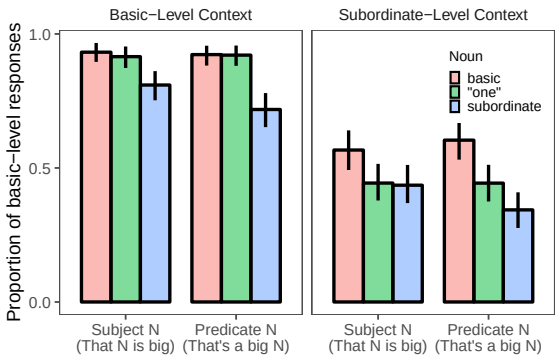

Figure 4.9: Experiment 3 results: Proportions of inferred comparison classes in terms of basic-level responses (e.g., "...big relative to other dogs"), depending on syntactic position of the noun (x-axis), noun-label (color), and context (facets). Context strongly modulated the comparison class (left 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.

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. [5] The predictors were sum-coded: predicate-N and basic-level context as 1, subject-N and subordinate context as -1, the basic-level and the subordinate noun-levels were coded against the baseline anaphoric 'one'. [pt: Include algebra in the appendix?]

The results indicate that participants flexibly adjust the inferred comparison class according to many factors. First and foremost, a large effect of visual context going above and beyond other factors was found ($\beta = 1.88[1.49, 2.31]$; Fig. 4.9, left vs. right facets), as indicated by the inferences drawn from the baseline condition anaphoric 'one' ($\beta = 0.37[0.10, 0.64]$; Fig. 4.9, green bars in the left vs. right facet): participants were appreciably more likely to infer basic-level comaprison classes given a basic-level context, compared to subordinate context. Furthermore, an effect of 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 ($\beta = 2.01[1.37, 2.71]$). The noun-effects can be observed on top of the baseline provided by the visual context: participants inferred more basic-level comparison classes from basic-level nouns that from 'one', and less from subordinate noun that from 'one' (basic-level vs. 'one': $\beta = 0.60[-0.47, 1.70]$) and subordinate vs. 'one' ($\beta = -1.40[-2.17, -0.66]$). Notably, the subordinate comparison class was the minority response even given a

[5] In brm-style syntax: `response_category ~ syntax*NP*context + (1 + syntax + NP + context || subject) + (1 + syntax*NP*context || target)`

subordinate noun in the basic-level context, speaking against a syntactic view of adjective comparison classes, wherein the noun would always set the comaprison class (Fig. 4.9; blue bars, left facet).

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 (red vs. blue bars X x-axis; $\beta = 0.47[0.02, 0.95]$). Examining the interaction by-noun, suggestive evidence was found that the interaction is primarily driven by the critical condition of interest involving the subordinate noun: a 90.0% probability that the subordinate-N vs. 'one' x 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' X Syntax interaction being greater than 0 (i.e., more basic-level comparison classes were 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: 95.6% of the sub N vs. 'one' x Syntax is less than 0. [6]

Exploiting the reference-predication trade-off hypothesis even further, another exploratory analysis suggests that participants initially consider referential utility of the noun irrespective of the syntactic position. That is, the subordinate context yields both the basic-level and the subordinate target label referenteially uninformative, such that listeners might reason about the presence of either noun as intended to convey the comparison class when it appears in both subject or predicate position; in line with this idea, a higher rate of basic-level comparison classes was inferred from the basic-level noun comapred 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. 4.9; red vs. green bars, right facet), and more subordinate comparison classes were inferred from subordinate labels comapred to 'one' (93.9% of the credible interval of the subordinate vs. 'one' estimate was smaller than 0: $\beta = -0.77[-1.88, 0.24]$;Fig. 4.9; blue vs. green bars, right facet), across syntactic frames. [7] The data observed in the basic-level context is also consistent with this hypothesis, but the referentially-uninformative basic-level label condition is subject to a ceiling effect and hence leaves no room for any effects beyond baseline.

These empitical results provide a comprehensive picture of syntactic and pragmatic effects contributing to comparison class inference. In particular, to our knowl-

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

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

edge, this is the first extensive experiment showing 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. 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 referential utility with syntactic cues when reasoning about the contribution of the noun to the comaprison class (as shown by the noun-syntax interaction). The conclusion can be drawn that one-dimensional theories do not account for comparison class inference, and humans integrate both pragmatic and syntactic information when interpreting gradable adjectives.

4.4 Experiment 4: Direct Modification Experiment

In order to keep a simple and interpretable 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 syntactically modified by the adjective (predicate-N condition) or not (subject-N condition). However, the reference-predication trade-off view predicts that referential pressure takes off some weight from the noun used for reference and decreases its strength in constraining the comparison class, independent of the syntactic modification, as informational goals are suggested to be the primary driving force above syntactic phenomena. This prediction was investigated in this direct-modification experiment.

In this experiment, 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. 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). 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

Table 4.2: 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

of the sentence, e.g., in the dog-context there were two other prize-winners. [pt: screenshot] 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 appears in the predicate of the sentence than in the subject.

The experimental set-up was similar to the set-up of experiment 3. Five different contexts were used in this experiment: there were two dog contexts, a flower, a bird and a tree context (Table 4.2). 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 sunflower and a daisy next to each other and read "This is a sunflower" and "This is a daisy", 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., flowers). 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 la-

beling 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 4.2). For example, participants saw a picture depicting the sunflower and the daisy in pots with bows, and read: "These flowers are gifts. Notice the bow on the pots.". Finally, participants completed a comparison class paraphrase practice trial, identical to the one used in experiment 3. The warm-up trials in 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. Participants read different context stories for each context (Table 4.2). For example, for a flower context, they read "You and your friend are at their garden 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 sunflower). The other possible target (i.e., the dandelion) then appeared in a filler trial in the same block. Filler trials were identical to main trials with basic-level contexts from experiment 3. The size of referent (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.

4.4.1 Participants

The number of participants was determined via a Bayesian power analysis, requiring a power of at least 0.8 (Kruschke, 2014; Kurz, n.d.). [pt: tbd]

4.4.2 Results

[pt: tbd]

Chapter 5

A Bayesian Reference-Predication Model

The vague context-dependent nature of gradable adjectives has been promisingly formalized in 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 the increasingly influential Bayesian cognitive modelling tradition, showing a great deal of flexibility to account for various pragmatic phenomena like scalar implicature, hyperbolic language or generics, among many others (e.g., Scontras et al., 2018; Tenenbaum et al., 2011). This chapter reviews the Rational Speech Act framework and prior models of gradable adjectives, to finally propose a minimal extension of existing models formalizing the reference-predication trade-off hypothesis, allowing to flexibly incorporate reasoning about context and role of the noun in comparison class inference.

5.1 Understanding Rational Speech Act Models

Language is fascinatingly flexible and efficient; this is largely due to the fact that interlocutors do not have to encode all information explicitly in utterances they produce, but instead rely on each other's ability to infer many aspects of meaning 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* under one important assumption: speakers are approximatly *rational* with respect to their intended goal (Frank & Goodman, 2012). The Rational Speech Act approach (henceforth: RSA) views communication as recursive reasoning between speaker and listener: in interpretation-oriented mod-

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

els, a pragmatic listener L_1 infers a state of the world intended to be conveyed by a rational speaker, by using *Bayesian inference* to reason about likely world states given the observed utterance, knowing that the rational speaker S_1 chooses the utterance according to its most likely semantic interpretation by a literal listener L_0.

The idea of language as a form of 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 proposes that coordination of intended meaning between interlocutors can be captured via iterative application of probabilistic mechanisms, and context-dependence of meaning can be captured as listener uncertainty about the message encoded in utterances by the speaker (i.e., the state of affairs in the world she wishes to communicate) - in Bayesian cognitive modelling spirit as *subjective beliefs* of the listener - that is, as a *probability distribution* over possible states of the world (Tenenbaum et al., 2011). The listener agent can then update her beliefs about the world upon learning a proposition via Bayes' rule - namely upon hearing an utterance u produced by an informative speaker S_1 (Frank & Goodman, 2012).

These mechanisms of RSA are best illustrated by a simple example from a reference game, as 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 such 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, let us assume she has

a finite set of utterances $U = \{blue, green, square, circle\}$. [1] A listener then tries to recover the intended referent (i.e., the blue square) upon receiving an utterance (e.g., "blue"). As briefly mentioned above, standard RSA models consist of three layers: a pragmatic speaker S_1 who 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 every RSA model is the naïve literal listener agent L_0 that S_1 reasons about when choosing an optimal utterance to communicate the blue square, that computes a probability distribution over states consistent with the received utterance u (i.e., conditioning on $[\![u]\!](s) = 1$):

$$P_{L_0}(s|u) = \frac{[\![u]\!](s) \times P(s)}{\sum_{s' \in C} [\![u]\!](s') P(s')}$$

Given that the denominator is a constant, it can be dropped for simplicity, so that the probability of a particular state s given u is *proportional* to the literal meaning of $[\![u]\!](s)$ and the prior probability of s:

$$P_{L_0}(s|u) \propto [\![u]\!](s) \times P(s)$$

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).

So for our example utterance 'blue' the literal listener L_0 infers the following distribution (Table 5.1), since the utterance equally applies to two objects in the example context:

Table 5.1: The probability distribution over states inferred by L_0 when hearing the utterance 'blue'

State	Probability
blue circle	0.5
blue square	0.5

One crucial component of the L_0 is the *literal meaning* of the observed utterance

[1] The finite set of alternative utterances is a crucial assumption made in RSA. It is a highly relevant question for future research how human interlocutors actually determine the set of relevant alternatives.

u. In RSA, literal semantics computation is based on a form of Montague's compositional semantics, classically 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 in context Fig. 5.1, applying the utterance 'square' to the blue circle would return `false`, but 'blue' would be `true`:

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

$$[\![blue]\!](blue - circle) = 1$$

In information-theoretic terms, L_0 provides a hook to compute the *informativeness* of particular utterances as communicating particular states, where informativeness is quantified by the utterance's 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' (Frank & Goodman, 2012):

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

$I_{\tilde{u}(s')}(s')$ measures how much information is gained when hearing the utterance u, assuming a known distribution $\tilde{u}(s')$ over states of the world that are coveyed by the literal interpretation $[\![u]\!]$, implying the probability of s'; i.e., it measures how *surprising* it would be to observe s' upon observing 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. 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.

The next RSA layer, $P_{S_1}(u|s)$, incorporates the notion of a cooperative speaker. Specifically, it can be captured 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 uttering the particular utterance over other available options:

$$U_{S_1}(u; s) = logL_0(s|u) - C(u)$$

Given the L_0 derivation above, it can be observed that speaker utility is anti-proportional to the surprisal of the utterance. The cost function $C(u)$ is also an important hook for integrating psychologically plausible information about speaker-tendencies, like frequency or complexity of particular utterances comapred to others. Now the rational speaker S_1 strives to maximize 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 (see above) described by a *softmax* function:

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

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 utterance applicable to the target 'blue square' (Table 5.2):

Table 5.2: The 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

The parameter α 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).

Finally the top-level layer, the pragmatic listener L_1, reasons about this speaker behaviour given a particular utterance u ('blue'), using Bayes' rule: [2] [3]

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

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 *blue* in order to communicate about 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'. Since the denominator is a constant, it can be dropped, resulting in the probability of a particular s given u being *proportional* to the speaker production probability $P_{S_1}(u|s)$ times the state prior $P(s)$:

$$P_{L_1}(s|u) \propto P_{S_1}(u|s)P(s)$$

[2] This recursive depth of three levels has been argued to be cognitively plausible, because it implements first-order reasoning of an agent about other agent's intentions, and requires a reasonable amount of computational resources (Frank & Goodman, 2012). Yet this is just a practical approximation, and some models (e.g., production-oriented models) employ additional levels.(Scontras et al., 2018)

[3] Usually, the context C is assumed to be shared and knwon to both speaker and listener, so C will be dropped for simplicity in further explanations, [pt: incorporate somewhere else]

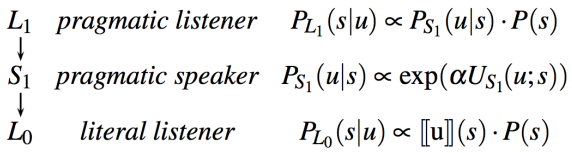

Figure 5.2: A schematic depiction of a vanilla RSA model (Scontras et al., 2018)

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). So in this example, upon hearing 'blue', L_1 would infer that the speaker is more likely to have meant the blue square (Table 5.3):

Table 5.3: The 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 receiving the same utterance 'blue'. The social reasoning about a speaker-agent incorporated in L_1, which differentiates it from L_0 who 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 - that is, L_1 *explains away* the other potential intended state (Table 5.3). In contrast, L_0 infers equal probability of both the blue square and the blue circle (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 (Frank & Goodman, 2012).

"Speech acts are actions; thus, the speaker is modeled as a rational (Bayesian) actor. He chooses an action (e.g., an utterance) according to its utility. The speaker simulates taking an action, evaluates its utility, and chooses actions based on their utility. Rationality of choice is often defined as choice of an action that maximizes the agent's (expected) utility. Here we consider a generalization in which speakers use a softmax function to approximate the (classical) rational choice to a variable degree" (problang) The speaker chooses utterance u to communicate a state s to L0 , by trying to minimize effort for L0 to arrive from u at s, i.e., by minimizing surprisal of s given u, while trying to also keep utterance cost minimal. Having this utility function in mind, the S1 computes a probability distribution over utterances having an s in mind, in proportion to the speaker's utility function Us (above), where alpha may control for speaker optimality "To interpret the utterance, the pragmatic listener considers the process that generated the utterance, in the first place" in the form of the S1.

threshold semantics, where the threshold is probabilistically inferred (Lassiter & Goodman, 2017) for a given comparison class.

Lassiter & Goodman (2013, 2017) first provided a model of gradable adjective interpretation within the RSA-framework, showing that a Bayesian approach can capture their vague meaning via inference over the latent threshold variable θ underlying the adjective semantics. Importantly, probabilistic reasoning provides tools to capture uncertainty over certain aspects of the message, in this particular case - the speaker's intended meaning of the adjectival utterance. In the proposed model, the listener jointly infers the value of the threshold along with the state of the world - i.e., the degree of the propoerty under discussion. The literal meaning of adjectives is therefore formalised in terms of degree-semantics, assuming that the lexical entry of the adjective specifies the underlying scale and its polarity. The authors assume a standard RSA model with three levels, adding one crucial component - the threshold, entering the literal semantics of the adjective on the level of L0. In order to allow specifying the compositional semantics of the utterance via the L0 which requires the computation of the truth of an utterance for a given world state, the authors propose L1 consider all possible assignments of values the latent variable, given a prior over that variable. The assumed values are then iteratively passed down through the model, such that it can be computed how likely it is that S1 would produce the observed utterance if the threshold took on a particular value:

$$P_{L_0}(s|u, V) = P_{L_0}(s|[\![u]\!]^V = 1)$$

$$P_{S_1}(u|s, V) \propto exp(\alpha \times ln(P_{L_0}(s|u, V) - C(u)))$$

Via Bayes' rule, L1 can then infer the joint posterior distribution over all possible combinations of states and values of the latent threshold:

$$P_{L_1}(s, V|u) \propto P_{S_1}(u|s, V) \times P_{L_1}(s) \times P_{L_1}(V)$$

In this work, the relevant comparison class was assumed to be implicitly supplied.

QUD stuff

Tessler et al. 2017: Listeners use their world knowledge to infer the comparison class about what worlds are plausible given a specific comparison class, what comparison classes are likely to be talked about, and how a rational speaker would bhave in a given world and given a comparison class.

5.2 Refpred-RSA

[pt: tbd]

Chapter 6

Discussion

Declaration

I declare that..

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 indicated a native language other than English. Data from 3 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: 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.1.3 E2, E3 Response Classification

The following free-production responses were excluded from analysis in Experiment 2 (Noun Production): "last", "oak", "finch", "duck", "lavender", "salmon", "goldfinch", "dahlia", "poetry", "one", "lablador", "geony".

The following provided responses (corrected for misspellings, capitalization and number) were classified as subordinate: "goldfish", "hummingbird", "canary", "doberman", "sunflower", "swordfish", "sparrow", "tuna", "peony", "chihuahua", "daisy", "clownfish", "Great Dane", "goose", "bonsai", "pug", "dandelion", "swan", "eagle", "redwood", "blue swordfish", "eagle that is landing", "red clownfish", "Dandelion with seeds", "sequoia", "redwood tree", "peony flower".

The following provided responses (also corrected) were classified as basic-level: "birds", "dogs", "great dog", "fish", "fishes", "flowers", "trees", "animal", "plants", "weeds".

The following free-production responses were excluded from analysis in Experiment 3 (Comparison Class Inference): "that man is big", "that's small boy", "that is big one", "that's a small doberman", "that bird compare small", "that boy is small", "me is big", "but some one small", "that pug small", "you are small", "beauty for fish", "yes", "aim", "growth", "honest", "medicine", "heathy", "dogs name", "trees name", "big", "tall", "small", "cute", "good", "bushes", "why it in land", "what is this flower", "fish nose", "labrador".

The following provided responses (corrected for misspellings, capitalization and number) were classified as subordinate: "chihuahuas", "bonsai", "pugs", "great danes", "sunflowers", "dobermen", "swordfish", "dandelions", "goldfish", "eagles", "redwoods", "hummingbirds", "sequoias", "bonsai trees", "redwood trees", "other swordfish", "the other sunflowers".

The following provided responses (also corrected corrected) were classified as basic-level: "birds", "dogs", "fish", "flowers", "birds in the sky", "big dogs", "things", "objects", "dogs", "dogs that we see", "fish that we see", "birds that we see", "flower", "trees", "animal", "the other birds", "the other dogs", "weeds", "small flowers", "dog", "large dogs", "giant trees", "breeds", "plant", "variety of dogs",

"long trees".

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), 10.1016/j.jml.2012.11.001. https://doi.org/10.1016/j.jml.2012.11.001. *Journal of Memory and Language*, *68*(3).

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.

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*.

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), SI77–SI103.

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.

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.

Graf, C., Degen, J., Hawkins, R. X., & Goodman, N. D. (2016). Animal, dog, or dalmatian? level of abstraction in nominal referring expressions., In *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*.

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

Ilieva, S., Ji, X., Rautenstrauch, J., & Franke, M. (n.d.). Minimal architecture for the generation of portable interactive experiments. 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*.

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

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. (n.d.). Bayesian power analysis: Part iii.b. what about 0/1 data? https://solomonkurz.netlify.app/post/bayesian-power-analysis-part-iii-b/

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.

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.

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.

Scontras, G., Tessler, M. H., & Franke, M. (2018). Probabilistic language understanding: An introduction to the rational speech act framework. 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

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), https://academic.oup.com/jos/article-pdf/3/1-2/1/9835673/1.pdf, 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), SI105–SI122.

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., Lopez-Brau, M., & Goodman, N. D. (2017). Warm (for winter): Comparison class understanding in vague language, In *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 *42th annual meeting of the 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/