NORTHWESTERN UNIVERSITY

Polarity and Comparison at the Interface of Language and Cognition

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Linguistics

By

Daniel M. Tucker

EVANSTON, ILLINOIS

May 2021

Abstract

This dissertation investigates the nature of the interface between morphosyntax and cognition. My goal is to connect formal semantic theories of meaning with theories of cognition, drawing on the initial hypothesis that the interface between language and cognition is transparent. I look at different forms of adjectival comparatives—positive and negative—and their interplay as a case study for understanding the general mechanisms at work at this interface. Specifically, I leverage formal semantic proposals and the transparency thesis to generate predictions about (i) how long the evaluation of different statements is supposed to take; and (ii) what mechanisms might lead to the differences in the behavioral responses that I observe. This work contributes to our growing understanding of how the nature of the interface might constrain the sorts of structures that occur in natural language.

Throughout this dissertation, I use formal semantics as a bridge between linguistic representations (i.e., morphosyntactic objects) and nonlinguistic representations (e.g., representations of line lengths). In thinking about this interface, I posit a close connection (potentially one-to-one) between the atoms of meaning on the linguistic side, and representations and operations on the nonlinguistic side. My case study will be positive and negative adjectives (their antonyms) when they occur in comparative sentences. These cases are interesting because semanticists have posited internal structure to what otherwise appears to be a word, e.g. *short* equals something like *not tall*. The operating idea is that formal semantics can provide suggestions about which expressions plausibly invoke transformations on representations, which may induce measurable processing costs.

I begin this dissertation by setting out to link a decompositional account of *shorter* with processing by adopting the hypothesis that each linguistic unit is linked explicitly to a cognitive operation. In two experiments—a sentence-to-picture verification task and a picture-to-sentence verification task—I find that *shorter* comparatives take longer to process than *taller* comparatives, in line with the decompositional analysis that posits the former as representationally more complex than the latter. Next, I extend this analysis to analytic comparatives with *less*, and ask whether the behavioral evidence is consistent with decomposition here as

well. Importantly, I ask whether there is an additive effect of processing multiple instances of negation in a single comparative statement (e.g. *less short*). My results suggest that a decompositional analysis of *less* comparatives is tenable given the behavioral evidence I find. Finally, I extend my psycholinguistic investigation to adjectival comparatives to include consideration of evaluativity, following recent theories on the distribution and interpretation of gradable adjectives that posit silent morphosyntactic elements. My experimental evidence suggests that evaluativity as I operationalized it may capture some of the psychological realia associated with evaluative comparatives, but further research will be necessary to unpack precisely what realia these results correspond to.

This dissertation is important for researchers interested in the interface between linguistic and non-linguistic cognition, and more specifically, in the prospects of linking morphosyntactic units to cognitive operations. My emphasis is on the explanatory value that can be gleaned from such a study by positing explicit linking hypotheses between our formal semantic theories and our models of cognitive processing. While my investigation focuses on the representation and processing of gradable adjectival comparatives in English, the methods and analyses I use can be applied more broadly to other constructions and other languages.

Contents

List of Tables

List of Figures

1 Introduction

In this dissertation, I use formal semantics as a bridge between linguistic representations (i.e., morphosyntactic objects) and nonlinguistic representations (e.g., representations of line lengths). In thinking about this interface, I posit a close connection (potentially one-to-one) between the atoms of meaning on the linguistic side, and representations and operations on the nonlinguistic side. My case study will be positive and negative adjectives (their antonyms) when they occur in comparative sentences. These cases are interesting because semanticists have posited internal structure to what otherwise appears to be a word, e.g. *short* equals something like *not tall*.[1] The operating idea is that formal semantics can provide suggestions about which expressions plausibly invoke transformations on representations, which may induce measurable processing costs.

Formal, truth-conditional semantic approaches to meaning strive to satisfy two desiderata: empirical adequacy and compositionality. On these approaches, a theory of meaning is considered empirically adequate to the extent that it pairs statements of a language L with truth conditions in such a way that it accords with the truth/falsity intuitions of native speakers of L. The second component concerns a theory's capacity to capture the unbounded creativity of human language. Speakers are able to combine a finite number of elements in novel ways, and are able to interpret complex, never-before-encountered utterances; as such, a theory that seeks to capture speakers' ability to interpret composites must be compositional—it must, in other words, capture the generative process of composing and decomposing the meanings of complex sentences. Taken together, these desiderata constitute two important goals for many contemporary semantic theories.

Another possible desideratum that a semantic theory might aim to satisfy concerns capturing how people understand the meaning of sentences. One way to do this while holding

[1] While I will not be offering any new claims or hypotheses about the derivation of these expressions, e.g. whether their decomposition is lexical or syntactic, I take them to offer an interesting test case for the transparency thesis offered in the main text.

to a compositional, truth-conditional account is to make a distinction between what Church (1936) calls 'functions in extension' and 'functions in intension' (cf. Pietroski 2010).

(1) a. $y = |(x-1)|$

b. $y = \sqrt[+]{x^2 - 2x + 1}$

The functions in (1) are equivalent in extension (i.e. they are characterized by the same domain and range), but they are unique in intension—each functional description stipulates a different set of computations. (1a) calls for subtracting 1 from the input, and computing the absolute value of the result; (1b) calls for squaring the input, subtracting its double, adding one, and computing the resultant positive square root. As Pietroski (2010, pg.251) suggests, a human mind might be capable of implementing (1a), but not (1b) (cp. Marr 1982).

In much the same fashion as (1), contemporary truth-conditional semantic theories assign functional interpretations to morphosyntactic expressions, and these interpretations may be understood either as functions in extension or as functions in intension. To unpack what this difference might look like in application, consider the statement expressed in (2): *Most of the dots are blue*. As discussed in Pietroski et al. (2009), the truth of (2) can be in (at least) any of the ways given in (3), where DOT and BLUE stand for $\lambda x.dot(x)$ and $\lambda x.blue(x)$, respectively.

(2) Most of the dots are blue.

(3) a. $> (|DOT \cap BLUE|, |DOT - BLUE|)$

b. $OneToOnePlus(DOT \cap BLUE, DOT - BLUE)$

c. $> (|DOT \cap BLUE|, \frac{1}{2}|DOT|)$

d. $> (|DOT \cap BLUE|, |DOT| - |DOT \cap BLUE|)$

In both (3a) and (3b), *most* indicates the same relation, but only (3a) specifies this relation in terms of cardinalities. (3c) represents another possibility, due to Hackl (2009), which is truth-conditionally equivalent to (3a), but which allows for the computation of rational numbers. A further possibility, noted by Lidz et al. (2011), is given in (3d); the representation calls for

the computing the cardinality of all the dots and subtracting from this number the cardinality of the blue dots.

Given many possible truth-conditional equivalences, one wants to know if there is a fact of the matter about which specification of truth conditions is better than others. Are they simply notational variants, like the difference between measuring temperature in Fahrenheit or Celsius? Or can at least some contrasts be regarded as alternative psychological hypotheses about speakers? I suggest that we can fruitfully think of semantic description in the latter sense, and thereby gain insight into the nature of human language understanding. A smaller question such an investigation can address is: how do we decide between extensionally-equivalent descriptions? And a larger question it raises is: which functions is the human mind able to biologically implement?

To think about how formal theories of meaning may be made more amenable to a theory of language understanding, consider Marr's (1982) levels of analysis in information-processing systems: (1) computational, (2) algorithmic and (3) implementational. Seeing little hope, at present, of linking formal semantic description with Level 3, we might focus instead on the first two levels. Analyses at the computational level describe what a system does, and why. Analyses at the algorithmic level describe how the system does what it does: specifically, the types of representations it uses and the processes it employs to manipulate those representations. A step in between, straddling the line between Marr's computational and algorithmic levels, would be to delimit the class of possible algorithms by looking at what information the system draws on. Peacocke (1986) calls this 'Level 1.5'. In this dissertation, I use behavioral measures (reaction times, truth/falsity judgments, etc.) to probe the classes of algorithms that might be involved in speakers' language understanding.

Specifically, I investigate semantic description at 'Level 1.5' by examining negation and English comparatives as a test case. The results of early cognitive psychology studies (e.g. Just and Carpenter 1971, Clark and Chase 1972, Trabasso et al. 1971, Clark et al. 1973, *inter alia*) report longer processing times for 'negative' statements vis-à-vis their positive analogues. These effects have been found both for sentences with overt sentential negation

(e.g. *The dots are not red* versus *The dots are red*), as well as sentences featuring 'linguistic negation' (e.g. *Few of the dots are red* versus *Many of the dots are red*; *A minority of the dots are red* versus *A majority of the dots are red*; cf. Klima 1964). Throughout this early literature, 'negative features' were consistently found to impact the time it took to process a sentence—an observation which led Clark and Chase (1972) to postulate explicit algorithms linking morphosyntactic constituents to cognitive operations. My goal is to consider the interplay between explicit algorithms like these ones along with semantic descriptions that posit decomposition (i.e., morphosyntactic analysis below the word level), to better understand how linguistic units align with cognitive operations.

To do this, I adopt the hypothesis that the mapping between linguistic structure and mental representation is transparent (Lidz et al. 2011). Lidz et al.'s Interface Transparency Thesis (ITT) holds that "the verification procedures employed in understanding a declarative sentence are biased towards algorithms that directly compute the relations and operations expressed by the semantic representation of that sentence" (p.233). To understand how ITT works, let's suppose that speakers understand, e.g., *Most of the dots are blue* as a set of computations that can be described symbolically as $>(|DOT \cap BLUE|, |DOT| - |DOT \cap BLUE|)$. What ITT implies is that speakers of English who evaluate the sentence in question in this way will be biased towards verification procedures that involve representing the number of red dots, the total number of dots, and the result of subtracting the first cardinality from the second. In line with ITT, I hypothesize that to understand the meaning of a morpheme, word, or sentence is to recognize it as an instruction to generate a particular algorithm (or class of algorithms) which may be used in the construction of simple or complex non-linguistic representations (cf. Pietroski 2010). This linking hypothesis crucially allows me to use decompositional analyses to generate explicit predictions about the mapping between linguistic structure and cognitive operations, which I test by looking at how people understand sentences in a laboratory setting.

What might this linking hypothesis look like when applied to different (competing) semantic analyses of gradable adjectives like *tall* and *short*? I consider two views: atomic and

decompositional. An atomic view of antonymy might analyze *tall* as in (4a) and *short* as in (4b): where the meaning of *tall* involves mapping individuals x to their height (here, an interval subset of a scale representing height), and that of *short* involves mapping individuals to the near-complement of their height. There is an imaginable operation that could relate these two denotations (e.g., a complementation operation that preserves $H(x)$ in both intervals), but there is no meaningful part of the morphosyntactic representation of either adjective that points to such an operation. Instead, both members of such adjectival pairs are morphosyntactically indivisible, and stand in no representationally transparent containment relation. And while native speakers of English have the intuition that *tall* and *short* are importantly semantically related, this intuition is not represented compositionally.[2]

(4) a. $[\![\text{TALL}]\!] = \lambda x.(0, H(x))$

b. $[\![\text{SHORT}]\!] = \lambda x.(H(x), \infty)$

An atomic view on antonymic pairs stands in contrast with what I will call a decompositional theory of antonymy (cf. Heim (2008)). On such a view, there is no atomic negative gradable adjective like *short*. Instead, the surface form *short* spells out a collocation of two meaningful morphosyntactic units, one of which is the same as that which spells out *tall*, (5a), and the other is a sort of negation operator, as in (5b), which, when combined with TALL, yields *short*.[3]

(5) a. $[\![\text{TALL}]\!] = \lambda x.(0, H(x))$

b. $[\![\text{LITTLE}]\!] = \lambda A.\neg A$

Thus, in Chapter 2 of this dissertation, I explore whether processing and behavioral evidence (i.e., response latency and accuracy) can be leveraged to adjudicate between these two types of theories of antonymy. Adopting the ITT of Lidz et al. (2011), I predict that,

[2]The decompositional approach underwrites speaker knowledge of the relevant entailments via a derivational or structural relationship, whereas the non-decompositional approach does not.

[3]I will need to say a bit more about how the interpretations in (5) combine to deliver extensionally-equivalent representations of the meaning of *short*, but this is not difficult.

if indeed the representation of negative adjectival comparatives like *shorter* contains a morpheme that quietly introduces an operation like that specified in (5b), this should be evident in the processing of comparatives that hypothetically contain that morpheme as part of their morphosyntactic representation. Specifically, I predict that comparatives containing the negative *short* should take longer to process than comparatives with the positive counterpart *tall*.[4] To preview, I find that this is the case. These results are in line with, and extend, the results of the early cognitive psychology literature.

Taken together with interface transparency, this initial result suggests that a decompositional view of negative gradable adjectives is tenable. More generally, these results suggest that explicitly hypotheses about linkages between morphosyntactic units and cognitive operations may be promising for adjudicating between competing representational theories.

Building on my initial results, Chapter 3 investigates whether similar evidence can be leveraged to address competing decompositional analysis of *less*-comparatives. Heim's (2008) decompositional analysis posits two variant 'negative' morphemes: a scopally-fixed LITTLE—the same as Büring posits—whose spellout in combination with TALL yields *short* as in (7b), and a scopally-mobile LITTLE*, (6b), whose spellout in combination with -ER yields *less* as in (7b).

(6) a. $[\![\text{LITTLE}]\!] = \lambda A.\neg A$

b. $[\![\text{LITTLE*}]\!] = \lambda d.\lambda A.d \notin A$

(7) a. ER LITTLE* $>$ *less*

b. LITTLE TALL $>$ *short*

Addressing whether processing evidence can bear on these theories involves two steps. The first is simply whether there is evidence for something like LITTLE (whether Büring's LITTLE or Heim's LITTLE*) in *less* comparatives at all. If so, given the assumptions I have maintained, participants should take longer to evaluate sentences with *less* than their *more*

[4]Specifically, I will not be looking at the step-by-step timecourse of processing, but rather, the global time to process.

counterparts. The second is whether there is evidence for the distinction between LITTLE and LITTLE*. To address these questions, I present two sentence verification tasks with synthetic and analytic (*more* and *less*) comparatives. In both experiments, the RT evidence points to a robust asymmetry between *less* and *more*, as expected by decompositional analyses in general. Moreover, the effect of *less* processing has an additive effect on response latency with that of *short*, suggesting that the two hypothetical instances of LITTLE contribute independent processing costs. These results suggest the viability of a decompositional account for *less*, but do not yet decide between the Büring and Heim variants. The initial statistical probes that I report in this chapter are, as I discuss, inconclusive.

Moreover, any approach to deciding between a theory involving a unitary LITTLE and one that posits a distinction between LITTLE and LITTLE* will require further fleshing out of the morphosyntactic differences of the relevant comparatives. In particular, recent theories of the interpretation of gradable adjectives (e.g. Rett 2015; Moracchini 2018) observe 'evaluative' effects in certain of their comparative occurrences, and posit a silent morphosyntactic element like EVAL in (8) to account for these effects. Such investigations dovetail with morphosyntactic approaches like Embick (2007), and semantic investigations of 'metalinguistic' (McCawley 1988; Morzycki 2011; Giannakidou and Yoon 2011) or 'categorizing' comparatives (Wellwood 2014).

(8) $[\![\text{EVAL}]\!] = \lambda D \lambda d.D(d) \wedge d > s$

(9) a. ER TALL $>$ *taller*

b. ER EVAL TALL $>$ *more tall*

Thus, Chapter 4 extends my psycholinguistic investigation of decomposition to include the study of evaluativity, and the hypothetical distribution and interpretation of EVAL. First, I conduct a post-hoc analysis on the data presented in Chapter 3, to see whether including a modeling predictor corresponding to EVAL resulted in a better-fitting model. This initial exploratory phase offers promising prospects for the study of evaluativity. However, no additional light is shed on teasing apart the subtleties that differentiate Heim's approach

from Büring's, at least for the present.[5]

Following this initial exploration, Chapter 4 presents an experiment designed to test different hypotheses about the meaning contribution of EVAL, all of which are designed to operationalize the relevance of comparing a mentioned object (e.g., *Box A* in *Box A is taller/more tall than Box B*) to a class of objects in the context. In particular, I test whether participants' responses to analytic comparatives are sensitive to rank comparison (Bale 2006, 2008, 2011), simple length comparison, or something I dub 'goodness-of-fit', building on (Wellwood 2014). This last measure, novel to this dissertation, corresponds to a participant's likelihood of labeling an object as *tall*, *short*, etc., in a given context, as measured empirically in my experiments.

To weigh the predictions made by a 'goodness-of-fit' hypothesis against two other hypotheses—length comparison and rank comparison—my participants evaluated the truth/falsity of subcomparatives (e.g. *Box B is taller than Box C is wide*). Analysis of the judgment data reveals that whenever goodness-of-fit and another hypothesis made disparate predictions, my goodness-of-fit measure consistently outperformed the prediction accuracy of that alternative measure. These results suggest that people consider how well an object fits to a category in analytic comparatives, yet leave open the question of precisely what my 'goodness-of-fit' metric captures concerning their evaluation procedures. I present the results of a brief post-hoc investigation in which I fit a series of deep neural networks to the judgment data in hopes of revealing variable importance biases introduced by the best fitting model. These model weights reveal a number of interesting differences between, e.g. sensitivity to different statistical properties of the stimuli that distinguished evaluation of *short* vs. *tall*. Thus, these model weights may help to unpack precisely what goodness-of-fit captures in terms of participant evaluation strategies.

This work will thus be important for researchers interested in the interface between lin-

[5]It may be possible to do this leveraging the data from the novel experiment reported in Chapter 4, but this awaits future research.

guistic and non-linguistic cognition, and more specifically, in the prospects of linking morphosyntactic units to cognitive operations. I emphasize, along the way, the added explanatory value we get by positing explicit linking hypotheses between our formal theories and cognitive processing. While my investigations focus on the representation and processing of gradable adjectival comparatives in English, the methods and analyses I use can potentially be applied much further.

Finally, while some of the decompositional analyses I cite here (e.g. those of Büring 2007a; Heim 2008; Embick 2007; Moracchini 2018) may assume or be particularly amenable to explanations in Distributed Morphology (Marantz 1997), I see this work as independent of such theoretical frameworks, and thus remain neutral between the relevant competing options. All that matters, theoretically, for my dissertation project is that we have competing hypotheses about whether unitary surface forms—e.g., *short* and *less*—obscure non-atomic morphological or syntactic structure. That is, I require only a distinction between the abstract, morphosyntactic representation of a given word and its surface realization, such that the former may be complex and the latter appear simple. Whether the internal structure of the form is composed in the lexicon or in the syntax need not concern us here.

The remainder of this dissertation is organized as follows. Chapter 2 seeks processing evidence supporting a decompositional account of negative adjectival comparatives (e.g. *shorter*). Chapter 3 builds on the results of Chapter 2, turning to *less*-comparatives to probe whether a similar evidence can be found in favor of related decompositional analyses of such forms (e.g., Heim 2008). Chapter 4 explicitly addresses potential semantic differences between putative analytic and synthetic variants of such comparatives (e.g., the comparison between *more tall/taller* and *more short/shorter*, and experimental evidence for how those semantic differences (i.e., whether it is evaluative or not) are understood. Chapter 5 concludes.

2 Negation, decomposition and transparency

How does formal semantics relate to language understanding? And, how can linguistic processing bear on questions about the atoms of compositional interpretation? Recent proposals in the literature on superlatives (Hackl 2009, Szabolcsi 2012), negative comparatives (Rullmann 1995, Büring 2007a,b; cp. Heim 2008), and positive comparatives (Solt 2015, Wellwood 2012, 2015) have highlighted the compositional role of units below the word level. With negative comparatives, much recent debate has centered on whether forms like *shorter* decompose into LITTLE-TALL plus -ER. I look for evidence of such decomposition in processing, by investigating the time it takes to judge sentences containing *taller* and *shorter* as true or false of simple pictures.

The results of early cognitive psychology studies (e.g. Just and Carpenter 1971, Clark and Chase 1972, Trabasso et al. 1971, Just and Clark 1973, *inter alia*) report longer processing times for 'negative' statements vis-à-vis their positive analogues. These effects have been found both for sentences with overt sentential negation (e.g. *The dots are not red* versus *The dots are red*), as well as sentences featuring 'linguistic negation' (e.g. *Few of the dots are red* versus *Many of the dots are red*; *A minority of the dots are red* versus *A majority of the dots are red*; cf. Klima 1964). Throughout this early literature, 'negative' features were consistently found to impact the time it took to process a sentence.

This chapter contributes to early results in comprehending negation, but links the processing of negative sentences directly to how the meanings of these sentences are characterized in contemporary formal semantics. Specifically, I test for these effects with *taller* (positive) and *shorter* (negative), and examine the possibility of an additional effect of 'congruence'—whether a statement is true or false of a picture (Just and Carpenter 1971, Trabasso et al. 1971). Congruence played an important role in the construction of early cognitive models of sentence-picture verification with negative statements, and can thus support a finer-grained picture of the underlying cognitive processes involved in these tasks.

2.1 Background and motivation

Positive gradable adjectives like *tall* are morphemes—they are not amenable to further morphological analysis. However, Büring's (2007a) theory decomposes negative gradable adjectives like *short* into two parts, glossed LITTLE and TALL (cf. Heim 2008). Evidence for decomposition is seen explicitly on the surface in some languages; in Hixkaryana, for example, the antonym of an adjective like *long* is formed by two pieces, i.e. *kawo-hra*, which Bobaljik (2012) glosses as 'long-not'. My research brings to bear a new kind of evidence for these questions through an examination of gradable adjectives like *tall* and *short* in English, seeking a different kind of evidence for decomposition in sentence processing.

2.1.1 Decomposition of adjectival comparatives

In the contemporary degree semantics tradition, *tall* is analyzed as involving a relation between individuals and their heights, and a sentence like (10a) is interpreted as a comparison between those heights. 'Heights' are formalized as degrees or sets of degrees, and gradable adjectives like *tall* as relations between individuals and those degrees (Cresswell 1976, Heim 1985, 2001, Kennedy 1999, among many others). The question for this section is: how does the analysis of comparatives with *tall* relate to those with *short*, as in (10b)?

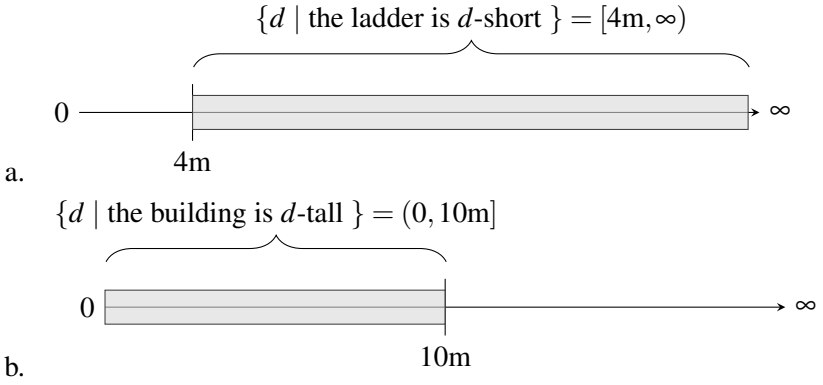
(10) a. Al is taller than Bill is.

b. Bill is shorter than Al is.

(10a) and (10b) stand in a mutual entailment relationship: competent speakers of English intuitively infer that if (10a) is true, (10b) is guaranteed to be true, and vice versa. Is this entailment relation due to their shared *forms*, or something else? On the traditional view, speakers' intuitive awareness of this relationship is not a matter of logic, per se: if both *tall* and *short* are atomic, then their dual nature isn't syntactically 'visible'. Kennedy captures the mutual entailment relation by way of something like a meaning postulate: where S is a scale, pos_S is a positive adjective associated with S and neg_S is its antonym, $pos_S(x) > pos_S(y) \Leftrightarrow neg_S(y) > neg_S(x)$ (Kennedy 2001, p.56).

Büring's (2007a) decompositional approach, in contrast, supports an analytic relationship between (10a) and (10b). His analysis begins by considering Kennedy's (2001) explanation of the oddity of (11), which is argued to follow from the hypothesis that *tall* and *short* relate individuals to incommensurable sorts of degrees, positive and negative. More formally, the measure function expressed by the negative antonym, SHORT, maps the entity referred to by *the ladder* to a set of degrees like that in (11a), while TALL maps *the building* to a set of degrees like that in (11b).[6] What Heim (2008) calls *Kennedy's constraint* is that -ER cannot compare positive and negative degrees.

(11) ? The ladder is shorter than the building is tall. ?HEIGHT



Büring points out that, as given, Kennedy's explanation for (11) incorrectly predicts that (12) should be odd as well. Since, as Kennedy suggests, a negative adjective like *short* introduces a negative set of degrees, and a positive adjective like *wide* introduces a positive set of degrees, (12) should also be anomalous.

(12) The ladder is shorter than the building is wide. LENGTH

Büring suggests that decomposition is critical to understanding this pattern. By decomposing *short* into the pieces TALL and LITTLE (where LITTLE TALL is semantically equiv-

[6]Note that Kennedy's analysis differs from Rullmann's in that Rullmann had the negative antonym 'flip' what was otherwise a positively-oriented scale (i.e. reverse the ordering relations). In contrast, Kennedy (and subsequent authors presupposing his ontology) proposes that negative antonyms introduce sets of degrees that extend from a point *d* to infinity, the complement of the set introduced by the positive antonym (see especially Kennedy 2001, p55, examples (46) and (48), for discussion).

alent to Kennedy's SHORT), he is able to argue that the component LITTLE is also shared with the decomposed form of *less* (i.e. LITTLE-ER; Heim 2006). This raises the potential for (10b) to be analyzed as ambiguous between two structures, one containing the bundling [LITTLE-ER] TALL and the other -ER [LITTLE TALL]. (12) would be interpretable on the first bundling as a less-than relation between the positive degrees introduced by TALL and WIDE. It would not be interpretable on the other bundling, since that would express a greater-than relation between the negative degrees introduced by LITTLE TALL and the positive degrees introduced by WIDE, which is barred by Kennedy's constraint.

This analysis can account for the contrast between (11) and (12) as follows. In principle, there could be two bracketings for (11), but either would be problematic. On the bundling -ER [LITTLE TALL] for *shorter*, (11) would express a greater-than comparison between positive TALL and negative LITTLE TALL, barred by Kennedy's constraint. If *shorter* were bundling [LITTLE-ER] TALL, (11) would express a less-than comparison between two instances of positive TALL. This last structure is, presumably, barred by an independent rule or preference that the second of a pair of identical adjectives delete in the *than*-clause of a comparative (cf. Bresnan 1973).

In addition to accounting for (11) and (12), Büring's account extends to cases of ambiguity with *less high* and *lower* that are not evidenced by comparatives with their antonym *higher*, (13a)-(13c) (Seuren 1973, Rullmann 1995). (13a) describes a helicopter flying some degree higher than the maximal height a plane can safely fly, while both (13b) and (13c) can describe a helicopter flying some degree lower than the maximal height a plane can safely fly, or some degree lower than the minimal height a plane can safely fly. This pattern is predicted if LITTLE is able to Quantifier Raise (Lakoff 1970, May 1977, Heim and Kratzer 1998, *inter alia*) in the *than*-clause higher or lower than *can*. (See also Rullmann 1995 for relevant data involving NPI licensing.)

(13) a. The helicopter was flying higher than a plane can fly. NOT AMBIGUOUS

b. The helicopter was flying less high than a plane can fly. AMBIGUOUS

c. The helicopter was flying lower than a plane can fly. AMBIGUOUS

Though promising, such an account faces challenges. As Heim (2008) points out, an account like Büring's would seem to predict that adjectives with *less* should always be substitutable with their negative antonym and *-er* without a change in meaning. So far this prediction is not correct in the general case. Heim shows that, while (14a) can be judged true if Polly's speed may, but needn't, exceed Larry's (perhaps because she has more time to get to her destination), (14b) cannot be read this way: (14b) only has the reading where whatever speed Polly drives, it *has to* be less than Larry's.

(14) a. Polly needs to drive less fast than Larry needs to drive. AMBIGUOUS

b. Polly needs to drive more slowly than Larry needs to drive. NOT AMBIGUOUS

Nonetheless, rolling-back the decompositional analysis for *short* entirely would, as Heim notes, have trouble explaining contrasts like that between (11) and (12). In light of this and other data, Heim posits that there are in fact two distinct LITTLEs, a scopally-mobile one for the decomposition of *less*, and a scopally-immobile one for the decomposition of *short*. One question that potentially arises for this part of her proposal is why the sentences in (15) 'feel different'; if (15a) has an instance of a covert LITTLE, and (15b) results from LITTLE morphologically exerting itself on the adjective, why does (15b) seem more difficult to understand than (15a)?[7]

(15) a. The ladder is shorter than the doorway is wide.

b. ? The ladder is shorter than the doorway is narrow.

Distinguishing the finer details of these proposals is not my focus here. Rather, I assume that the linguistic evidence amassing in favor of a decompositional analysis of *shorter* is

[7]Possibly more importantly, Beck (2013) has found some slipperiness in the judgments of speakers for the relevant scope data. Thus, so far it seems that the evaluation of decompositional analyses from the perspective of semantic theory should not yet hang on the data in (14).

strong, at least strong enough to warrant further investigation. My interest is in the fact that decompositional proposals can be seen to make explicit predictions about sentence comprehension.

2.1.2 Relating language and vision

How can the decompositional approach be tested in processing? In what follows, I draw a link with research in classic and contemporary research concerning how semantic representations might make contact with extralinguistic cognition. Of primary interest is early research on the processing of different types of 'linguistic negation', as well as recent results targeting similar questions. Ultimately, I suggest that decompositional approaches explicitly predict that negative adjectival comparatives should take longer to judge true or false than positive comparatives.

Beginning with the cognitive psychology literature, many proposals in the late 1960s and early 1970s were made as to what sorts of processing mechanisms would need to be deployed when people considered the truth or falsity of a sentence against a picture. While this literature is broad, I can draw some important conclusions from it. The first is that positive statements are more readily processed than negative (polarity effects), and that it is easier to verify a statement when it is true of its accompanying scene than when it is false (congruence effects).

A core assumption from this early work is that "perceptual events are interpreted" (Clark and Chase, 1972), specifically into a sort of propositional format. One motivation for this idea is the simplicity that it affords to understanding how, ultimately, a sentence meaning and a representation of a picture can be compared. If sentence meanings and perceptual events are encoded in a common representational format, the comparison can simply be one of identity—not merely truth-conditional identity, though this ultimately plays a role—specifically, *identity of representation*. I will be more explicit about this shortly.

Separately from the representational assumptions, models of sentence-picture match-

ing were designed to account for the response latencies of judgments in extremely simple tasks.[8] Typically, this type of task would involve a participant reading a sentence, considering a picture, and indicating whether they understand the sentence to be true or false of the picture. Two importantly different types of tasks were found to make different demands on the participant, and the models were designed to make the right predictions accordingly: the Sentence-to-Picture verification task and the Picture-to-Sentence verification task, which differ only in whether the picture or the sentence is presented first. I will discuss each model, in turn.

On the 'Sentence-First' model Clark and Chase (1972), the process of comparing a sentence with a picture proceeds in four stages, summarized in (16). Stage 1 involves linguistic decoding/encoding, and Stage 2 involves nonlinguistic perceptual/conceptual processing that eventuates in a representation given in the same general format as the sentence. This general format is thought to be important for comparison to proceed at Stage 3, which might also involve *transformations* of a given representation before the final check for identity. At Stage 4, participants record their judgment, typically using a button press.

(16) 'Sentence-First' processing stages (Clark and Chase 1972)

i. **Stage 1**: form a mental representation of the sentence

ii. **Stage 2**: form a mental representation of the picture

iii. **Stage 3**: compare the two representations

iv. **Stage 4**: produce a response

Stage 3 is thus crucial. In this model, it involves checking whether two representations 'mean' the same thing, where 'meaning the same' is cashed out in terms of representational identity (Clark 1969 calls this the 'principle of congruence'). However, it would be

[8] The most explicit overview of the methodology and models is given by Clark and Chase (1972), who cite Clark (1970), Trabasso et al. (1971) as important precursors, as well as an extensive list of even earlier results that informed their view. Subsequent research suggests complications in the pursuit of additive effects from simple group differences in reaction times (see Roberts and Sternberg (1993), Van Zandt and Ratcliff (1995), and Stafford and Gurney (2011) for further discussion). Though I note the potential conflicts here, I will interpret independent additive effects as indicative of individual processing stages.

overly simplistic to assume that this amounts merely to truth-conditional equivalence, or mere representational equivalence based on the initial representation of the sentence or picture. Checking for mere truth-conditional equivalence would predict that evaluating *A is above B* and *B is below A* should take the same amount of time in the same contexts. However, studies have repeatedly shown that there is a cost to sentences with *below* compared to *above*. On the other hand, merely checking whether the two representations match would be overly restrictive: comparing linguistic BELOW(A,B) and visual ABOVE(B,A) should then be judged as 'false', which would be incorrect.

Thus, according to (Clark and Chase, 1972, p.472), "Stage 3 must be endowed with a series of comparison operations, each checking for the identity of the subparts of the two representations, and each adding to the computation of *true* and *false*." There are many different ways, in the modern era of computational analogies in semantics research, to conceptualize such 'comparison operations' (e.g., reduction to a canonical form, comparison of evaluation consequences, etc.); I will attempt to remain at a fairly informal level here.

To remark briefly on the 'Picture-First Model' Clark and Chase 1972, its major surface difference from the 'Sentence-First' model is its assumption of a default, positive encoding of the picture at Stage 1, which is then checked against whatever the sentence encoding is. The default encoding in Clark and Chase's experiment is specified in terms of ABOVE. In cases where the sentence and the picture encodings don't immediately match (i.e. whenever it is not the case that the encoding of the picture is ABOVE(A,B) and the sentence is *A is above B*), one will have to transform the sentence to put it in a format that the comparison operations can understand.

Having discussed the 'Sentence-First' and 'Picture-First' models proposed by Clark and Chase (1972), I will conclude this section by briefly touching on the parameters thought to affect response latency. Clark and Chase (1972) posit a number of parameters, each of which additively contributes (citing Sternberg 1969) to the total response time. The parameters relevant to the present study are summarized in (17). A cost of $+a$ should be observed for evaluating sentences with the 'marked' or 'negative' member of a pair of linguistic opposites

(per the hit observed for *below*). And, a cost of $+b$ should be observed for the operations required to determine that the linguistic and visual encodings mismatch (the time for performing operations at Stage 3, i.e. *falsification*). In previous work, these two factors did not interact (Clark and Chase 1972, p.487). Finally, there is an overall and independent cost of t_0 for the time to plan and execute the response. This "wastebasket parameter" corrects for the execution not accounted for theoretically by any of the other parameters, and will be assumed both here and throughout the remainder of the dissertation.

(17) Parameters affecting response latency

i. a - cost of 'linguistic negation'; *Below* time

ii. b - cost of comparison operations; *Falsification* time

iii. t_0 - 'wastebasket parameter'; *Base* time

Differing somewhat methodologically from these early studies are the recent papers in the Interface Transparency suite (Pietroski et al. 2009, Lidz et al. 2011). These studies all made use of the Sentence-to-Picture verification task, but limited the viewing time for the picture to 150ms or 200ms, whereas the classic studies tended to give participants essentially as much time with the picture as was necessary to make the judgment. With a restricted viewing time, it was assumed that participants' response latencies reflect operations over the initial representation of the scene in memory.

More recently, Deschamps et al. (2015) tested similar hypotheses but with different linguistic stimuli, and a different experimental set-up. They investigated polarity contrasts with the quantifiers *more/less* and *many/few* versus quasi-mathematical expressions in a verification task that required numerical estimation and comparison. My study differs in that I test comparative adjectives, provide a shorter viewing time for the picture (theirs was 2500-2800ms), and I include tests for congruence effects.[9]

[9] A further difference is that Deschamps et al. (2015) presented their linguistic statements auditorily, rather than visually.

2.1.3 Hypotheses and predictions

I assume the decompositional analysis of English negative comparatives in line with Büring (2007a), and combine these assumptions with the predictions of the Sentence-First model of Clark and Chase (1972). In what follows, I discuss the predictions for English statements.

Sentence-first. On the decompositional analysis, the semantic representation of a positive comparative is contained within the representation of a negative comparative. Abstracting away from many details, a proposal like Büring's can be summarized as in (18). The major operand of the semantic representation is ER, which specifies a greater-than relation between two quantities. These quantities are provided by TALL(A) and TALL(B) in (18a), and by an operation over such quantities (e.g. complementation) provided by LITTLE, (18b).

(18) a. $[\![\text{A is taller than B.}]\!] = \text{ER}(\text{TALL}(\text{A}),\text{TALL}(\text{B}))$

b. $[\![\text{A is shorter than B.}]\!] = \text{ER}(\text{LITTLE}(\text{TALL}(\text{A})), \text{LITTLE}(\text{TALL}(\text{B})))$

In light of the early cognitive psychology literature, I expected that the added presence of LITTLE should correspond to an increase in processing load: processing (18b) requires to processing something like (18a) in addition to the contributions of the two instances of LITTLE. Such additional processing steps should correspond to an increase in RTs. Furthermore, I expect an additional cost of evaluating the the semantic representation in situations where it is false of the scene—when the two are *incongruent*.

On the simplest version of the Sentence-First model, the effects of polarity and congruence are expected to be additive to RT: both negativity in the sentence and falsity of the sentence given scene induce independent processing costs. Thus I predicted the fastest RTs in the positive congruent condition, and the slowest in the negative incongruent condition. The expected results can be depicted as in Figure 1.[10]

[10]Indeed, this is the pattern found by Clark and Chase (1972), when participants evaluated the sentences *A is above B* and *A is below B* in a Sentence-Picture verification task. However, Trabasso et al. (1971) reported an interaction between polarity and congruence, in which RTs were greater for negatives in incongruent situations, yet greater for positives in congruent situations. These results, however, were found in a Picture-Sentence verification

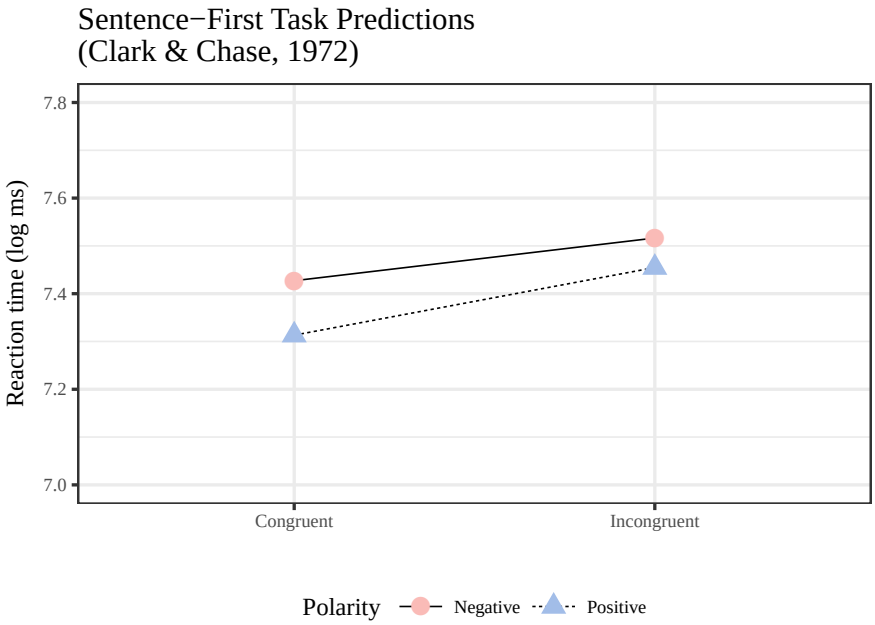


Figure 1: Predicted main effects of polarity and congruence for natural language, given the decompositional analysis of forms like *shorter* and the Sentence-First model of Clark and Chase (1972).

What about the predictions for accuracy? Clark and Chase (1972) report overall error rates of 9.7% in their task using *above* and *below*, but that these were unequally distributed between the 'positive' conditions with *above*, and the 'negative' conditions with *below*. They report that, in general, higher error rates were observed in conditions where 'more mental operations' needed to be carried out. I thus expected overall error rates to be similar in our task: broadly, higher RTs should pattern with higher error rates.

Picture-first. On the surface, the 'Sentence-First' processing model in (16) and the 'Picture-First' model should not look all that different; Stage 1 in a Picture-First model would involve forming a representation of the picture, and Stage 2 forming a representation of the sentence, as opposed to vice versa. Yet, Clark and Chase (1972) crucially assumed that, absent a linguistic cue, there was a default, positive encoding of a scene; when there was a linguistic

task where the contrast in negativity was sentential negation, e.g.: *The patch is/isn't orange*.

cue, sentence encoding could impact picture encoding. This default positive encoding of the scene associated with the Picture-First processing model explains why the results of Clark and Chase (1972) look as they do in Figure 2.

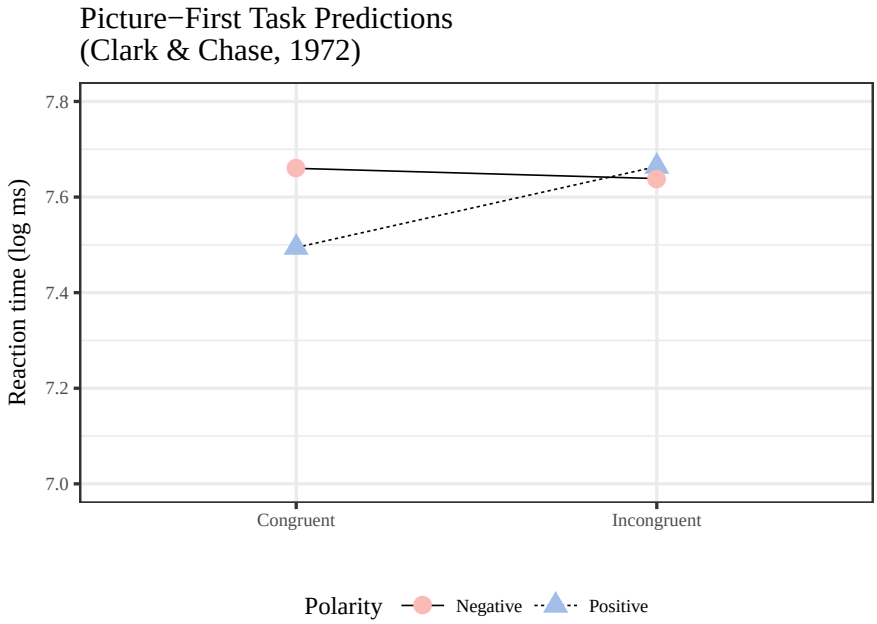


Figure 2: Results of Clark and Chase's (1972) Picture-to-Sentence verification task with *above* (positive) and *below* (negative), modeled after the presentation in Just and Clark (1973).

I anticipate that the reaction times associated with the interpretation of *taller* and *shorter* under a Picture-First experiment design will mirror the those of Clark and Chase (1972) in Figure 2. Statistically, I expect that there will be main effects of both adjectival negation (*short*) and congruence—just as in the Sentence-First design—but with the crucial addition of a significant interaction between the two factors.

2.2 Experiment 1a: Sentence-first verification task

Here I test the predictions of decompositional analyses of *shorter*, which posit that the semantic representation of sentences containing this form are strictly more complex than (and in fact contain) the representation of equivalent sentences with *taller*. In light of the early

and recent results indicating that the marked member of a positive-negative pair induces additional processing cost, I expected *shorter* should take longer to process than *taller*.

2.2.1 Design and stimuli

I designed a sentence-to-picture verification task following a 2x2 design. Participants were presented with a statement, followed by a picture, and asked to judge whether the statement accurately described the picture.

The factors manipulated were POLARITY (positive, negative) and CONGRUENCE (congruent, incongruent). In terms of statements presented, I considered the expressions that corresponded to a greater-than comparison as 'positive', and those which corresponded to a less-than comparison as 'negative'. Thus, the factor POLARITY varied whether the statement was positive (*taller than*) or negative (*shorter than*), for a total of 4 statements ("A/B is taller/shorter than B/A"). The factor CONGRUENCE varied whether the statement was true of the paired picture or not, corresponding to the congruent and incongruent conditions, respectively.

Stimuli consisted of 20 pictures featuring two lines marked A and B. The shorter line always appeared in one of two sizes (24 or 42 pixels, with a 160 pixel distance in between), and the longer line differed from the shorter by one of five different length ratios (.5, .75, .833, .875, .9). Figure 3 shows a subset of these visual stimuli: a ratio difference of .5 for an "A wins" picture (a) and a "B wins" picture (b); and a ratio difference of .75 for an "A wins" picture (c) and a "B wins" picture (d). In half of the pictures, the longer line was labeled 'A' and the shorter line was labeled 'B'; in the other half of the pictures, the shorter line was labeled 'A' and the longer line was labeled 'B'. Each of these pictures was paired with each of the 8 statements in Table 1. Every possible sentence-picture pair delivered a total of 80 trials.

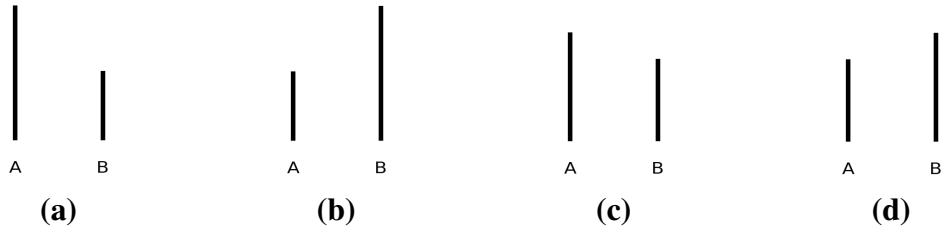


Figure 3: Sample picture stimuli used in Experiments 1a and 1b.

2.2.2 Procedure

The experiment was designed as a single-page application using JavaScript, HTML5 and CSS3. After consenting to participate, participants were presented with instructions for the experiment (see below). Following this, participants completed the 80 trials,[11] each of which was structured as follows. At the start of the trial, a statement was presented in the center of the screen, along with an indication that the statement would remain visible until the participant pressed the space bar. After pressing the spacebar, a center-oriented fixation cross appeared for 200ms, followed by a display of the picture for 200ms. 200ms after the display of the picture, a center-oriented '?' appeared, along with an indication to press 'f' if the statement matched the picture, or 'j' otherwise. Participants had a maximum of 5 seconds to record their judgment. Trials were organized into 2 blocks, each defined by one combination of comparator order (A first vs. B first). The order of presentation of the blocks and of the trials within the blocks was completely randomized.

The exact instructions given to participants were as below. As I was primarily interested in the timing of the response to our stimuli, I explicitly indicated that participants should attempt to make their judgments as quickly as possible.

> Welcome to the experiment!
>
> There are 80 trials in this experiment. Each trial will consist of a statement, an image, and your response. You will have as much time as you wish to view the

[11] No filler task items were used in this experiment or in the second experiment reported below.

statement, and then press spacebar to see the image. The image will be shown for only 1/5 of a second. Immediately afterwards, your task is to judge whether the statement accurately describes the image.

If the statement accurately describes the image, press the letter **f** on the keyboard.

If the statement does not accurately describe the image, press the letter **j** on the keyboard.

Please make this judgment as quickly as possible. The experiment will automatically advance to the next trial after 5 seconds of no response. The whole experiment should take no longer than 15 minutes to complete.

Ready? Press spacebar to begin the experiment.

I recruited 20 participants through a Human Intelligence Task (HIT) posted on Amazon's Mechanical Turk. I restricted eligibility to native speakers of English living in the United States who had completed at least 1000 HITs on Mechanical Turk with a HIT approval rate of at least 99%. Participants were compensated $2.50 for participating, and took an average of 13.5 minutes to complete the HIT.

2.2.3 Results

Here I report the results of linear and logistic mixed effects model comparisons with maximal random effects structures (i.e. including random intercepts and slopes by subject and item; best generalization for LMEMs, Barr et al. 2013). For all analyses, I used an orthogonal contrast coding scheme that assigned values of -.5 and .5 to each level of POLARITY and CONGRUENCE, respectively. The significance levels (p-values) that I report are derived from comparison of the maximal model in each case, against the same model minus the relevant parameter.[12]

[12]To add more detail here: I performed model comparisons *R*'s implementation of model likelihood ratio tests. Each model's log likelihood was compared using the χ^2 distribution.

I conducted two separate linear mixed effects model comparisons on the log-transformed RT data. Analyses for RT measures were conducted on the log-transformed RT data to respect the normality assumptions of linear mixed effects models (Gelman and Hill 2007). I plot the log-transformed RT measure, and report both the results in both logRT and milliseconds (ms) for readability. Analyses for response accuracy were summarized by participant by condition and are reported as mean percent correct.

All analyses reported in this section were conducted using R's *lme4* package (Bates et al. 2015).

REACTION TIMES

	χ^2	p-value	β
POLARITY	18.00	$p < 0.001$	-0.18
CONGRUENCE	7.53	$p = 0.006$	-0.11
POLARITY*CONGRUENCE	2.48	$p > 0.1$	-0.10

ACCURACY

	χ^2	p-value	β
POLARITY	2.08	$p > 0.1$	0.41
CONGRUENCE	1.31	$p > 0.1$	-0.34
POLARITY*CONGRUENCE	0.02	$p > 0.1$	0.07

Table 1: Summary of model comparison results for Experiment 1a (Limited VT, Sentence-first)

Reaction times. In the sentence-first paradigm, participants took longer to evaluate sentences with *shorter* than with *taller*. This was reflected in a robust main effect of POLARITY (means: negative 6.28, positive 6.10, $\beta = -0.18$, $\chi^2 = 18.00$, $p < 0.001$) in the predicted direction: RTs in the negative conditions were longer than in the positive conditions (means, in ms: negative 686.01ms, positive 589.95ms).

Additionally, participants took longer to reject false statements than to accept true statements. This was reflected in a strong main effect of CONGRUENCE (means: congruent 6.13, incongruent 6.25, $\beta = -0.11$, $\chi^2 = 7.53$, $p = 0.006$), in accord with our predictions: a state-

ment's truth or falsity with respect to its accompanying picture had a non-trivial impact on associated RTs (means, in ms: congruent 593.10ms, incongruent 627.53ms).

In terms of interaction between the two factors, both POLARITY and CONGRUENCE appear to have had independent, additive effects on response times. This was reflected in a lack of significant interaction between POLARITY and CONGRUENCE ($\beta = -0.13$, $\chi^2 = 2.48$, $p > 0.1$)[13]: negative statements always had longer associated RTs in both congruent (means: negative 6.25, positive 6.02; means, in ms: negative 661.08ms, positive 557.36ms) and incongruent conditions (means: negative 6.31, positive 6.19; means, in ms: negative 710.95ms, positive 622.53ms).

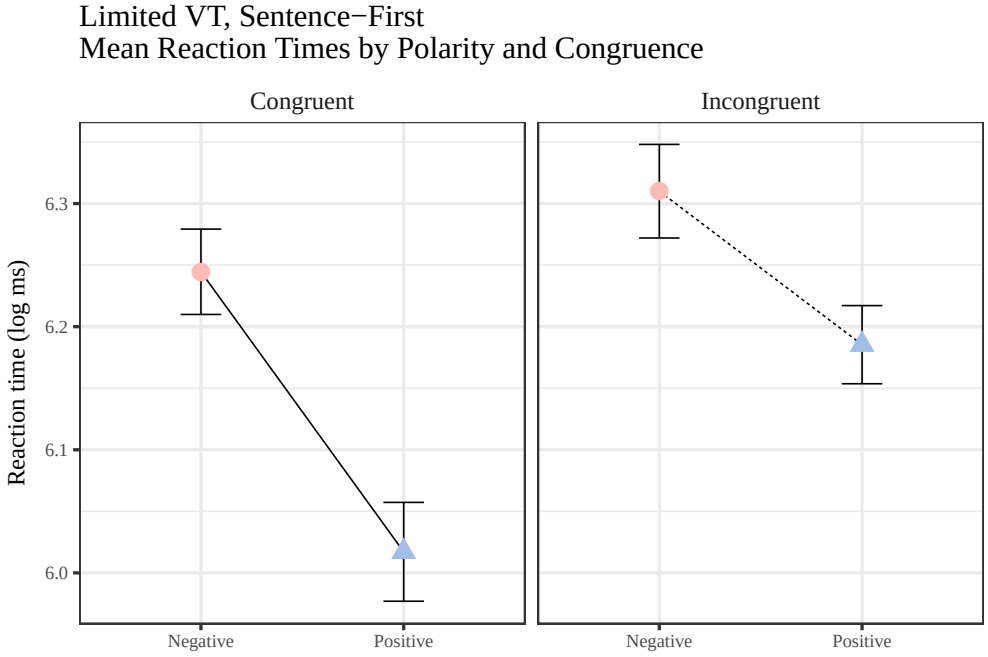


Figure 4: Experiment 1a: Limited VT; Sentence-first. Reaction times (log) by condition. Error bars represent standard error of the mean.

Response accuracy. Participants' response accuracy was not significantly worse for sentences with *shorter* than for those with *taller*. This was reflected in the lack of effect of

[13]The standard error of the fixed effect of the interaction bewteen POLARITY and CONGRUENCE was 0.062, while the standard error for the main effects was 0.032 and 0.037 for POLARITY and CONGRUENCE, respectively.

POLARITY on mean response accuracy (means: negative 92.0%, positive 94.5%, $\beta = 0.41$, $\chi^2 < 2.08$, $p > 0.1$). This result is unexpected in light of the early cognitive psychology literature, which found an inverse correlation between reaction time and response accuracy.

Additionally, participants were no less accurate at rejecting false statements than at accepting true statements. I found no effect of CONGRUENCE on mean response accuracy (means: congruent 92.4%, incongruent 94.4%, $\beta = -0.34$, $\chi^2 = 1.31$, $p > 0.1$): whether a statement was true or false given its accompanying picture made no significant difference to verification accuracy.

Analyses revealed no interaction between POLARITY and CONGRUENCE ($\beta = 0.07$, $\chi^2 = 0.02$, $p > 0.1$); there was no difference in mean response accuracy in the negative versus positive congruent conditions (means: negative 91.0%, positive 93.8%). Such was also the case in the negative and positive incongruent conditions (means: negative 93.5%, positive 95.3%).

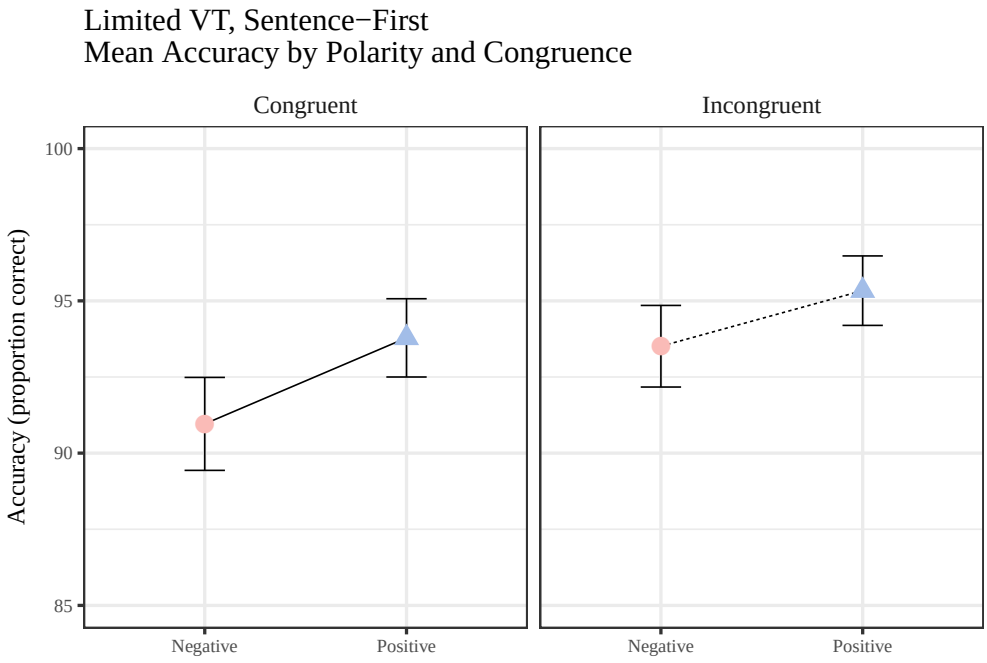


Figure 5: Experiment 1a: Limited VT; Sentence-first. Accuracy (proportion correct) by condition. Error bars represent standard error of the mean.

Statistical power. A post-hoc power analysis was conducted on the reaction times model for Experiment 1a using the R-package *pwr* (Champely, 2018). The effect size associated with this model was 0.018 with a sample size of 1390 observations. The estimated power of the model was 0.95, which indicates a more than sufficiently powered analysis. Given the results of this power estimation, I conclude with good certainty that the effect seen were not false positives.

2.2.4 Discussion

In Experiment 1a, I found that sentences with *shorter* took longer to process than sentences with *taller*, supporting the decompositional analysis on which *shorter* is strictly more representationally complex than *taller*. Furthermore, evaluating false statements took longer than evaluating true statements, and there was no interaction between adjectival negation and congruence effects—both as predicted. These results are in line with the earlier results for *above* and *below* and other pairs reported for previous Sentence-to-Picture matching tasks (cf. Clark and Chase 1972).

2.3 Experiment 1b: Picture-first verification task

Having reported the results of the Sentence-first experiment, I now turn my attention to a Picture-first task making use of the same stimuli and experimental manipulations. The crucial difference here concerns the procedure, which mirrors the picture-first task procedure used by Just and Clark 1973.

2.3.1 Design and stimuli

The experiment design and stimuli details for Experiment 1b were identical to those of Experiment 1a.

2.3.2 Procedure

The procedure for Experiment 1b was identical to that of Experiment 1a, with one important exception. At the start of each trial, a center-oriented fixation cross appeared for 200ms, followed by a display of the picture for 200ms. 200ms after the display of the picture, the statement to be evaluated appeared, along with an indication to press 'f' if the statement matched the picture, or 'j' otherwise. Participants had a maximum of 5 seconds to record their judgment. To allow for a pause between trials, a splash screen was displayed prompting the participant to press spacebar to continue. Trials were organized into 2 blocks, each defined by one combination of comparator order (A first vs. B first). The order of presentation of the blocks and of the trials within the blocks was completely randomized.

The exact instructions given to participants were as below. As I was primarily interested in the timing of the response to our stimuli, I explicitly indicated that participants should attempt to make their judgments as quickly as possible.

> Welcome to the experiment!
>
> There are 80 trials in this experiment. Each trial will consist of an image, an statement, and your response. The image will be shown for only 1/5 of a second. You will then have 5 seconds to read a statement and judge whether the statement accurately describes the image.
>
> If the statement accurately describes the image, press the letter **f** on the keyboard.
>
> If the statement does not accurately describe the image, press the letter **j** on the keyboard.
>
> Please make this judgment as quickly as possible. The experiment will automatically advance to the next trial after 5 seconds of no response. The whole experiment should take no longer than 15 minutes to complete.
>
> Ready? Press spacebar to begin the experiment.

I recruited 20 participants through a Human Intelligence Task (HIT) posted on Amazon's Mechanical Turk. I restricted eligibility to native speakers of English living in the United States who had completed at least 1000 HITs on Mechanical Turk with a HIT approval rate of at least 99%. Participants were compensated $2.50 for participating, and took an average of 15 minutes to complete the HIT. No Mechanical Turk master workers were recruited for this study.

2.3.3 Results

Here I report the results of linear and logistic mixed effects model comparisons with maximal random effects structures (i.e. including random intercepts and slopes by subject and item; best generalization for LMEMs, Barr et al. 2013). For all analyses, I used an orthogonal contrast coding scheme that assigned values of -.5 and .5 to each level of POLARITY and CONGRUENCE, respectively. The significance levels (p-values) that I report are derived from comparison of the maximal model in each case, against the same model minus the relevant parameter.

I conducted two separate linear mixed effects model comparisons on the log-transformed RT data. Analyses for RT measures were conducted on the log-transformed RT data to respect the normality assumptions of linear mixed effects models (Gelman and Hill 2007). I plot the log-transformed RT measure, and report both the results in both logRT and milliseconds (ms) for readability. Analyses for response accuracy were summarized by participant by condition and are reported as mean percent correct.

One noteworthy difference in the data trimming protocol for Experiment 1b was the choice remove observations with reaction times less than 200ms. The rationale underlying this additional trimming step is due to the experimental procedure. Recall that in the sentence-first task, participants saw a statement (until keypress), followed by a picture (for 200ms), followed by a verification screen. By the time the verification screen appeared, participants were already 200ms past the presentation of the final stimulus. In the present

(picture-first) design, participants were presented with a picture (for 200ms), followed immediately by the presentation of the statement (the final stimulus). All RTs less than this threshold were therefore filtered as uncoordinated responses.

REACTION TIMES			
	χ^2	p-value	β
POLARITY	16.03	$p < 0.001$	-0.08
CONGRUENCE	6.50	$p = 0.011$	-0.06
POLARITY*CONGRUENCE	32.47	$p < 0.001$	-0.19

ACCURACY			
	χ^2	p-value	β
POLARITY	1.02	$p > 0.1$	0.27
CONGRUENCE	0.00	$p > 0.1$	-0.01
POLARITY*CONGRUENCE	10.65	$p = 0.001$	0.48

Table 2: Summary of model comparison results for Experiment 1b (Limited VT, Picture-first)

Reaction times. In the picture-first paradigm, participants took longer to evaluate sentences with *shorter* than with *taller*. This was reflected in a robust main effect of POLARITY (means: negative 6.99, positive 6.91, $\beta = -0.08$, $\chi^2 = 16.03$, $p < 0.001$) in the predicted direction: RTs in the negative conditions were longer than in the positive conditions (means, in ms: negative 1219.52ms, positive 1123.06ms).

Additionally, participants took longer to reject false statements than to accept true statements. This was reflected in a main effect of CONGRUENCE (means: congruent 6.92, incongruent 6.98, $\beta = -0.06$, $\chi^2 = 6.50$, $p = 0.011$), in accord with my predictions: a statement's truth or falsity with respect to its accompanying picture had a non-trivial impact on associated RTs (means, in ms: congruent 1144.67ms, incongruent 1197.90ms).

Analyses also revealed that accepting true sentences with *taller* was much faster than could be accounted for with just the main effect of congruence. This was reflected in an interaction between POLARITY and CONGRUENCE ($\beta = -0.19$, $\chi^2 = 32.47$, $p < 0.001$).

RTs in the positive congruent condition were shorter than in the negative congruent condition (means: negative 7.01, positive 6.84; means, in ms: negative 1248.69ms, positive 1040.65ms), while there was little difference between the negative incongruent condition and the positive incongruent condition (means: negative 6.97, positive 6.99; means, in ms: negative 1190.34ms, positive 1205.49ms).

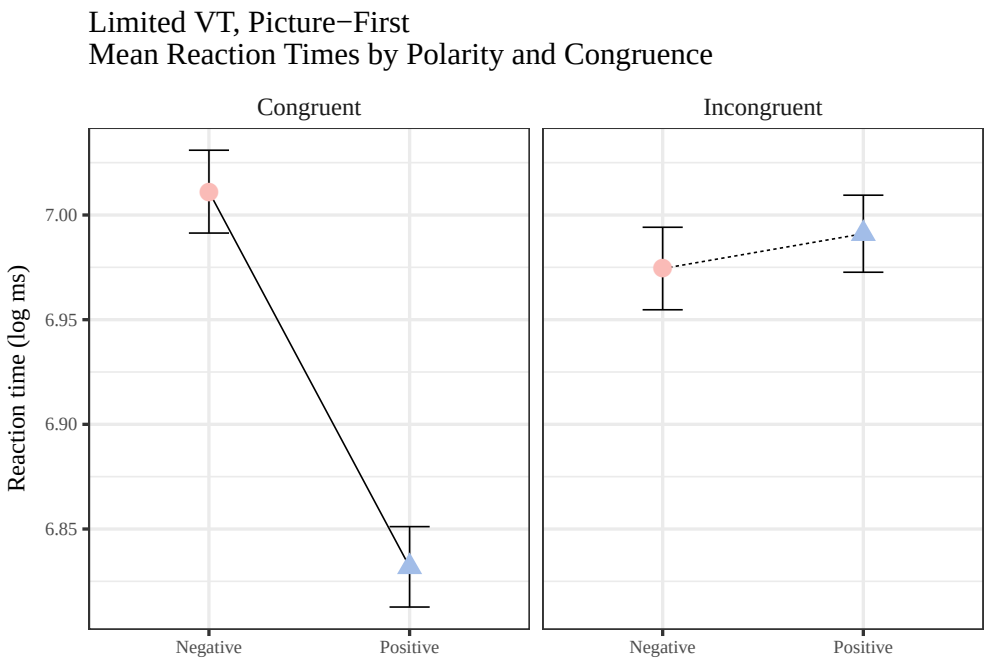


Figure 6: Experiment 1b: Limited VT; Picture-first. Reaction times (log) by condition. Error bars represent standard error of the mean.

Response Accuracy. Mirroring performance in the sentence-first paradigm, participants' response accuracy was not significantly worse for sentences with *shorter* than for those with *taller*. This was reflected in the lack of effect of POLARITY on mean response accuracy (means: negative 93.1%, positive 94.4%, $\beta = 0.27$, $\chi^2 < 1.02$, $p > 0.1$).

Additionally, participants were no less accurate at rejecting false statements than at accepting true statements. I found no effect of CONGRUENCE on mean response accuracy (means: congruent 93.6%, incongruent 93.9%, $\beta = -0.01$, $\chi^2 = 0.00$, $p > 0.1$): whether a statement was true or false given its accompanying picture made no significant difference to

verification accuracy.

In terms of variable interactions, accuracy mirrored RTs in that a robust interaction was found between POLARITY and CONGRUENCE ($\beta = 0.48$, $\chi^2 = 10.65$, $p > 0.1$): accuracy in the positive congruent condition was significantly higher than accuracy in the negative congruent condition (means: negative 90.9%, positive 96.3%), while there was less difference in accuracy between positive and negative incongruent conditions (means: negative 95.2%, positive 92.6%).

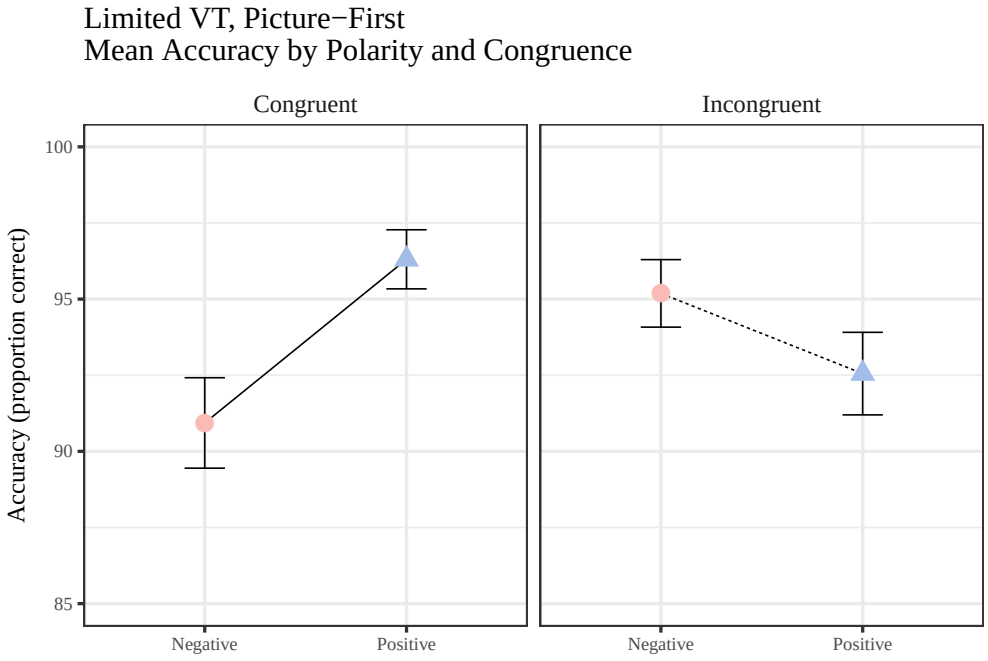


Figure 7: Experiment 1b: Limited VT; Picture-first. Accuracy (proportion correct) by condition. Error bars represent standard error of the mean.

Statistical power. A post-hoc power analysis was conducted on the RT model for Experiment 1b using the R-package *pwr* (Champely, 2018). The effect size associated with this model was 0.022 with a sample size of 1504 observations. The estimated power of the model was 0.99, which indicates a more than sufficiently powered analysis. Given the results of this power estimation, I conclude with good certainty that the effect seen were not false positives.

2.3.4 Discussion

In Experiment 1b, I found that sentences with *shorter* took longer to process than sentences with *taller*, supporting the decompositional analysis on which *shorter* is strictly more representationally complex than *taller*. Furthermore, evaluating false statements took longer than evaluating true statements, and there was a significant interaction between adjectival negation and congruence effects. These results are in line with the earlier results reported for previous Picture-to-Sentence matching tasks (cf. Clark and Chase 1972).

Taken together, the results of Experiments 1a and 1b lend reliability to the processing models proposed by Clark and Chase (1972). I will proceed specifically under the assumption that the Sentence-First model (which will be used again in Chapters 3 and 4) is methodologically valuable. Further, these results validate the value of interface transparency (Lidz et al. 2011) in allowing me to make explicit predictions about the relationship between morphosyntax and cognition.

Finally, I note that the designs of Experiments 1a and 1b were characterized by limited viewing time windows (200ms) for the picture stimuli. One question that emerges from this discussion is whether effects seen under this limited viewing time condition can be replicated under a less restrictive viewing time requirement. It is entirely possible that the statistical effects seen in 1a and 1b are, at least in part, the result of automatic cognitive operations (cf. Clark and Chase 1972). And if this is the case, one might expect these effects to be diminished under an experimental design in which a much larger viewing time window is permitted. In the section that follows, I consider precisely this open viewing time question.

2.4 Impact of viewing time on polarity and congruence effects

To address the question of whether the results of Experiments 1a and 1b can be replicated under a larger viewing time window, I designed two follow-up experiments—Experiments 1c (Sentence-First) and 1d (Picture-First)—whose results I report below. Both experiments were identical to their counterparts above, with the exception that the picture stimulus was

displayed for 5s or until keypress, as opposed to being displayed for only 200ms as in the previous experiments. To foreshadow, the results of these follow-up experiments fail to show significant effects of polarity, and a significant effect of congruence is found only in the picture-first task. Taken together, these results suggest that the effects reported above disappear when the task is not speeded, with statistical power and effect sizes likewise diminishing in suite.

2.4.1 Experiment 1c: Sentence-first, unlimited view time

REACTION TIMES

	χ^2	p-value	β
POLARITY	0.60	$p > 0.1$	0.04
CONGRUENCE	3.63	$p > 0.1$	0.13
POLARITY*CONGRUENCE	0.96	$p > 0.1$	0.10

ACCURACY

	χ^2	p-value	β
POLARITY	0.41	$p > 0.1$	0.34
CONGRUENCE	0.82	$p > 0.1$	-0.26
POLARITY*CONGRUENCE	0.24	$p > 0.1$	0.25

Table 3: Summary of model comparison results for Experiment 1c (Unlimited VT, Sentence-first)

Reaction times. Under the unlimited viewing time condition, participants in the sentence-first paradigm showed no evidence of taking longer to evaluate sentences with *shorter* than with *taller*. This was reflected in a lack of a main effect of POLARITY (means: negative 6.46, positive 6.49, $\beta = 0.34$, $\chi^2 = 0.41$, $p > 0.1$): RTs in the negative conditions were not statistically different from RTs in the positive conditions (means, in ms: negative 921.15ms, positive 931.26ms).

Additionally, participants took no longer to reject false statements than to accept true statements. This was reflected in a lack of main effect of CONGRUENCE (means: congruent

6.54, incongruent 6.41, $\beta = 0.13$, $\chi^2 = 3.63$, $p > 0.1$): a statement's truth or falsity with respect to its accompanying picture had no significant impact on associated RTs (means, in ms: congruent 960.54ms, incongruent 892.14ms).

Analyses revealed no interaction between POLARITY and CONGRUENCE ($\beta = 0.10$, $\chi^2 = 0.96$, $p > 0.1$); there was no difference in mean RTs in the negative versus positive congruent conditions (means: negative 6.50, positive 6.59; means, in ms: negative 940.01ms, positive 980.70ms). Such was also the case in the negative and positive incongruent conditions (means: negative 6.42, positive 6.40; means, in ms: negative 902.30ms, positive 881.98ms).

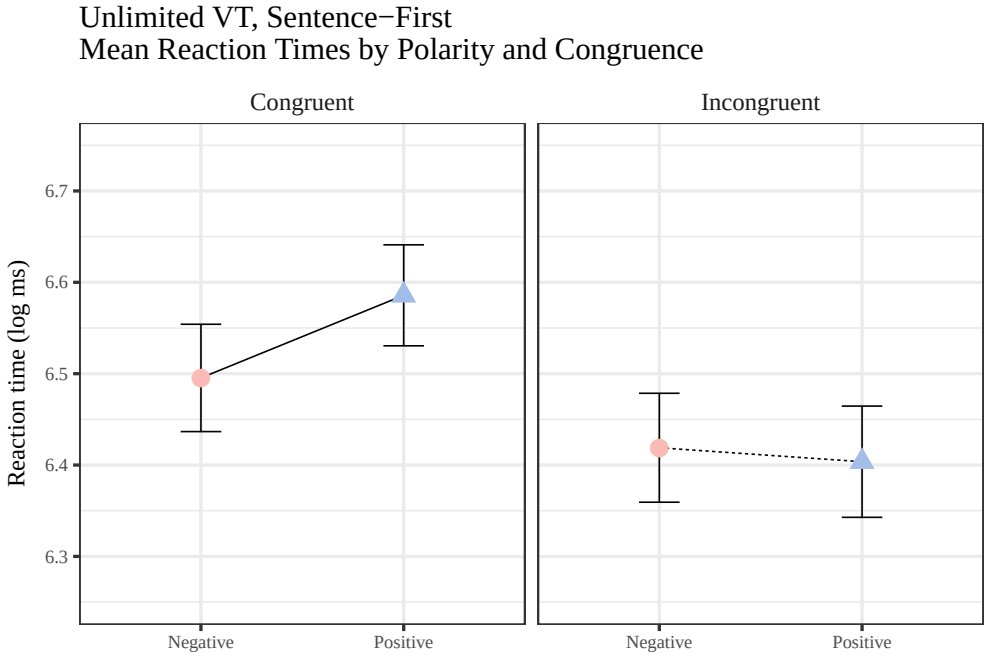


Figure 8: Experiment 1c: Unlimited VT; Sentence-first. Reaction times (log) by condition. Error bars represent standard error of the mean.

Response accuracy. Response accuracy for the unlimited viewing time variant of the sentence-first paradigm largely mirrored patterns seen in the reaction time data. Participants' response accuracy was no significantly worse for sentences with *shorter* than for those with *taller*. This was reflected in the lack of effect of POLARITY on mean response accuracy

(means: negative 94.8%, positive 96.2%, $\beta = -0.34$, $\chi^2 < 0.41$, $p > 0.1$).

Additionally, participants were no less accurate at rejecting false statements than at accepting true statements. I found no effect of CONGRUENCE on mean response accuracy (means: congruent 94.9%, incongruent 96.1%, $\beta = -0.26$, $\chi^2 = 0.82$, $p > 0.1$): whether a statement was true or false given its accompanying picture made no significant difference to verification accuracy.

Analyses revealed no interaction between POLARITY and CONGRUENCE ($\beta = 0.25$, $\chi^2 = 0.24$, $p > 0.1$); there was no difference in mean response accuracy in the negative versus positive congruent conditions (means: negative 93.9%, positive 95.9%). Such was also the case in the negative and positive incongruent conditions (means: negative 95.7%, positive 96.5%).

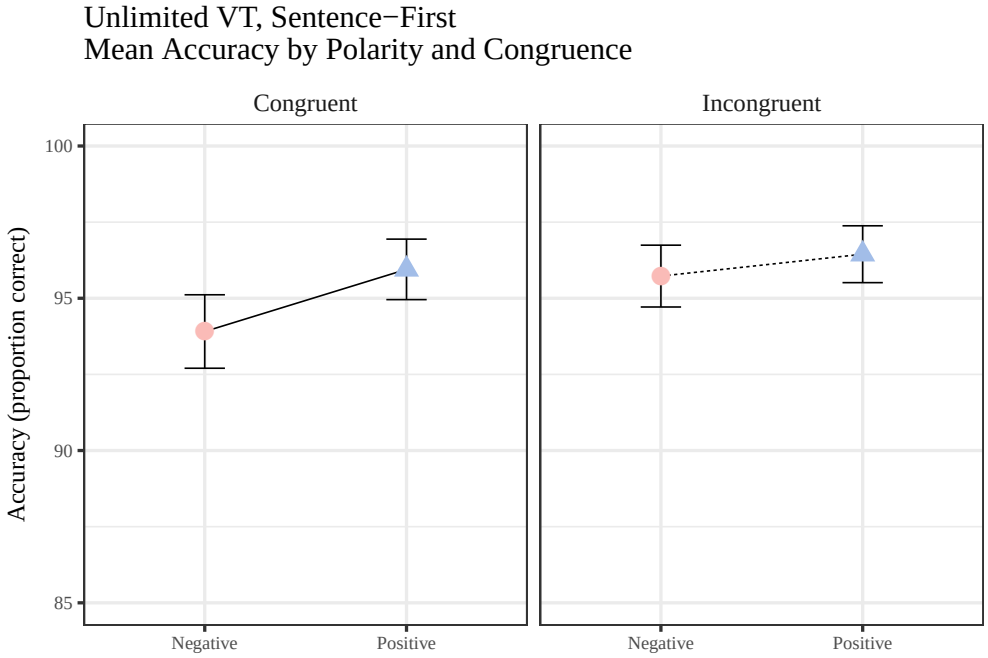


Figure 9: Experiment 1c: Unlimited VT; Sentence-first. Accuracy (proportion correct) by condition. Error bars represent standard error of the mean.

Statistical power. A post-hoc power analysis was conducted on the RT model for Experiment 1b using the R-package *pwr* (Champely, 2018). The effect size associated with this

model was 0.007 with a sample size of 1335 observations. The estimated power of the model was 0.53, which falls substantially below the conventionally desired statistical power of 0.8. However, it is noteworthy that no significant effects of POLARITY or CONGRUENCE were found, thus I use this power analysis only as an indicator of the categorically different nature of behavioral responses under the unlimited viewing time condition.

2.4.2 Experiment 1d: Picture-first, unlimited view time

REACTION TIMES

	χ^2	p-value	β
POLARITY	0.38	$p > 0.1$	-0.05
CONGRUENCE	7.36	$p = 0.007$	0.23
POLARITY*CONGRUENCE	2.04	$p > 0.1$	0.17

ACCURACY

	χ^2	p-value	β
POLARITY	0.41	$p > 0.1$	0.16
CONGRUENCE	0.82	$p > 0.1$	-0.26
POLARITY*CONGRUENCE	0.24	$p > 0.1$	-0.04

Table 4: Summary of model comparison results for Experiment 1d (Unlimited VT, Picture-first)

Reaction times. Under the unlimited viewing time condition, participants in the picture-first paradigm showed no evidence of taking longer to evaluate sentences with *shorter* than with *taller*. This was reflected in a lack of main effect of POLARITY (means: negative 6.06, positive 6.02, $\beta = -0.05$, $\chi^2 = 0.38$, $p > 0.1$): RTs in the negative conditions were not statistically different from RTs in the positive conditions (means, in ms: negative 664.17ms, positive 662.48ms).

Interestingly, participants took significantly longer to accept true statements than to reject false statements. This was reflected in a strong main effect of CONGRUENCE (means: congruent 6.16, incongruent 5.92, $\beta = 0.23$, $\chi^2 = 7.36$, $p = 0.007$): a statement's truth or falsity

with respect to its accompanying picture had a non-trivial impact on associated RTs, though in the opposite direction as expected based on limited viewing-time conditions (means, in ms: congruent 716.54ms, incongruent 610.23ms).

Analyses revealed no interaction between POLARITY and CONGRUENCE ($\beta = 0.17$, $\chi^2 = 2.04$, $p > 0.1$); there was no difference in mean RTs in the negative versus positive congruent conditions (means: negative 6.14, positive 6.18; means, in ms: negative 679.42ms, positive 753.18ms). Such was also the case in the negative and positive incongruent conditions (means: negative 6.18, positive 5.86; means, in ms: negative 649.17ms, positive 571.78ms).

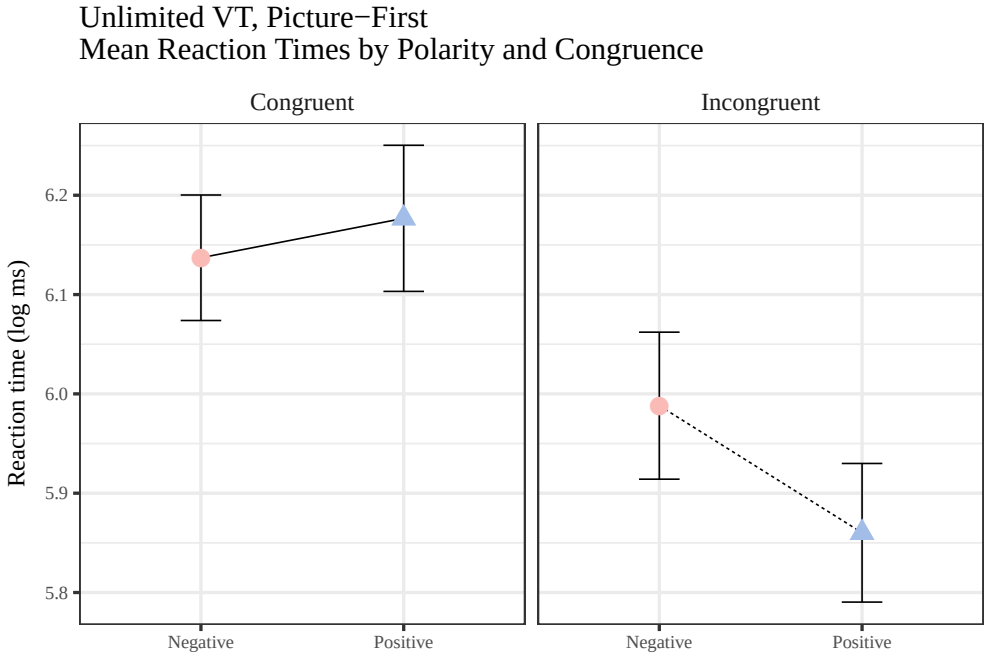


Figure 10: Experiment 1d: Unlimited VT; Picture-first. Reaction times (log) by condition. Error bars represent standard error of the mean.

Response accuracy. Response accuracy for the unlimited viewing time variant of the picture-first paradigm largely mirrored patterns seen in the response accuracy data for the unlimited VT variant of the sentence-first paradigm. Participants' response accuracy was no significantly worse for sentences with *shorter* than for those with *taller*. This was reflected in

the lack of effect of POLARITY on mean response accuracy (means: negative 95.7%, positive 96.2%, $\beta = -0.16$, $\chi^2 < 0.41$, $p > 0.1$).

Additionally, participants were no less accurate at rejecting false statements than at accepting true statements. I found no effect of CONGRUENCE on mean response accuracy (means: congruent 95.8%, incongruent 96.1%, $\beta = -0.26$, $\chi^2 = 0.82$, $p > 0.1$): whether a statement was true or false given its accompanying picture made no significant difference to verification accuracy.

Analyses revealed no interaction between POLARITY and CONGRUENCE ($\beta = -0.05$, $\chi^2 = 0.24$, $p > 0.1$); there was no difference in mean response accuracy in the negative versus positive congruent conditions (means: negative 95.5%, positive 96.1%). Such was also the case in the negative and positive incongruent conditions (means: negative 95.8%, positive 96.4%).

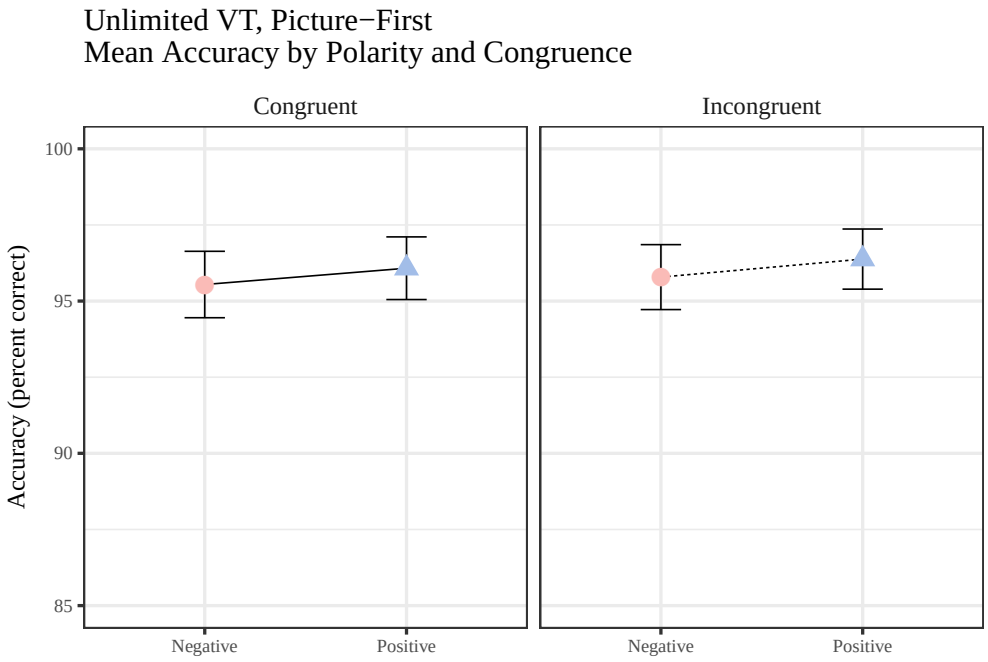


Figure 11: Experiment 1d: Unlimited VT; Picture-first. Accuracy (proportion correct) by condition. Error bars represent standard error of the mean.

Statistical power. A post-hoc power analysis was conducted on the RT model for Exper-

iment 1b using the R-package *pwr* (Champely, 2018). The effect size associated with this model was 0.007 with a sample size of 1335 observations. The estimated power of the model was 0.53, which falls substantially below the conventionally desired statistical power of 0.8. However, it is noteworthy that no significant effects of POLARITY or CONGRUENCE were found, thus I use this power analysis only as an indicator of the categorically different nature of behavioral responses under the unlimited viewing time condition.

2.5 General discussion

In this chapter, I set out to link the decompositional account of *shorter* with processing by adopting the hypothesis that each linguistic unit is linked explicitly to a cognitive operation. In two experiments—a sentence-to-picture verification task and a picture-to-sentence verification task—I found that *shorter* comparatives took longer to process than *taller* comparatives, in line with the decompositional analysis that posits the former is representationally more complex than the latter. In addition, I also found that verification took longer when the sentence and the picture were mismatched (i.e. incongruent). Both findings were as expected given previous results from Clark and Chase (1972), and in line with predictions made by the decompositional account.

This study leaves open the question of whether our data *could not* have been accounted for by positing a non-decompositional analysis in the first place, for instance that posited by Kennedy (2001). Given the other assumptions I made, such a view would hearken back to the early cognitive psychology literature, in which what was responsible for the additional processing cost of negation was some sort of linguistic 'negative feature'—in this case a negative lexical meaning. While such an approach could be made compatible with our findings, it would do so at the cost of transparency, at least in English. For example, the decompositional analysis renders the mapping from (abstract) syntax to conceptualization in English consonant with other languages, e.g. Hixkaryana, where the antonym of an adjective like *long* is formed by two pieces, i.e. *kawo-hra* ('long-not'; Bobaljik 2012). On the alternative

view, the mapping would be transparent in a language like Hixkaryana, but requires a detour through the lexicon in English.

As a final note, the effects of congruence were examined throughout this chapter in experiemnts where effects were expected based on previous work by Clark and Chase (1972). While I will continue to report congruence effects in subsequent experiments throughout the remainder of this dissertation, I will no longer make explicit hypotheses about the effects of congruence. The experiments reported in this chapter have been sufficient to establish a proof of concept.

In the chapter that follows, I extend the analysis presented here to *less* comparatives, and ask whether similar behavioral evidence can be found in support of Heim's (2008) modification to Büring's (2007a) proposal.

3 Additivity of Negation in English Categorizing Comparatives

Thus far in this dissertation, the discussion of adjectival comparatives has been principally concerned with the processing of synthetic comparatives (e.g. *taller*, *shorter*). The results of the previous chapter suggest that Büring's decompositional analysis of negative adjectival comparatives is consistent with behavioral evidence under the operationalization proposed there. I now seek to extend this analysis to analytic comparatives with *less*—whose putative decompositional analysis also contains LITTLE—and ask whether the behavioral evidence is consistent with this decomposition here as well. Importantly, I ask whether there is an additive effect of processing multiple instances of LITTLE in a single comparative statement, in this case *less short*.

This chapter presents two experiments, Experiments 2a and 2b, whose results suggest not only that LITTLE may introduce extra processing costs (as evidenced by the RT data), but also that the effects of introducing such a negative operation may be monotonically additive. I set out to test this hypothesis by examining analytic comparatives with *short* (e.g. *less tall/short*) in contrast to (putatively) extensionally-equivalent analytic comparatives with *more* (e.g. *more tall/short*). To preview: I find evidence that *less*-comparatives are subject to different processing constraints than *more* comparatives.

A further question that this chapter begins to explore is the question of whether *short* and *less* decompose in terms of the same LITTLE (as on Büring's analysis) or in terms of the variants LITTLE* and LITTLE (as on Heim's analysis), where the two differ in terms of their scopal behavior. While I will not be able to resolve this question here, I take important steps that delimit how that question might be resolved in future research.

3.1 Background and motivation

This chapter considers the first of two further pieces of morphosyntax that could appear in comparative sentences, and which could reveal themselves in downstream processing. The first instance involves cases where two LITTLEs might appear (e.g. *less short*), and the other a case where a silent morpheme κ blocks regular synthetic comparative formation. The latter will be the focus of Chapter 4.

The background to follow outlines a modification of Buring's proposal as it was laid out in Chapter 2, as spelled out by Heim (2008). Crucial to Heim's analysis is the presence of a silent morpheme LITTLE* in the decomposition of *less*-comparatives (e.g. *the tree is less tall than the house*). While this morpheme is essentially identical in function to Buring's LITTLE, its scopal properties differ. These differences will be described and motivated below.

This section concludes with a set of hypotheses and predictions for Experiments 2a and 2b. Behavioral predictions are spelled out from the decompositional analyses I outline here, in much the same fashion as they were in Chapter 2.

3.1.1 Heim's (2008) decomposition of *less*-comparatives

Under a lexical negation view of antonymy, denotations of *tall* (19a) and *short* (19b) are related by the operation of predicate negation, but there is no meaningful part of the syntactic representation of the negative adjective *short* that encapsulates this operation. Both adjective pairs are morphosyntactically indivisible, and stand in no representationally transparent entailment relation with each other. And while native speakers of English do have the intuition *tall* and *short* stand in an antonymous relation to one another, this intuition is not a strict matter of logic, per se.

(19) a. $[\![\text{TALL}]\!] = \lambda x.(0, H(x))$

b. $[\![\text{SHORT}]\!] = \lambda x.(H(x), \infty)$

This traditional lexical view of antonymy illustrated in (19) contrasts with what Heim (2008) refers to as a syntactic negation theory of antonymy. On this view, there is no mean-

ingful listing of an item like $\alpha-$ (e.g. *short*) in the lexicon. Instead, the lexicon generates the surface form *short* by spelling out a collocation of two meaningful units, one of which is the same as what spells out *tall*, and the other is a sort of negation operator. The proposal Heim offers (outlined below) builds on Büring's (2007a; 2007b) proposals as discussed in the previous chapter. I briefly review this proposal here before moving to Heim's modification to it.

Büring's (2007a) motivates his decompositional analysis of negative adjectival comparatives by considering what Kennedy (2001) identifies as "cross-polar anomalies" (20). One curious aspect of sentences like (20) is that their unacceptability is not pre-theoretically obvious: degrees of shortness and tallness are both instances of the same measure (height), and intuitively seem commensurable. Yet examples like (20) seem anomalous.

(20) ? The tree is shorter than the house is tall.

Kennedy 2001 proposed a constraint to account for such anomalies. Following Kennedy, a positive adjective like *tall* relates an individual to a positive length, while a negative adjective relates an individual to a negative length. A positive length is an interval (or set of degrees) that begins at 0 and extends to the object's height, while a negative length is an interval that begins just above the object's height and proceeds to infinity. The resulting lack of any sort of subset relation between the two intervals precludes any sort of comparison between positive and negative degrees, even if these are applied to the same scale.

(21) The tree is shorter than the house is wide.

At issue, however, is that there are instances of sentences such as (21) which are judged by native speakers of English as both well-formed and interpretable, and yet which clearly stand in violation of Kennedy's constraint. To account for these cross-polar nomalies, Büring offers a proposal whose essential thesis is that cross-polar nomalies are interpreted as less $\alpha+$ rather than more $\alpha-$. Building on proposals of Rullmann 1995 and Heim 2006, Büring capitalizes on the possibility of decomposition by analyzing sentences like (21) as involving a less-than comparison between positive degrees of tallness and wideness, rather than a

greater-than comparison between negative and positive degrees (which would violate Kennedy's constraint). On Büring's account, a lexicon akin to (22) is proposed, which contains nothing corresponding to a $\alpha-$; instead, as the subsequent spell-out rules in (23) indicate, *short* is non-atomic and generated at spell-out.

(22) a. $[\![\text{TALL}]\!] = \lambda x.(0, H(x))$

b. $[\![\text{ER}]\!] = \lambda f.\lambda A.\lambda B.f(B) \subset f(B)$

c. $[\![\text{LITTLE}]\!] = \lambda A.\neg A$

(23) a. TALL $>$ *tall*

b. ER $>$ *-er*

c. ER LITTLE $>$ *less*

d. LITTLE TALL $>$ *short*

Given the spell-out rules in (23), Büring notes that (21) has two possible spell-outs depending how the string ER LITTLE TALL is bracketed. Under one possible bracketing, [ER [LITTLE TALL]], the associated logical form is anomalous. However, the alternative bracketing, [[ER LITTLE] TALL], results in an interpretable spell-out (i.e. the *less* $\alpha+$ reading). As long as one of these two possible spell-outs respects Kennedy's constraint, Büring anticipates that the sentence will not be anomalous.

Against this backdrop, Heim (2008) observes that the lexical entries in (22) and spell-out rules in (23) engender a morphology-semantics mismatch in which the string ER LITTLE TALL can potentially spell out to *less tall* or *shorter*. Büring (2007a) does indicate that, while both bracketings, [[ER LITTLE] TALL] and [ER [LITTLE TALL]], may surface as either *less long* or as *shorter*, there is a preference for the former bracketing to surface as *less long*, and for the latter bracketing to surface as *shorter*. Precisely what engenders these markedness relations and spell-out preferences is left unclear in Büring's analysis. For Heim's modification to be motivated, however, it is sufficient to note that on Büring's analysis, both spell-outs are possible, modulo the ascribed markedness preferences.

To motivate her reprise of Büring's analysis, Heim invites us to consider the following scenario: two individuals, Polly and Larry, are both supposed to be in Boston by 8:00 PM at the latest. It is now 5:30, and Polly is just setting out from Providence, RI, while Larry is leaving from New Haven, CT (twice as far from Boston as Providence). Thus, of the statements in (24), only (24a) and (24b) hold true of these circumstances.

(24) a. Larry needs to drive faster than Polly needs to drive.

b. Polly needs to drive less fast than Larry needs to drive.

c. ? Polly needs to drive more slowly than Larry needs to drive.

Polly indeed needs to drive less fast than Larry because she does not need to cover as much distance. And while (24b) is a paraphrase of (24a), (24c) is not a paraphrase of (24a) and (24b)—it claims something quite different and is in fact false of the circumstances described. For (24c) to hold, there would have to be some penalty or disadvantage to arriving early. But this is not the case here: Polly may drive more slowly, but she doesn't need to. As Heim points out here, we would expect (24c) to share the same true reading as (24b) if Buring's analysis is adequate. But (24c) does not. Buring's analysis thus overgenerates.

To make sense of cross-polar nomalies like (24), Heim offers the following modification, as exemplified in the following lexicon and spell-out rules.

(25) a. $[\![\text{TALL}]\!] = \lambda x.(0, H(x))$

b. $[\![\text{ER}]\!] = \lambda f.\lambda A.\lambda B.f(B) \subset f(B)$

c. $[\![\text{LITTLE}]\!] = \lambda A.\neg A$

d. $[\![\text{LITTLE*}]\!] = \lambda d.\lambda A.d \notin A$

(26) a. ER LITTLE* > *less*

b. LITTLE TALL > *short*

While (25a) and (25b) correspond to the usual entries, (25c) is Büring's LITTLE (cf. 22c). The fourth element in Heim's lexicon is what she refers to as a 'scopally-mobile' variant of

LITTLE, LITTLE*. On Heim's account, LITTLE is scopally-fixed, and this allows her account to capture the observation that $\alpha-$ is bound to scope below any modal, thus ruling out the reading in (3).

As suggested by the term 'scopally-mobile,' Heim's LITTLE* exhibits different behavior from LITTLE. In general, when *less* surfaces, it is underlyingly LITTLE* and the negation in this interacts scopally with within-clause modals (see also Heim 2006; cf. Rullmann 1995; Büring 2007a).

This analysis will be further refined in the following chapter by considering Embick's analysis of alternations like *taller* and *more tall*. For present purposes, the principal take-away from Heim's 2008 modification of Büring's (2007a; 2007b) analysis is the decomposition of e.g. less tall into [[ER LITTLE*] TALL], and by extension, the decomposition of less short into [[ER LITTLE*] [LITTLE TALL]]. The experiments reported below test the predictions of Heim's decompositional account of less-comparatives, and serve as a natural extension of the findings reported in Chapter 2.

3.1.2 Hypotheses and predictions

In Chapter 2, explicit predictions were made concerning the response times associated with the comparisons involving *taller/shorter*. Such explicit predictions were made possible by espousing the assumption that the mapping between morphosyntax and non-linguistic cognition is transparent. Here, I again appeal to the Interface Transparency Thesis (Lidz et al. 2011) in spelling out my expectations for Experiments 2a and 2b.[14]

In Experiments 1a and 1b, I assumed a particular linking between morphosyntactic decomposition and operational parameters affecting response times in a sentence-first experiment paradigm. The parameters I assumed are repeated below in (27). This algorithmic model predicted that the presence of LITTLE in *shorter* would give rise to an RT hit not

[14] I will also be interpreting additive reaction time effects as indicative of discrete processing stages, with the caveat that this kind of interpretation is potentially problematic (see Footnote 8).

present in *taller* (cf. parameter t_b). The parameter t_a was associated with the cost of CONGRUENCE—whether the statement was true of the associated picture. The final parameter, t_0, was simply a baseline measure.

(27) Parameters affecting response latency

a. t_0 - 'baseline processing parameter'

b. t_a - cost of 'falsification' (cf. CONGRUENCE)

c. t_b - cost of 'linguistic negation' (LITTLE)

In contrast to Experiments 1a and 1b, which featured only the synthetic comparatives *taller/shorter*, Experiments 2a and 2b the analytic counterparts with *more tall/short*, as well as *less tall/short*, as shown in (28), along with the relevant parts of their putative decompositions (ignoring, for the moment, the possible difference of LITTLE versus LITTLE*).

(28) a. A is more tall than B. $\approx$ ... ER TALL ...

b. A is more short than B. $\approx$ ... ER LITTLE TALL ...

c. A is less tall than B. $\approx$ ... ER LITTLE TALL ...

d. A is less short than B. $\approx$ ... ER LITTLE LITTLE TALL ...

Building on the operationalization given in (27), I propose the set of parameters shown in (29). Here, in addition to the parameters given in (27), I have added an additional parameter t_c corresponding to the hypothetical cost of processing Heim's LITTLE*.

(29) Factors affecting response latency: *more short*

a. t_0 - 'baseline processing parameter'

b. t_a - cost of 'falsification' (cf. CONGRUENCE)

c. t_b - cost of 'linguistic negation' (LITTLE)

d. t_c - cost of 'comparative negation' (LITTLE*)

Deriving processing predictions from the operationalization of Heim's decomposition proposed in (29) is quite straightforward. While processing times for *taller* and *shorter*

are predicted to be analogous to the those reported in Chapter 2, the processing of *more*- and *less*-comparatives is predicted to behave somewhat differently. Here, crucially, I expect that the *less short* will have the longest response time of any comparative in (30) due to its putative containment of both LITTLE and LITTLE* (t_b and t_c).

(30) Predicted response latencies as additive sums over operations

a. *taller* $\rightarrow t_0 + t_a$

b. *shorter* $\rightarrow t_0 + t_a + t_b$

c. *more tall* $\rightarrow t_0 + t_a$

d. *more short* $\rightarrow t_0 + t_a + t_b$

e. *less tall* $\rightarrow t_0 + t_a + t_c$

f. *less short* $\rightarrow t_0 + t_a + t_b + t_c$

At this point, the reader may wonder whether any analysis predicts a distinction between *shorter*, *more short* and *less tall*, as all of these comparatives may be expected to have the same truth conditions. Heim's (2008) analysis leaves the precise operation underlying LITTLE* open for further precisification, positing both quantifier-raising and type-shifting as possible mechanisms. Whether one assumes either of these or some other possibility, what is clear is the need for scope manipulation to be invoked by LITTLE*. As such, I expect that the processing of LITTLE* (t_c) may involve a more demanding set of cognitive operations than the simple flipping of a truth value invoked by Büring's LITTLE. However, I note that this specific prediction—a difference between the processing of LITTLE and LITTLE*—will not be investigated directly until the post-hoc analysis following Experiments 2a and 2b.

In what follows, I set out to empirically test the predictions that (a) Heim's LITTLE* (i.e. comparative negation, or *less*) implicates the same sort of hit to response times associated with Büring's LITTLE in *shorter*; and (b) both comparative negation (*less*) and adjectival negation (*short*) stack additively in terms of the effect they have on processing times. Following Experiment 2b, I conduct a post-hoc analysis to investigate whether an empirical

distinction between LITTLE and LITTLE* is supported by the RT distributions associated with Experiment 2b.

3.2 Experiment 2a: Processing evidence for the decomposition of *less*

This section summarizes the materials, methods and results of an experiment designed to test for processing evidence of a decompositional account of *less* (following Heim 2008). To preview, I failed to find the anticipated effect of comparative negation. I consider the possibility that this failure to find an effect may have been due to strategizing on the part of the participants, and a lack of heterogeneity in the stimuli.

3.2.1 Design and stimuli

Design. Experiment 2a represents an aggregation of two separate 2x2x2 designs that were run concurrently. Both designs used the same experimental stimuli, and manipulated ADJECTIVAL NEGATION (positive, negative), COMPARATIVE NEGATION (positive, negative) and MORPHOLOGY (analytic, synthetic). However, while ADJECTIVAL NEGATION and COMPARATIVE NEGATION were manipulated within-subjects, MORPHOLOGY was manipulated between-subjects. In one group, participants saw the comparatives *taller* and *shorter* grouped with *more tall* and *more short*; in a second group, different participants saw the comparatives *taller* and *shorter* grouped with *less tall* and *less short*.

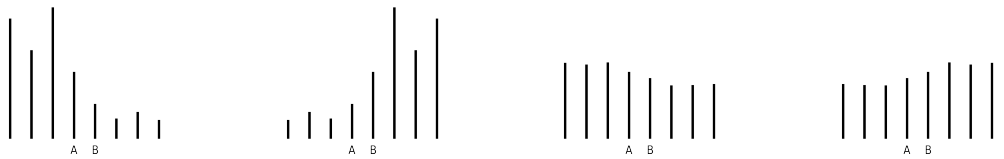
Stimuli. Sentence stimuli consisted of statements (e.g. *A is taller than B*) composed of the 6 comparatives given below in Table 5. The 4 analytic comparatives were split between two participant groups, i.e. (*more tall/short* and *less tall/short*), while the synthetic comparatives were tested within-group. Statement stimuli also varied by whether A or B came first in the matrix clause, i.e. *A/B is taller than B/A*.

Picture stimuli consisted of 40 images. Each image contained a total of 8 lines: 2 center

Comparative	Adjectival Negation	Comparative Negation	Morphology
taller	positive	—	synthetic
shorter	negative	—	synthetic
more tall	positive	positive	analytic
more short	negative	negative	analytic
less tall	positive	positive	analytic
less short	negative	negative	analytic

Table 5: Comparatives appearing in sentence stimuli in Experiment 2a. Analytic comparatives were between-participants: *more tall* and *more short* were used with one group, while *less tall* and *less short* were used with another.

lines labeled 'A' and 'B' with 3 context lines on either side, as exemplified in Figure 12. Lines 'A' and 'B' varied in relation to each other by a fixed set of proportions: (a) 0.5, (b) 0.75, (c) 0.83, (d) 0.875 and (e) 0.9. Half of the pictures were simply mirror images of the other half. For images in A was the taller line, the longest context line was always to the on the right side, and was always taller than a by one of the 5 aforementioned proportions. Similarly, the shortest context line was among the 3 context lines to the left of 'A' and varied by one of the aforementioned proportions in relation to 'B'. The remainder of the context lines were assigned random height values between the shortest context line height and B (for the shorter context lines), while the tallest context lines were assigned random values ranging from the height of A and the height of the tallest context line.



(a) (b) (c) (d)

Figure 12: Sample picture stimuli used in Experiment 2a.

Each item consisted of a combination of a single sentence (e.g. *A is more tall than B*) paired with a single picture. Due to the between-subjects nature of the experimental design, each participant saw 4 sentence types (2 analytic and 2 synthetic), each paired with

40 possible images, and each statement varied with 2 possible A/B orders. This yielded a total of 320 items or trials per participant.

3.2.2 Procedure

Following consent, participants saw a total 320 trials, each consisting of a statement (presented until keypress), followed by an image (presented either until response or for a maximum of 5s). To respond, participants pressed 'f' if they thought the statement was true of the associated image, 'j' if they thought it was not. The entire experiment took approximately 25 minutes to complete. Participants saw the following instructions upon beginning this study.

> Welcome to the experiment!
>
> In this study, you will be presented with 320 trials, each consisting of a sentence in the middle of the screen, followed by an image. Your task is to decide whether the sentence accurately describes the picture. You will have 5 seconds to make your decision before being advanced to the next trial.
>
> If the sentence accurately describes the picture, press the letter 'f' on the keyboard.
>
> If the sentence does not accurately describe the picture, press the letter 'j' on the keyboard.
>
> Ready? Press any key to begin.

40 undergraduate students at Northwestern University aged 18 years or older were recruited to participate in this experiment. Students received course credit in exchange for up to 1 hour of their time. No online participants were recruited for this study.

3.2.3 Analyses and exclusions

Data from each participant group were aggregated, and each analytic comparative was coded as a combination of two binary factors: COMPARATIVE NEGATION (positive, negative) and

ADJECTIVAL NEGATION (positive, negative). data from synthetic comparatives has been included in plots for visual comparison, but was not included in the statistical analyses here as it was not directly relevant to the empirical questions under consideration.

Following the same exclusionary procedure outlined in Chapter 2, participant responses with RTs less than 200ms were excluded. In addition, responses with RTs outside 2.5 standard deviations of the each participant's mean were excluded from analysis. This exclusionary procedure resulted in the of 1098 observations from the 12,800 observation collected, amounting to approximately 8.5% of the total data.

Below I report the results of linear mixed effects model comparisons with maximal random effects structures (i.e. including random intercepts and slopes by subject and item; best generalization for LMEMs, Barr et al. 2013). For all analyses, I used an orthogonal contrast coding scheme that assigned values of -.5 and .5 to each level of POLARITY and CONGRUENCE, respectively. The significance levels (p-values) that I report are derived from comparison of the maximal model in each case, against the same model minus the relevant parameter.

I conducted two separate linear mixed effects model comparisons on the log-transformed RT data. Analyses for RT measures were conducted on the log-transformed RT data to respect the normality assumptions of linear mixed effects models (Gelman and Hill 2007). I plot the log-transformed RT measure, and report both the results in both logRT and milliseconds (ms) for readability.

All analyses reported in this section were conducted using R's *lme4* package (Bates et al. 2015).

3.2.4 Results

The results of the LMEM comparisons conducted on Experiment 2a are summarized below in Table 6.

Participants took longer to evaluate sentences with *shorter* than with *taller*. This was

Factor	χ^2	p	β	Level means
ADJ	7.36	$p = 0.007$	0.23	*tall*: 6.91, *short*: 6.97
COMP	0.38	$p > 0.1$	0.01	*less*: 6.97, *more*: 6.94
ORDER	2.04	$p > 0.1$	0.17	AB: 6.94, BA: 6.94
ADJ:COMP	17.18	$p < 0.001$	0.17	*more tall*: 6.90, *less tall*: 6.98
				more short: 6.99, *less short*: 6.96

Table 6: Summary of model results and mean RTs (log ms) for Experiment 2a

reflected in a strong main effect of ADJECTIVAL NEGATION (means, in ms: short 1171.80ms, tall 1102.98ms; $\beta = -0.07$, $\chi^2 = 6.04$, $p = 0.014$) in the predicted direction: RTs in the negative conditions were longer than in the positive conditions (means, in ms: negative 686.01ms, positive 589.95ms).

Contra my expectations, participants took no longer on average to evaluate sentences with *less* than to evaluate sentences with *more*. No main effect of COMPARATIVE NEGATION was observed (means, in ms: less 1199.31ms, more 1125.54ms; $\beta = -0.06$, $\chi^2 = 0.11$, $p > 0.1$). RTs in the negative (*less*) conditions were no longer than those in the positive (*more*) conditions.

Also noteworthy was the unexpected interaction between ADJECTIVAL NEGATION and COMPARATIVE NEGATION: participants took longer to evaluate *less tall* than to evaluate *more tall* (means, in ms: less tall 1205.03ms, more tall 1076.75ms), but such a disparity was not evident in the processing of *less short* and *more short* (means, in ms: less short: 1193.61ms, more short: 1174.52ms). This was reflected in a significant interaction between the two polarity-based factors ($\beta = -0.25$, $\chi^2 = 17.18$, $p < 0.001$).

As noted above, I counterbalanced whether A or B was mentioned first in the sentence (ORDER). This difference was not expected to have any effect on responses, and analyses confirmed this expectation, as no statistical effect of ORDER was observed (means, in ms: AB 1135.74ms, BA 1138.94ms; $\beta = -0.01$, $\chi^2 = 0.02$, $p > 0.1$).

Statistical power. A post-hoc power analysis was conducted on the RT model for Experiment 2a using the R-package *pwr* (Champely, 2018). The effect size associated with this model

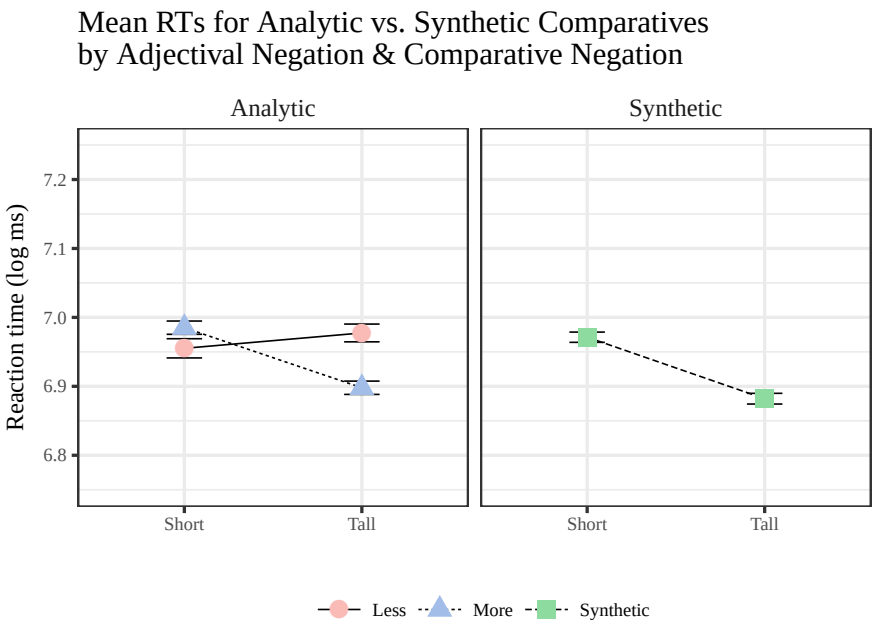


Figure 13: Experiment 2a: Mean RTs by adjectival negation and comparative negation. Error bars represent standard error of the mean.

was 0.0053 with a sample size of 5813 observations. The estimated power of the model was 0.99, which greatly exceeds the conventionally desired statistical power of 0.8 and indicates a more than sufficiently powered analysis. Given the results of this power estimation, I conclude with adequate certainty that the effects observed were not false positives.

3.2.5 Discussion

Experiment 2a set out to assess whether a decompositional analysis of *less*-comparatives is tenable given the RT evidence, in the same way that RT data was taken to support a decompositional analysis of negative adjectival comparatives (e.g. *short*) in Chapter 2. To summarize: while an expected main effect of adjectival negation (*short* vs *tall*) was found, I failed to find an anticipated main effect of comparative negation (*more* vs *less*). Here, I briefly speculate about the cause of this, in anticipation of Experiment 2b.

I suspect that this lack of anticipated effect may be due to a number of causes that can be addressed by modifying the design and stimuli of the experiment. First, it is possible that the

lack of an extensional difference between comparatives with *taller*, *more tall* and *less short* led participants to explicitly strategize a particular response to such sentences in the experiment, irrespective of their underlying representational differences. At 320 trials, as well, participants would have had ample opportunity to develop such a strategy. In addition, and due to the between-subjects nature of sentence stimuli in this experiment, participants only saw four different comparatives throughout the course of the entire experiment, which could have easily drawn participants' attention to (for example) truth conditional equivalences, and thus to task shortcuts. Meanwhile, the statistical interaction between ADJECTIVAL NEGATION and COMPARATIVE NEGATION suggests a potential effect of COMPARATIVE NEGATION that may have been obscured by this between-subjects design.

Thus, the next experiment addresses these considerations, while reprising the questions addressed with Experiment 2a.

3.3 Experiment 2b: Assessing additivity of adjectival and comparative negation

This section summarizes the materials, design and results of Experiment 2b, created as a direct follow-up to Experiment 2a to address possible task effects on the results of that experiment. And indeed, in contrast to those experimental results, here I find a main effect of both ADJECTIVAL NEGATION and COMPARATIVE NEGATION, with no interaction between these factors. I will use these results to suggest that a decompositional analysis of *less*-comparatives is tenable, given the apparently additive effects of morphemes with negative meanings.

3.3.1 Design and stimuli

Design. The design of Experiment 2b largely mirrored that of Experiment 2a, with a number of important exceptions. For 2a, comparative negation (*less* vs. *more*) was treated as a between-subject factor, while all other factors were within-subjects. Here, in 2b, all sentence

stimuli (summarized in Table 7) were manipulated within-subjects. In addition, baseline sentence stimuli (*A is tall/short*) were included in addition to analytic and synthetic comparatives. This increase in sentence stimuli created a potentially large number of trials for participants to complete. To offset this potential burden, picture stimuli were first divided into two balanced lists. Given the lack of effect of manipulating the order of comparators in Experiment 2a (that is, whether A or B was mentioned first in the sentence), that manipulation was dropped. Thus instead of 320 trials there were 160, and all sentence stimuli were presented in the form 'A is COMPARATIVE X than B'.

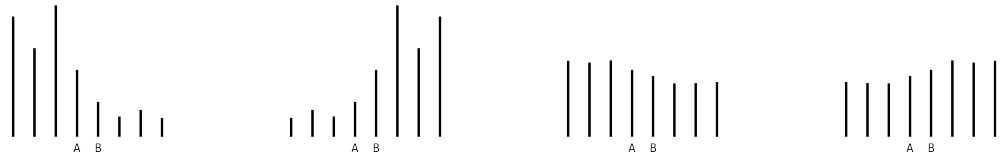
Stimuli. Sentence stimuli consisted of statements (e.g. *A is taller than B*) composed of the 6 comparatives and 2 adjectives (*tall* and *short*) given below in Table 5. In contrast to Experiment 2a, all participants saw all sentence stimuli; there was no between-subject manipulation of sentences.

Comparative	Adjectival Negation	Comparative Negation	Morphology
tall	positive	—	—
short	negative	—	—
taller	positive	—	synthetic
shorter	negative	—	synthetic
more tall	positive	positive	analytic
more short	negative	negative	analytic
less tall	positive	positive	analytic
less short	negative	negative	analytic

Table 7: Adjectives and adjectival comparatives appearing in sentence stimuli in Experiment 2b.

Picture stimuli consisted of the same 40 images used in Experiment 2a. Each image contained a total of 8 lines: 2 center lines labeled 'A' and 'B' with 3 context lines on either side, as exemplified in Figure 12. Lines A and B varied in relation to each other by a fixed set of proportions: (a) 0.5, (b) 0.75, (c) 0.83, (d) 0.875 and (e) 0.9. Half of the pictures were simply mirror images of the other half. For images in which A was the taller line, the longest context line was always to the on the right side, and was always taller than a by one of the 5 aforementioned proportions. Similarly, the shortest context line was among the 3 context

lines to the left of A and varied by one of the aforementioned proportions in relation to B. The remainder of the context lines were assigned random height values between the shortest context line height and B (for the shorter context lines), while the tallest context lines were assigned random values ranging from the height of A and the height of the tallest context line.



(a) (b) (c) (d)

Figure 14: Sample picture stimuli used in Experiment 2b.

In contrast to Experiment 2a, participants in Experiment 2b were divided into 2 groups based on which list of picture stimuli they were presented. Each list of pictures was counterbalanced by the set of proportions given above. The full set of stimuli used for this experiment may be found in the Appendix.

3.3.2 Procedure

Following consent, participants saw a total 160 trials, each consisting of a statement (presented until keypress), followed by an image (presented either until response or for a maximum of 5s). To respond, participants pressed 'f' if they thought the statement was true of the associated image, 'j' if they thought it was not. The entire experiment took approximately 25 minutes to complete. Participants saw the following instructions upon beginning this study.

> In this experiment, you will see pictures of lines with different heights, and, for each picture, you will be asked to verify whether a sentence correctly describes that picture.

> This experiment will consist of 160 trials. In each trial, you will be shown a sentence, which you can look at as long as you need to. After pressing spacebar you will be shown a picture of lines, and asked to decide whether the sentence correctly describes the picture you see. You will decide by clicking yes or no, respectively. You will have 5 seconds to make your decision.
>
> This task should take no longer than 20 minutes, and you will be compensated $3.32 for completing it. Once you've completed this task, you'll be given a unique completion code. When you're ready, click below to proceed.

40 participants aged 18 years or older were recruited through Amazon's Mechanical Turk platform to participate in this experiment. Participants received $3.32 in exchange for an estimated 20 minutes to complete the experiment. (Up to one hour was allotted for participants to complete the HIT after accepting it.)

3.3.3 Analyses and exclusions

Data from each participant group were aggregated, and each analytic comparative was coded as a combination of two binary factors: COMPARATIVE NEGATION (positive, negative) and ADJECTIVAL NEGATION (positive, negative). In the plots below, I include the data from synthetic comparatives for visual comparison, but I did not include these data in the statistical analyses I report directly, as they were not directly relevant to my target empirical questions.

Following the same exclusionary procedure outlined in Chapter 2, participant responses with RTs less than 200ms were excluded. In addition, responses with RTs outside 2.5 standard deviations of the each participant's mean were excluded from analysis. This exclusionary procedure resulted in the of 1098 observations from the 12,800 observation collected, amounting to approximately 8.5% of the total data.

Below I report the results of linear mixed effects model comparisons with maximal random effects structures (i.e. including random intercepts and slopes by subject and item; best generalization for LMEMs, Barr et al. 2013). For all analyses, I used an orthogonal

contrast coding scheme that assigned values of -.5 and .5 to each level of POLARITY and CONGRUENCE, respectively. The significance levels (p-values) that I report are derived from comparison of the maximal model in each case, against the same model minus the relevant parameter.

I conducted two separate linear mixed effects model comparisons on the log-transformed RT data. Analyses for RT measures were conducted on the log-transformed RT data to respect the normality assumptions of linear mixed effects models (Gelman and Hill 2007). I plot the log-transformed RT measure, and report both the results in both logRT and milliseconds (ms) for readability.

All analyses reported in this section were conducted using R's *lme4* package (Bates et al. 2015).

3.3.4 Results

The results of the LMEM comparisons conducted on Experiment 2b are summarized below in Table 8.

Factor	χ^2	p	β	Level means
ADJ	33.13	$p < 0.001$	0.23	*tall*: 7.17, *short*: 7.27
COMP	73.08	$p < 0.001$	0.36	*more*: 7.15, *less*: 7.29
ADJ:COMP	0.05	$p > 0.1$	-0.02	*more tall*: 7.10, *less tall*: 7.25
				more short: 7.20, *less short*: 7.34

Table 8: Summary of model results and mean RTs (log ms) for Experiment 2b

Reaction times. Participants took longer to evaluate sentences with *short* than with *tall*. This was reflected in a strong main effect of ADJECTIVAL NEGATION ($\beta = 0.23$, $\chi^2 = 33.13$, $p < 0.001$) in the predicted direction: RTs in the negative conditions were longer than in the positive conditions (means, in ms: negative 1586.50ms, positive 1417.38ms).

In addition, participants took no longer on average to evaluate sentences with *less* than to evaluate sentences with *more*: a robust main effect of COMPARATIVE NEGATION was observed ($\beta = 0.35$, $\chi^2 = 73.08$, < 0.001). RTs associated with the negative (*less*) conditions

were longer than those in the positive (*more*) conditions (means, in ms: negative 1628.09ms, positive 1375.04ms).

Also noteworthy was a lack of interaction between ADJECTIVAL NEGATION and COMPARATIVE NEGATION ($\beta = -0.01$, $\chi^2 = 0.05$, $p > 0.1$). This suggests that processing effects of each type of negation are independent (means, in ms: *less short*: 1727.36ms, *more short*: 1445.50ms, *less tall*: 1529.40ms, *more tall*: 1305.36ms).

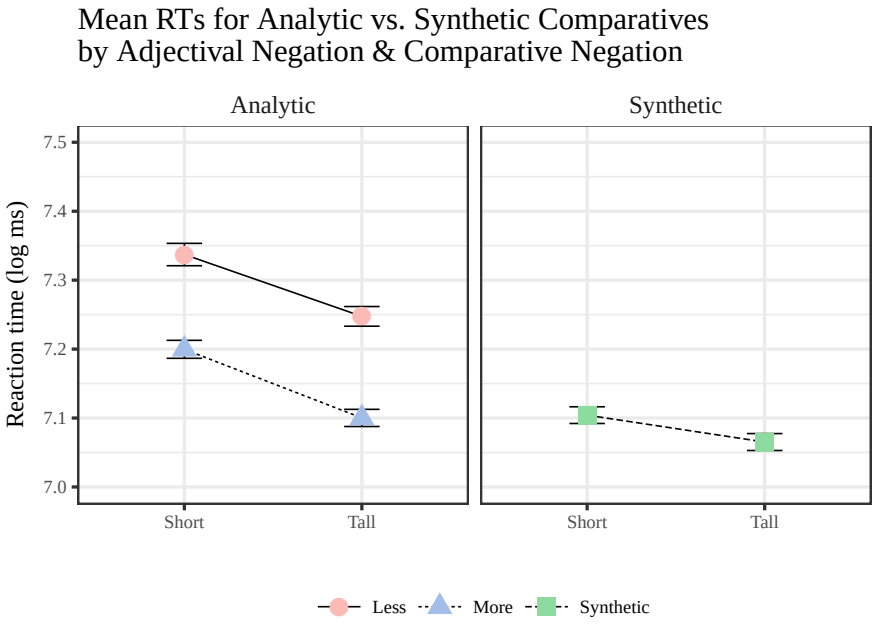


Figure 15: Experiment 2b: Mean RTs by adjectival negation and comparative negation. Error bars represent standard error of the mean.

Statistical power. A post-hoc power analysis was conducted on the RT model for Experiment 2a using the R-package *pwr* (Champely, 2018). The effect size associated with this model was 0.042 with a sample size of 3183 observations. The estimated power of the model was 0.99, which greatly exceeds the conventionally desired statistical power of 0.8 and indicates a more than sufficiently powered analysis. Given the results of this power estimation, I conclude with adequate certainty that the effects observed were not false positives.

3.3.5 Discussion

Experiment 2b was a direct follow-up to Experiment 2a. It attempted to address the question of whether task effects in Experiment 2a obscured what would otherwise be seen as an underlying additive effect of negation. In particular, we suspected that the lack of variety between sentences, the extensional equivalence between some of them, and the high number of trials lead participants to decide on a conscious strategy of evaluation that they deployed in lieu of the sorts of evaluation suggested directly by the comparative sentences.

To address whether this was the case, Experiment 2b was designed as an 'omnibus' experiment of sorts: each participant was exposed to every target comparative form, in addition to the corresponding positive forms (*tall* and *short*). This increased the variety of sentences and types of evaluations needed to successfully complete the task, which was predicted to eliminate the utility of strategizing, in tandem with a reduction in the total number of trials.

And indeed, the results of this experiment stand in contrast with those obtained for Experiment 2a. While I found no significant main effect of COMPARATIVE NEGATION, and a significant interaction between COMPARATIVE NEGATION and ADJECTIVAL NEGATION in Experiment 2a, in Experiment 2b I found a strong main effect of COMPARATIVE NEGATION and no meaningful interaction. In light of the reasonable possibility of strategizing, I believe that the design of Experiment 2b gives its results greater validity than those of Experiment 2b.

3.4 Post-hoc analyses

Here I describe the methods and results of two post-hoc analyses conducted on the data obtained from Experiment 2b. The first analysis seeks to find evidence for a behavioral distinction between Heim's (2008) scopally-fixed LITTLE and scopally-mobile LITTLE* via a series of RT distribution comparisons. The second analysis assesses the independent RT contribution of 'falsification time'—a factor which was explicitly manipulated in the experiments reported in Chapter 2, but not in Experiments 2a and 2b above.

3.4.1 Assessing behavioral evidence for Heim's scopally-mobile LITTLE

Building on Experiment 2b, are the RT distributions associated with the evaluation of the sentence stimuli more consistent with a Heim decomposition, or a Buring decomposition. Specifically, I ask whether RT components predicted to be associated with Büring's LITTLE consistent with each other, and are RT values associated with Heim's LITTLE* (and thereby inconsistent with Büring's LITTLE).

The set of equalities given in (31) outlines the differences in RT distributions as predicted under a Büring decompositional analysis are contrasted with those predicted under a Heim analysis. To unpack these predictions in greater detail: for both Büring and Heim, I expected that equalities (31a-31e) should be equivalent to one another, as these are all predicted to be due to the processing cost of LITTLE. Under Heim's analysis, equalities (31d-31e) are due to a scopally-mobile variant of LITTLE, which Heim denotes as LITTLE*. Crucially, as operationalized here, Heim's analysis will be consistent two equivalence classes: (31a-31c) and (31d-31e). Conversely, Büring's analysis will be taken as consistent with a single equivalence class subsuming all inequalities enumerated in (31).

(31) Differences in RT distributions as expected by Büring's vs. Heim's analyses

a. RT(*shorter*) − RT(*taller*) = RT(LITTLE) Büring

b. RT(*more short*) − RT(*more tall*) = RT(LITTLE) Büring

c. RT(*less short*) − RT(*less tall*) = RT(LITTLE) Büring

d. RT(*less tall*) − RT(*more tall*) = RT(LITTLE*) Heim

e. RT(*less short*) − RT(*more short*) = RT(LITTLE*) Heim

This analysis was conducted by computing the difference in RTs between each pair in (31). For example, to calculate the difference in (31a), I created two subsets of the data: one corresponding with RTs for the comparative *shorter*, and a second vector corresponding to RTs for the comparative *taller*. I then subtracted the first vector corresponding with *shorter* from the latter vector corresponding with *taller* to obtain a difference vector. To ensure that

vectors were cross-comparable—i.e. that, at any index, both vectors contained the observation for the same picture stimulus from the same participant—all observations in the data were ordered by picture stimulus and by participant. After carrying out vector subtraction, I then took the absolute value of each value in the resulting vector.

Once all of the difference vectors were obtained for every item in (31), I then cross-compared all of the difference distributions to see which vectors were (dis)similar to each other. To conduct this comparison, I performed a total of 10 two-samples Kolmorgorov-Smirnov (K-S) tests. The K-S test is a non-parametric test of the equality of distributions, and was chosen in this case due to the continuous, one-dimensional nature of the RT data involved. The results of these tests are given in Table 9.

	MS − MT	LS − LT	LS − MS	LT − MT
S − T	0.393	0.000*	0.000*	0.010*
MS − MT		0.000*	0.000*	0.071
LS − LT			0.745	0.026*
LS − MS				0.009*

Table 9: Results of 10 two-samples Kolmogorov-Smirnov tests. Values denote probability values, with asterisks denoting significant differences between the paired distributions.

The results of this cross-comparison suggest neither Büring's nor Heim's decompositional analysis, as operationalized in (31), is supported by the equivalence relations between difference distributions. Under my operationalization of Büring's decomposition, I expected all distributions to be equivalent, which would have been borne out statistically by non-significant p-values for every comparison. Under my operationalization of Heim's decomposition, I expected the differences in (31d) and (31e) to form one equivalence class (associated with LITTLE*), and the differences in (31a-31c) to form a separate equivalence class (corresponding to LITTLE). Neither prediction was born out by the distributional data. However, the results do suggest that more operations may be involved that what I have considered thus far. This consideration will be further investigated in the next chapter.

3.4.2 Assessing independent impact of falsification time

In Chapter 2, I proposed a set of putative parameters affecting response latency in the evaluation of comparative statements. In the present chapter, I added an additional parameter—one corresponding to a cost of 'comparative negation'—and found processing evidence associated with its predicted impact on verification. In the analyses associated with Experiments 2a and 2b, however, I did not assess the cost of 'falsification' (labeled in Chapter 2 as CONGRUENCE) in my RT models.

Consider the set of parameters in (32), which was introduced earlier in this chapter. In Experiment 2b, I found independent (additive) costs of both linguistic negation (*short*) and comparative negation (*less*)—denoted t_b and t_c in (32).

(32) Factors affecting response latency: *more short*

a. t_0 - 'baseline processing parameter'

b. t_a - cost of 'falsification' (CONGRUENCE)

c. t_b - cost of 'adjectival negation' (LITTLE)

d. t_c - cost of 'comparative negation' (LITTLE*)

To ascertain whether 'falsification' incurs a processing cost similar to that of linguistic and comparative negation, I conducted a linear mixed effects model comparison with the model used in Experiment 2b, with the addition of CONGRUENCE, coded orthogonally as 0.5 ('true') and -0.5 ('false'). The data modeled here was identical to the data used in the RT model for Experiment 2b, with the same exclusions applied. The results of this post-hoc analysis are summarized in Table 10.

In addition to the expected robust main effects of ADJECTIVAL NEGATION (denoted ADJ) and COMPARATIVE NEGATION (denoted COMP), I found a main effect of CONGRUENCE (denoted CONG) ($\beta = 0.20$, $\chi^2 = 36.69$, $p < 0.001$). In terms of interactions between these three factors, I failed to find significant interactions between ADJECTIVAL NEGATION and COMPARATIVE NEGATION ($\beta = -0.02$, $\chi^2 = 0.07$, $p > 0.1$), ADJECTIVAL NEGATION

Factor	χ^2	p	β
ADJ	37.09	$p < 0.001$	0.22
COMP	77.98	$p < 0.001$	0.34
CONG	36.69	$p < 0.001$	0.20
ADJ:COMP	0.07	$p > 0.1$	−0.02
ADJ:CONG	2.41	$p > 0.1$	0.10
COMP:CONG	1.12	$p > 0.1$	0.07

Table 10: Summary of model results and mean RTs (log ms) for Experiment 2a

and CONGRUENCE ($\beta = 0.10$, $\chi^2 = 2.41$, $p > 0.1$), and COMPARATIVE NEGATION and CONGRUENCE ($\beta = 0.07$, $\chi^2 = 1.12$, $p > 0.1$).

Taken together, the results of this post-hoc analysis (visualized in Figure 16) point to independent (additive) costs associated with the processing of adjectival negation, comparative negation, and falsification.

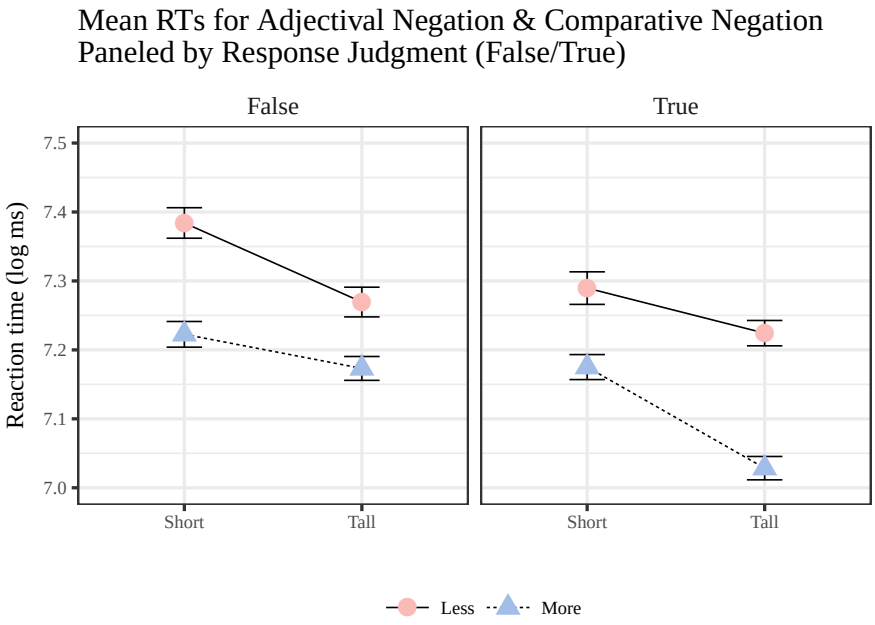


Figure 16: Mean RTs by adjectival negation and comparative negation, paneled by whether the participant responded 'false' or 'true' for each item. Error bars represent standard error of the mean.

3.5 General discussion

In this chapter, I set out to assess whether, like in Chapter 2, behavioral evidence could be found in support of a decompositional analysis of *less* comparatives (following Heim 2008). To this end, I introduced Heim's (2008) modification of Büring's (2007a) proposal, and made explicit predictions about the RT distributions I expected to be associated with the interpretation of synthetic and analytic comparatives. I then presented the results of two experiments designed to probe behavioral evidence for an effect of comparative negation.

The results of Experiment 2a were not quite as expected—specifically, I failed to find an anticipated main effect of COMPARATIVE NEGATION, although a significant interaction between COMPARATIVE NEGATION and ADJECTIVAL NEGATION suggested this main effect may have been obscured. Suspecting design issues may have been at play in obscuring this effect, I designed a direct follow-up experiment—Experiment 2b—with several design choices made explicitly to address concerns believed to be at issue with Experiment 2b. This follow-up experiment yielded the expected results.

Finally, in the first of two post-hoc analyses, I probed whether RT evidence could be leveraged to support Heim's (2008) theoretical distinction between scopally-fixed LITTLE (putatively present in *short*) and scopally-mobile LITTLE* (putatively present in *less*). I carried out a series of RT distribution comparisons, and concluded that, while the RT data failed to align with a Büring decomposition, it was not wholly consistent with what would be predicted by a Heim decomposition. In the second post-hoc analysis, I assessed whether 'cost of falsification' could be individuated as an independent RT component in the data from Experiment 2b. This proved to be the case: I found a main effect of 'falsification' when this predictors was added to my RT models along with predictors coinciding with adjectival and comparative negation.

Taken together, these results suggest that a decompositional analysis of *less* comparatives is tenable given the behavioral evidence presented here. Moreover, the results of Experiment 2b and the second post-hoc experiment strongly suggest that adjectival negation, comparative

negation, and the cost of falsification can be individuated as independent, additive RT effects. In what follows, Chapter 4 continues the established vein of inquiry by seeking behavioral evidence for another silent piece of morphosyntax, proposed to explain the distribution and interpretation of evaluativity in English comparatives.

4 Assessing processing effects of evaluativity in categorizing comparatives

Chapter 3 left open an important question concerning the assumptions made thus far about the interpretational equivalence of synthetic and analytic comparatives. Indeed, recent theories on the distribution and interpretation of gradable adjectives (e.g. Rett (2015); Moracchini (2018)) have posited the role of silent morphosyntactic elements like EVAL to account for 'evaluativity' effects: that is, in some cases, we interpret a comparative as involving evaluation of the subject of the local clause relative to the context. Importantly here, such theories are relevant to consider in the interpretation of what Wellwood (2014) calls categorizing comparatives, which distributionally co-occur with analytic comparatives in English.

In this chapter, I extend my psycholinguistic investigation of decompositional approaches to adjectival comparatives to include consideration of evaluativity. First, I leverage the possibility of the presence of EVAL in a subset of the comparatives tested in Chapter 3 to conduct a post-hoc analysis on data presented there, to see whether including a modeling predictor corresponding to EVAL results in a better-fitting model. This initial exploratory phase offers promising prospects: my model indicates the inclusion of a predictor encoding the distribution of EVAL results in a substantially better fitting model than the final model assessed in Chapter 3.

Second, I present an experiment that allows me to test different semantic proposals for the evaluative interpretation where it occurs. Building on Solt & Gotzner 2012, I test a simple comparative hypothesis (just compare lengths with *tall* and *wide*), a rank order hypothesis (compare ordinal positions of individuals relative to their heights and widths), and a novel 'goodness-of-fit' hypothesis, which correlates judgments on a categorizing comparative like *A is more X than B is Y* with judgments of the independent likelihood that A is tall, or B is wide in the same contexts. The results of a three-way prediction accuracy comparison test suggest that goodness-of-fit offers a better predictive fit than the other proposals considered

here.

4.1 Background and motivation

In this background, I begin by offering labels to describe the types of comparatives that have appeared in Chapters 2 and 3. The conventional labels 'synthetic' and 'analytic' describe the morphological composition of these comparatives, but no mention has yet been made about how, e.g., the interpretation of *taller* should differ from that of *more tall*. I highlight the patterns of interpretation and implication that differentiate these two comparatives. I adopt Wellwood's (2014) labeling scheme in which the analytic variant are called 'categorizing' comparatives (e.g. *Al is more tall than Bill*) and the synthetic variant are called 'commensurating' or 'regular' comparatives.

I then review two proposals concerning the morphology and semantics of comparatives that are consistent the distribution of commensurating and categorizing comparatives. The first is that of Embick (2007), who proposes that morphological alternations characteristic of analytic/synthetic morphology are due to a covert morphosyntactic element, which he calls κ. The second proposal I examine is that of Rett (2015) (cf. Moracchini 2018), who proposes a silent morphosyntactic element called EVAL to account for 'evaluative' interpretations in English analytic comparatives. Both proposals will be reviewed in turn.

I conclude this section by outlining my hypotheses and predictions concerning two analyses: a post-hoc analysis revisiting the data Experiment 2b (from Chapter 3), and an analysis of belonging to an as-yet undiscussed experiment, Experiment 3.

4.1.1 Analytic and synthetic comparatives

In this dissertation so far, I have made reference to both synthetic comparatives (e.g. *taller*) and analytic comparatives (*more tall*), but I have done so without citing any meaningful differences of interpretation between them. In this section, I will begin open by discussing how differences in comparative structure pattern with consistent differences in interpretation

and implication.

In Chapters 2 and 3, I examined positive and negative adjectival comparisons involving synthetic comparatives such as those illustrated in (33) and (34). I will follow Wellwood (2014) in calling these 'commensurating' comparatives. Sentences containing such comparatives express greater-than relations between various sorts of degrees, understood as measures along referenced dimensions. Framed in this way, (33) expresses that Box A's height strictly exceeds Box B's height, while (34b) expresses that the measure of Box A's height strictly exceeds Box B's width. Comparisons such as the latter ('subcomparatives') are possible because the dimensions in question share a common measure, length.

(33) Box A is taller than Box B.

(34) a. Box A is taller than it is wide.

b. Box A is taller than Box B is wide.

Commensurating comparatives like (33) involve comparing the measure of two objects along a single dimension. Consider the arrangement of boxes illustrated in Figure 17. The comparative in (33) expresses that the vertical extent of Box A is greater than the vertical extent of Box B (from their common base point of reference). We can likewise make comparisons of different dimensions sharing a common measure (e.g. extent in space) for one and the same object. For instance, (34a) compares A's vertical extent to its horizontal extent, in much the same way that (34b) compares these dimensions between A and B.

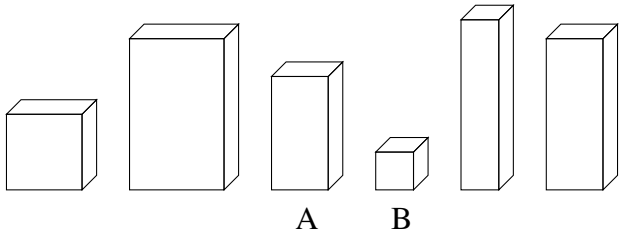


Figure 17: A collection of boxes; A is strictly taller (in terms of its y dimension) than B.

One intuition specific to commensurating comparatives is that they are intuitively just as true in the scenario illustrated in Figure 17 as they would be in a scenario lacking the

context boxes surrounding A and B. That is, only the objects referenced in the sentences are relevant for determining the truth or falsity of the statement. But for comparatives like (35) and (36), the interpretations appear to be different. Instead of inviting us to consider a direct comparison of A's height to B's, (35) invites us to think that A might be a better fit to the category 'tall' than B is. I will follow Wellwood (2014) in calling these 'categorizing comparatives' due to their denoted concern with fit to a category.

(35) Box A is more tall than Box B.

(36) a. Box A is more tall than it is wide.

b. Box A is more tall than Box B is wide.

With this understanding in mind, I return to the scenario depicted in Figure 17. The statement expressed in (35) suggests that A is a better fit to the category of 'tall' than B is. And indeed, as a competent speaker of English, I would judge (35) as true in this context since, if anything, A would count as tall in this context, but B would not. And following the same evaluation process, I would judge (36a) as false and (36b) as true.

As I have shown here, the labels 'commensurating' and 'categorizing' are intended to describe the interpretation of a given comparative, namely whether that interpretation says something about the object in question independent of its context (commensurating), or relative to its context (categorizing). For Wellwood (2014), there is one type of comparative with two distinct types of interpretation—commensuration and categorization. While both commensuration and categorization involve the strict comparison of measures, the categorizing variety differs from the commensurating variety in its measure of the speaker's degree of confidence in an object or entity's fit to a category.

I have already pointed out differences in the context-sensitivity that set categorizing comparatives apart from their commensurating counterparts. Before proceeding, I will highlight a second difference, which concerns their distribution. Commensurating comparatives require more structure than simply a bare adjective in the *than*-clause, as opposed to categorizing comparatives (37) (di Sciullo and Williams, 1987). A similar difference, is the

observation that only the commensurating comparative is compatible with a bare measure phrase (e.g. *6 inches*) in the standard clause (38) (Giannakidou and Yoon, 2011).

(37) a. ? Box A is taller than wide.

b. Box A is more tall than wide.

(38) a. Box A is taller than 6 inches.

b. ? Box A is more tall than 6 feet.

These examples serve to highlight the differences in the distributions of commensurating and categorizing comparatives in English. What remains to be seen is a structural account that maps the distributional and interpretational differences of these comparatives to differences in the morphosyntax. To explore this, I will turn to two separate proposals, made by Embick (2007) and Rett (2015), respectively.

4.1.2 Embick (2007) on κ-comparatives

To account for patterns of analytic/synthetic alternations in English comparatives, in which two-word (e.g. *more tall*) and one-word (e.g. *taller*) comparative forms alternate with each other, Embick (2007) proposed a silent morphosyntatic constituent κ as part of the morphosyntactic representation of analytic forms. I will briefly review the motivation for Embick's proposal here, and how it will be leveraged toward explaining differences in interpretation characteristic of commensurating and categorizing comparatives.

In laying out his account of analytic/synthetic alternations in English, Embick's (2007) principal concern was to offer a morphologically-motivated explanation for the distributional differences between the comparative form *-er* from the analytic form *more*. Embick observed that synthetic forms like (39a) can alternate with analytic forms like (39b), while (39c) appears with only one possible form. Embick argues that these patterns cannot be explained as instances of optionality if something in (39b) blocks the normal rules of comparative formation in (39a).

(39) a. Al is smarter than Bill is.

b. Al is more smart than Bill is.

c. Al is more intelligent than Bill is.

Embick (2007), building on earlier work by Bresnan (1973), posits that differences in alternations between English synthetic comparatives (e.g. *taller*) and analytic comparatives (e.g. *more/less tall*) are due to the presence of a phonologically unpronounced morpheme that he dubs κ. The presence of this morpheme crucially blocks a process whereby the comparative (or superlative) morpheme attaches to the adjectival root to form the synthetic comparative in forms that undergo comparative synthetic formation.[15] Embick's proposed structures of the analytic and synthetic comparative are provided in (40).

(40) a. ER TALL $>$ *taller*

b. ER κ TALL $>$ *more tall*

Importantly for my purposes, on Embick's morphological analysis, the analytic variants feature the morpheme κ, but the synthetic variants do not. The question now is, if this morpheme is present, what difference in interpretation comes about? A relevant observation here (following McCawley 1988; Embick and Noyer 2007; Morzycki 2011; Wellwood 2014) is that competent speakers of English have an intuitive semantic distinction between these variants. For example, in the case of *A is more tall than B*, the comparison is not strictly about the length of A as compared to the length of B, but instead, about some other property. Researchers have disagreed on exactly how to characterize this difference. For McCawley (1988), this difference concerns how 'appropriate' it is to say *tall* over some other descriptor; for Morzycki (2011), it is a concern about how 'precise' the difference is; for Giannakidou and Yoon (2011), the difference is about 'desirability'; and for Wellwood (2014), 'confidence' in the assertion. Despite their differences, these analyses share positing

[15] A pre-theoretic generalization about comparatives that can undergo synthetic comparative formation is the phonological restriction of having at most two syllables. Thus, the type of alternation Embick was interested in accounting for can only be investigated for a subset of the gradable adjectives in English.

an 'evaluative' (if not truth conditional) difference—that is, analytic comparatives involve evaluating an entity with respect to a category or relevance class. Rett (2015) proposes a particular way of cashing out this 'evaluative' property, and it is to her proposal that I now turn.

4.1.3 Rett (2015) on evaluativity in gradable adjectives

To offer a general intuition about what evaluativity is and why its treatment is so central to theories of gradable adjectives (and to theories of degree semantics, more generally), I begin by considering the contrast in (41).

(41) a. John is tall.

b. John is five feet tall.

Sentences like (41a) involve an unmodified gradable adjective, whereas sentences like (41b) involve a measure phrase composed of a numeral and optional measure noun (e.g. *five feet*). Sentences (41a) and (41b) contrast as follows: the adjective *tall* relates the subject (John) to a specific degree of tallness (five feet). In contrast, (41a) lacks a measure phrase, and intuitively appears to require that John's degree of tallness exceed some relevant or salient standard of height. Rett refers to this semantic property as evaluativity, and calls an adjectival construction evaluative if and only if the construction makes reference to a degree which exceeds a contextually marked standard.

The semantic contrast between (41a) and (41b) appears to stand at odds with a basic premise of compositional semantics, Frege's (1884) Principle of Compositionality, which assumes that the meaning of a sentence depends on the meaning of its morphemes and the syntax used to combine them. Hence, if a gradable adjective like *tall* means something like 'counts as tall in some context C,' this raises the question of why this meaning disappears in constructions like (41b) with added presence of a measure phrase. Conversely, if *tall* means something less than 'counts as tall in some context,' how do we then account for the evaluative property in (41a)?

A traditional approach to resolving this dilemma is Cresswell's (1976) semantics of gradable adjectives, in which they denote relations between individuals and degrees. On Cresswell's approach, an adjective like *tall* denotes a relation between individuals and degrees of tallness, and thus maps each individual (x) with several degrees of tallness (d), rather than to a single degree of maximum height (e.g. five feet). In (42), *tall(x,d)* should be read as "x is tall to at least degree d."

(42) $[\![\text{TALL}]\!] = \lambda x \lambda d.tall(x,d)$

This intuitive approach so far falls short of predicting the truth conditions of positive constructions like (41a) on two accounts: first, it incorrectly predicts that the sentence denotes a degree property instead of a proposition. Second, it fails to predict that the sentence is evaluative.

This tension between a relational analysis of gradable adjectives (i.e. one that maps individuals to degrees) and an adequate treatment of positive constructions is usually resolved with a null morpheme called 'POS' (Bartsch and Vennemann 1972; Kennedy 1999), which is defined in (43). POS denotes a function from gradable adjective meanings (type $< e,< d,t >>$) to individual properties (type $< e,t >$) by restricting ($d > s$) and binding ($\exists d$) the gradable adjective's degree argument.

(43) $[\![\text{POS}]\!] = \lambda G \in D_{<e,<d,t>>} \lambda x \exists d[G(x,d) \wedge d > s]$

The introduction of POS solves both the extra argument and evaluativity problems. Evaluativity is introduced by relating the degree argument of the gradable adjective to a contextual standard, s, while the addition of a quantifier, which existentially binds the degree argument, resolves the extra argument problem. As a consequence, a constituent POS for a gradable adjective like *tall* has the semantic type of a property of individuals and an evaluative meaning. On this theory, JOHN IS POS TALL has the meaning given in (44).

(44) JOHN IS POS TALL $= \exists d[tall(John,d) \wedge d > s]$

In lieu of POS, Rett proposes EVAL as in (45). Rett defines EVAL as a degree modifier that denotes a relation between sets of degrees D (type $< d,t >$). As a degree modifier, EVAL addresses the evaluativity problem by binding the degree argument with the existential closure step. Under this formulation, constructions like (46) are predicted to denote a true proposition if and only if there is a degree to which John is tall which exceeds the contextual standard (i.e. $\exists d[tall(John,d) \wedge d > s_{tall}]$).[16] It is worth noting that Rett's EVAL has a slightly different semantics from POS in that it can be used compositionally in environments where POS cannot be used compositionally. I consider these distributional differences below.

(45) $[\![\text{EVAL}]\!] = \lambda D \lambda d.D(d) \wedge d > s$

(46) JOHN IS EVAL TALL $= \exists d[tall(John,d) \wedge d > s] \rightarrow \exists d[tall(John,d) \wedge d > s_{tall}]$

Concerning the distribution of EVAL, Rett points out that POS cannot account for the absence of evaluativity in comparative constructions like (47a) or in equative constructions like (47a). The comparative in (47a) does not presuppose that John or Bill is short, and the equative construction in (47b) does not presuppose that either is tall.

(47) a. John is shorter than Bill.

b. John is as tall as Bill.

Elsewhere, Rett Rett, 2015[22] states that "neither sentence in [(48)] requires that Adam or Doug be tall or short, respectively. [(48a)] could be truthfully and felicitously uttered in a context in which Adam and Dough are clearly short, and [(48b)] could be truthfully and felicitously uttered in a context in which they are clearly tall." While I agree with Rett's observation concerning (48a), her observation about (48b) is not in accord with my intuitions. As a native speaker of English, my intuition is that (48a) and (48b) stand in contrast to one anther: the former does not imply that either Adam or Doug is tall, while the latter does indeed imply that at least one is short.

[16] On Rett's analysis, the semantic type assigned to EVAL will require, for sentences containing it, some form of abstraction to bind the degree argument of the lexical adjective. I gloss over these details here.

(48) a. Adam is taller than Doug.

b. Adam is shorter than Doug.

I distill this debate down to the assumption that wherever *short* appears, EVAL is present in the morphosyntactic representation. This assumption follows work by Moracchini (2018), who looked at a much wider range of cases than either Embick or Rett, including the cases at issue here (the analytic/synthetic alternations along with their positive and negative variants). In her analysis, EVAL appears whenever the analytic variant appears (which includes forms like *less tall*). Semantically, EVAL is a modifier of the adjectival morph, TALL; precisely what that means will be taken up later. But for present purposes, I will assume the morphological structures associated with the strings given in (49). In addition, I will assume Heim's use of LITTLE* and this analysis is also adopted by Moracchini (2018).

(49) a. *taller* = ER TALL

b. *shorter* = ER LITTLE TALL

c. *more tall* = ER EVAL TALL

d. *more short* = ER EVAL LITTLE TALL

e. *less tall* = ER LITTLE* EVAL TALL

f. *less short* = ER LITTLE* EVAL LITTLE TALL

As a final point, it is important to acknowledge that Embick, Rett, and Moracchini were coming at the problem discussed here from different perspectives, using different tools, and considering different subsets of the strings in (49). For my purposes, the decompositions in (49) unify these perspectives, where I write EVAL where Embick would write κ. I see no harm in this substitution, because Embick attributes to κ whatever is responsible for the semantic difference between, e.g., *taller* and *more tall*, and Moracchini's use of the EVAL label indeed ties the silent morpheme to a specific semantics.

4.1.4 EVAL and the psycholinguistics of degree comparisons

The preceding discussion focused on formal approaches to the meaning of the silent morphosyntactic element EVAL. In light of the general similarity in meaning between POS and EVAL, one question that this discussion leaves open is precisely what the truth conditional contribution of that morpheme is in our target cases. In other words, how might native speakers of English operationalize evaluativity, and how might experimental evidence be leveraged to narrow this down?

Recent work by Solt and Gotzner (2012) has investigated this question experimentally. In their work, they presented participants with sentences like that in (50), along with scenes featuring relevant objects (including that mentioned in the sentence) that instantiate the relevant property to varying degrees. Against this background, they were able to evaluate various proposals for how (50) might be judged, e.g. (50a)-(50c).

(50) $[\![\text{JOHN IS TALL}]\!]^C = 1$ iff

a. John is among the tallest $n\%$ of Cs

b. $\text{HEIGHT}(John)$ is among the top $n\%$ of heights of Cs

c. $\text{HEIGHT}(John) > mean_{x \in C}(\text{HEIGHT}(x))$

On (50a), John might be considered tall if he is among the tallest n percent of the comparison class (e.g. in the top quartile; cf. Bale 2011). On (50b), John might be considered tall if his degree of height falls within some specified subsegment of the range of heights corresponding to the comparison class—e.g. the top quartile of this range (cf. Bale 2008). On (50c), the standard for tallness might be derived as an average over the heights of the individuals in the comparison class (Solt and Gotzner 2012, cf. Bartsch and Vennemann 1972; von Stechow 1984).

Solt & Gotzner's studies were aimed primarily to differentiate whether people made use of scales that track mere rank order information—i.e., a scale that encodes whether one individual x exceeds another individual y by tallness, but does not encode by how much x exceeds

y (cf. Bale's theory)—as opposed to the richer structure typically associated with scales in degree semantic theories. Their evidence suggests that the richer structure is needed. In my study, I retain consideration of rank order comparison, as well as introduce a new operationalization of evaluativity in comparatives—what I call 'goodness of fit'—that is derived from people's evaluation of positive sentences like (50a).

4.1.5 Hypotheses and predictions

This section outlines the hypotheses and subsequent predictions that feed into the post-hoc analysis (revisiting Experiment 2b) and experiment (Experiment 3) that follow.

Post-hoc analysis: Experiment 2b. In the post-hoc analysis described in the forthcoming section, I investigate whether independent processing evidence can be found in support of the explicit parameters itemized in (51). Crucially, the parameter whose inclusion has not yet been assessed up to this point is t_d—the cost of 'context checking'—putatively associated with the evaluativity.

(51) Factors affecting response latency

a. t_0 - 'baseline processing parameter'

b. t_a - cost of 'falsification' (CONGRUENCE)

c. t_b - cost of 'adjectival negation' (LITTLE)

d. t_c - cost of 'comparative negation' (LITTLE*)

e. t_d - cost of 'context checking' (EVAL/κ)

How the parameters in (51) map to the morphosyntactic elements in the comparatives of interest is illustrated in Table 11. Here I assume that the morphosyntactic constituents TALL and ER are present in every comparative representation. These constituents are therefore not included to consolidate the table. Consequently, I have omitted t_0 and t_a from the listed parameters are these are common to all comparatives.

Comparative	Morphosyntax	Processing Parameters
shorter	ER LITTLE TALL	t_b
more tall	ER EVAL TALL	t_d
more short	ER EVAL LITTLE TALL	$t_b + t_d$
less tall	ER LITTLE* EVAL TALL	$t_c + t_d$
less short	ER LITTLE* EVAL LITTLE TALL	$t_b + t_c + t_b$

Table 11: Summary of putative processing parameters corresponding to the evaluation of adjectival comparatives

In addition to itemizing the parameters and mapping them to the morphosyntactic elements in representations, it is also instructive to illustrate which decompositional theories predict which representations, and therefore, which decompositional theories will be supported by processing evidence for particular morphosyntactic elements. Table 12 illustrates precisely this.

	Decompositional Analysis		
Comparative	Büring (2007a)	Heim (2008)	Morachini (2018)
shorter	LITTLE	LITTLE	LITTLE
more tall	–	–	EVAL
more short	LITTLE	LITTLE	EVAL LITTLE
less tall	LITTLE	LITTLE*	LITTLE* EVAL
less short	LITTLE LITTLE	LITTLE* LITTLE	LITTLE* EVAL LITTLE

Table 12: Summary the morphosyntactic units of interpretation predicted by each decompositional theory

Experiment 3. Building on the previous work that examined the interpretation of evaluativity in gradable adjectives (e.g. Bale 2008, 2011; Solt and Gotzner 2012), Experiment 3 examines the interpretation of subcomparative statements (e.g. *Box B is taller than Box C is wide*). Specifically, I investigate three possible algorithmic interpretations of a subcomparative like (52), as given in (53).

(52) $[\![\text{Box B is taller than Box C is wide}]\!]^C = 1$ iff

(53) a. $\text{HEIGHT}(BoxB) > \text{WIDTH}(BoxC)$ LENGTH COMPARISON

b. $rank(\text{HEIGHT}(BoxB)) > rank(\text{WIDTH}(BoxC))$ RANK COMPARISON

c. $GOF(\text{HEIGHT}(BoxB)) > GOF(\text{WIDTH}(BoxC))$ GOF COMPARISON

(53a) denotes a strict length comparison between the height of Box B and the width of Box C, and is true just in case the vertical extent of Box B is greater than the horizontal extent of Box C. (53b) denotes a rank comparison between the height of Box B and the width of Box C with respect to the relevant comparison class, being true just in case the rank of Box B's height exceeds the rank of Box C's width.[17] Finally, (53c) denotes a comparison of the goodness-of-fit of Box B to the category 'tall' with the goodness-of-fit of Box C to the category 'wide.' Precisely how goodness-of-fit predictions were computed will be discussed in detail in the coming section.

4.2 Revisiting Experiment 2b: Processing evidence for EVAL/κ

This section summarizes the methods and results of a post-hoc analysis conducted on the data from Experiment 2b. The question I attempt to address here is whether I can find processing evidence for Rett's EVAL/Embick's κ. To preview, I find evidence that adding a predictor (corresponding to EVAL/κ) to the model significantly increases the model's fit to the data. Similarly, the statistical impact of this predictor is shown to be robust.

4.2.1 Methodology

To conduct this post-hoc analysis, I returned to the data collected from Experiment 2b, and conducted a series of linear mixed effects model (LMEM) comparisons. I added an additional predictor this analysis, dubbed EVAL, and applied orthogonal contrasts: -0.5 for non-evaluative comparatives, 0.5 for evaluative comparatives (following Moracchini's 2018 morphosyntactic taxonomy). All comparative-related parameters included in the full LMEM, along with their level codes, are shown in Table 13. It is also instructive to note here that

[17] Here, 'rank' was operationalized in such a way that it is not sensitive to a difference between 'tallest/widest among all boxes' (cf. Bale 2008) and 'widest/tallest among all box height/width instances' (cf. Bale 2011)

CONGRUENCE was included in this model; however, this factor was not a property of the comparative (or sentence stimulus). Rather, it was an item-level property, pertaining to a sentence stimulus in conjunction with a picture stimulus.

Comparative	ADJ	COMP	EVAL
taller	−0.5	−0.5	−0.5
shorter	0.5	−0.5	−0.5
more tall	−0.5	−0.5	0.5
more short	0.5	−0.5	0.5
less tall	−0.5	0.5	0.5
less short	0.5	0.5	0.5

Table 13: Summary of comparative/sentence-level predictors included in this post-hoc analysis (with CONGRUENCE omitted).

Prior to running a series of LMEM comparisons to determine the statistical robustness of each of the model's predictors, an pair of model evaluations which of two models fit the data better: (1) a Büring/Heim model, consisting of only ADJ, COMP and CONGRUENCE; and (2) a Büring/Heim & Rett/Embick model, consisting of the aforementioned predictors, plus EVAL. The purpose of this pair of model evaluations was to demonstrate whether the latter (more complex) model indeed provided a better fit to the RT data than the former. This evaluation is similar to evaluating difference in model fit via LMEM comparisons, but is intended to provide additional, supplementary information.

4.2.2 Results

Model	RMSE	log likelihood
Büring/Heim	0.0995	5360.9
Büring/Heim + Rett/Embick	0.0996	5376.5

Table 14: Summary of model evaluation comparisons

Reaction times. As expected from results in previous chapters, participants took longer to evaluate sentences with *shorter* than with *taller*. This was reflected in a strong main effect of ADJECTIVAL NEGATION ($\beta = -0.02$, $\chi^2 = 24.62$, $p < 0.001$) in the predicted direction:

Predictor	χ^2	p	β	means (log ms)
ADJ	24.62	$p < 0.001$	-0.02	*tall*: 7.14, *short*: 7.21
COMP	61.27	$p < 0.001$	-0.04	*more*: 7.14, *less*: 7.29
EVAL/κ	31.11	$p < 0.001$	-0.03	non-evaluative: 7.09, evaluative: 7.21
CONGRUENCE	39.75	$p < 0.001$	-0.02	congruent: 7.14, incongruent: 7.21

Table 15: Summary of model results

RTs in the negative conditions were longer than in the positive conditions (means, in ms: *tall* 1395.13ms, *short* 1498.40ms).

Also as expected, participants took longer to evaluate sentences with *less* than those with *more*. This was reflected in a robust main effect of ADJECTIVAL NEGATION ($\beta = -0.04$, $\chi^2 = 61.27$, $p < 0.001$) in the predicted direction: RTs in the negative conditions were longer than in the positive conditions (means, in ms: *more* 1386.58ms, *less* 1628.01ms).

The novel predictor to this analysis, EVAL/κ, proved to be statistically significant: participants took longer to respond to sentences with evaluative comparatives than those without evaluative comparatives ($\beta = -0.03$, $\chi^2 = 31.11$, $p < 0.001$). Reaction times were longer for evaluative comparatives than for non-evaluative comparatives (means, in ms: non-evaluative 1283.97ms, evaluative 1501.29ms).

Finally, and as anticipated, participants took longer to false statements than to evaluate true statements. This was reflected in a strong main effect of CONGRUENCE ($\beta = -0.02$, $\chi^2 = 39.75$, $p < 0.001$): RTs in for statements judged false were on average longer than RTs for statements judged false (means, in ms: congruent 1388.85ms, incongruent 1500.12ms).

Statistical power. A post-hoc power analysis was conducted on the RT model for this analysis using the R-package *pwr* (Champely, 2018). The effect size associated with this model was 0.0437 with a sample size of 6376 observations. The estimated power of the model was 0.99, which greatly exceeds the conventionally desired statistical power of 0.8 and indicates a more than sufficiently powered analysis. Given the results of this power estimation, I conclude with reasonable certainty that the effects observed were not false positives.

4.2.3 Discussion

In the hypotheses and predictions for this chapter, I advanced the set of parameters in (54) as potential factors affecting response latency. The results of the post-hoc analysis discussed above suggest that EVAL may play a non-trivial role in the interpretation of comparatives in which it has been hypothesized to occur. It does not yet address whether or how this effect relates to the parameters listed in ()54).

(54) Parameters affecting response latency

a. t_0 - 'baseline processing parameter'

b. t_a - cost of 'falsification' (CONGRUENCE)

c. t_b - cost of 'adjectival negation' (LITTLE)

d. t_c - cost of 'comparative negation' (LITTLE*)

e. t_d - cost of 'context checking' (EVAL/κ)

One salient limitation of this post-hoc analysis is the confounding of EVAL with morphology. Because evaluativity always co-occurs with the analytic form (and non-evaluativity with the synthetic), this analysis cannot disambiguate between a processing effect of evaluativity and one due merely to the morphology. Experiment 3 directly addresses this concern by probing people's interpretation of subcomparatives.

4.3 Experiment 3: Assessing processing evidence for EVAL in subcomparatives

This section summarizes the methods and results of Experiment 3. The question I investigate here is how evaluativity is understood in subcomparative statements. I compare a number of number of different ways that the truth conditions for evaluativity might be computed. In my statistical analyses, I compare the performance of these hypotheses with respect to participants' responses in cases where the two hypotheses make the opposite predictions.

4.3.1 Design and stimuli

Design. Experiment 3 featured a 2x2 true/false judgment task design in which two factors were explicitly manipulated: LENGTH COMPARISON prediction and RANK COMPARISON prediction. This design generated four conditions: two in which LENGTH and RANK made identical true/false predictions, and two in which they made disparate true/false predictions. Picture stimuli were designed to carry out the condition contrasts for each type of comparative statement. More details about the stimuli are given below.

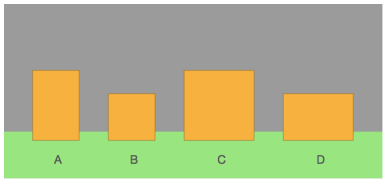
Sentence stimuli. Sentences containing comparatives were divided into two groups: baseline and experimental. Baseline items were of the form *Box B is tall* and as such were not subcomparatives, they were simply absolute or positive occurrences of the relevant gradable adjective. In contrast, all sentences in the experimental group were subcomparatives, e.g. *Box B is taller than Box C is wide*. The adjective in the *than*-clause was always *wide*, while the adjective and the comparative form appearing in the matrix clause were explicitly manipulated. All comparative sentences appearing in the sentence stimuli used in Experiment 3 are shown in Table 16.

Comparative	Group	Evaluativity
tall	baseline	+EVAL
short	baseline	+EVAL
wide	baseline	+EVAL
taller	experimental	-EVAL
shorter	experimental	-EVAL
more tall	experimental	+EVAL
more short	experimental	+EVAL
less tall	experimental	+EVAL
less short	experimental	+EVAL

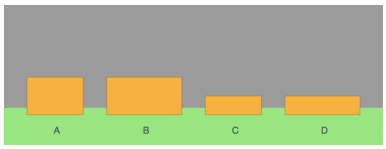
Table 16: All comparatives appearing in the sentence stimuli used in Experiment 3, along with the hypothetical distribution of EVAL, which were derived from Moracchini (2018).

Picture stimuli. 16 unique picture stimuli were generated in-browser by updating variables nested within SVG-drawing commands. Variables nested within the vector graphics were

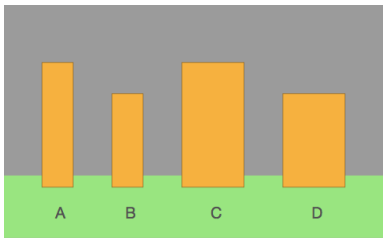
updated immediately upon transitioning from item to item, thus instantly changing the picture in the browser's view. The variables manipulated in each picture were the height and width values for the boxes, as well as the maximum height and width of the container. The height of the container was always 80px plus the height of the tallest box, while the width of the container was always 250px plus the combined widths of all four boxes. (250px was the sum of a consistent spacing of 50px between each box, and between each box and the left/right sides of the image container.)



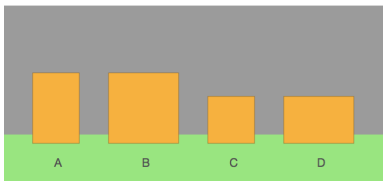
(a) Length: $B < C$, Rank: $B < C$



(b) Length: $B < C$, Rank: $B > C$



(c) Length: $B > C$, Rank: $B < C$



(d) Length: $B > C$, Rank: $B > C$

Figure 18: Sample picture stimuli used in Experiment 3.

Figure 18 illustrates a representative subset of the 16 image stimuli. Pictures were designed to counterbalance consistency with strict Length comparison and Rank comparison predictions: (a) 25% Length: $B < C$, Rank: $B < C$; (b) 25% Length: $B < C$, Rank: $B > C$; (c) 25% Length: $B > C$, Rank: $B < C$; (d) Length: $B > C$, Rank: $B > C$. As such, when paired with any comparative, half of picture stimuli were designed to receive identical true/false judgments, while the other half were designed to receive disparate judgments.

4.3.2 Procedure

Following consent, participants saw a total 144 trials, each consisting of a statement and image, simultaneous presented with the statement above the picture, until a response was selected. To respond, participants clicked 'YES' if they thought the statement was true of the associated image, 'NO' if they thought it was not. Participants saw the following instructions after consenting to participate in this study.

> Welcome to the experiment!
>
> In this experiment, you're going to see pictures with sentences above them. Your task is to decide whether each sentence accurately describes the picture it comes with. Select YES if you think the sentence accurately describes the picture; NO if you think it does not. Please respond as quickly and as accurately as you can.
>
> This experiment will consist of 144 yes/no picture+sentence trials, and should take approximately 15 minutes to complete. At the end, you will receive a unique completion code to submit back on Mechanical Turk.
>
> Your browser window should be at maximum height & width for the duration of this experiment.
>
> Ready? Press the button below to begin.

20 participants, native speakers of English aged 18 years old or older, were recruited via Amazon's Mechanical Turk platform to participate in this experiment. Participants were compensated $2.49 in exchange for up to 15 minutes of their time.

4.3.3 Analyses and exclusions

Following the same exclusionary procedure outlined in previous chapters, participant responses with RTs less than 200ms were excluded. In addition, responses with RTs outside 2.5 standard deviations of each participant's mean were excluded from subsequent analysis.

In contrast to previous analyses, here I examined binary ('yes/no') response judgments, which were contrast coded as 0 ('no') and 1 ('yes') respectively. In the results reported below, participant responses were taken as ground truth and compared directly to the responses (coded with the same scheme) predicted by the hypotheses LENGTH COMPARISON, RANK COMPARISON and GOODNESS-OF-FIT COMPARISON.

LENGTH COMPARISON predictions were computed by taking the vertical extent of Box B (in pixels) and comparing it directly to the horizontal extent of Box C. RANK COMPARISON predictions were computed by comparing the height rank of Box B (always either Rank 1 or Rank 2) with the width rank of Box C (also always either Rank 1 or Rank 2). Finally, GOODNESS-OF-FIT COMPARISON was computed for each picture by calculating the mean log likelihood of a participant responding 'yes' to *Box B is tall/short*, and subtracting from it the mean log likelihood of a participant responding 'yes' to *Box C is wide*.

For the comparatives *taller*, *more tall* and *less short*: LENGTH COMPARISON predicted an outcome of 1 iff the height of Box B exceeded the width of Box C; otherwise, the predicted outcome was 0. RANK COMPARISON predicted an outcome of 1 iff the height rank of Box B exceeded the width rank of Box C; otherwise, the predicted outcome was 0. GOODNESS-OF-FIT COMPARISON predicted an outcome of 1 iff the goodness-of-fit difference between Box B and Box C was positive; otherwise, the predicted outcome was 0. For the comparatives *shorter*, *more tall* and *less tall*: all predictions were the inverse of those indicated for the comparatives *taller*, *more tall* and *less short*.

The results below report the average accuracy of each type of prediction, which amounts to the mean of the true positive and true negative predictions.

4.3.4 Results

The results of Experiment 3 are presented in the following series of figures, with each figure showing a specific subset of the data. Figure 19 compares the prediction accuracy for all three prediction types, for only cases in which LENGTH COMPARISON and RANK COMPARISON

made disparate predictions, i.e. the former predicting 1 while the latter predicted 0, and vice versa. Figure 20 compares prediction accuracy for cases in which RANK COMPARISON and GOODNESS-OF-FIT made disparate predictions. Finally, Figure 21 compares prediction accuracy for cases in which LENGTH COMPARISON and GOODNESS-OF-FIT COMPARISON made contrasting predictions.

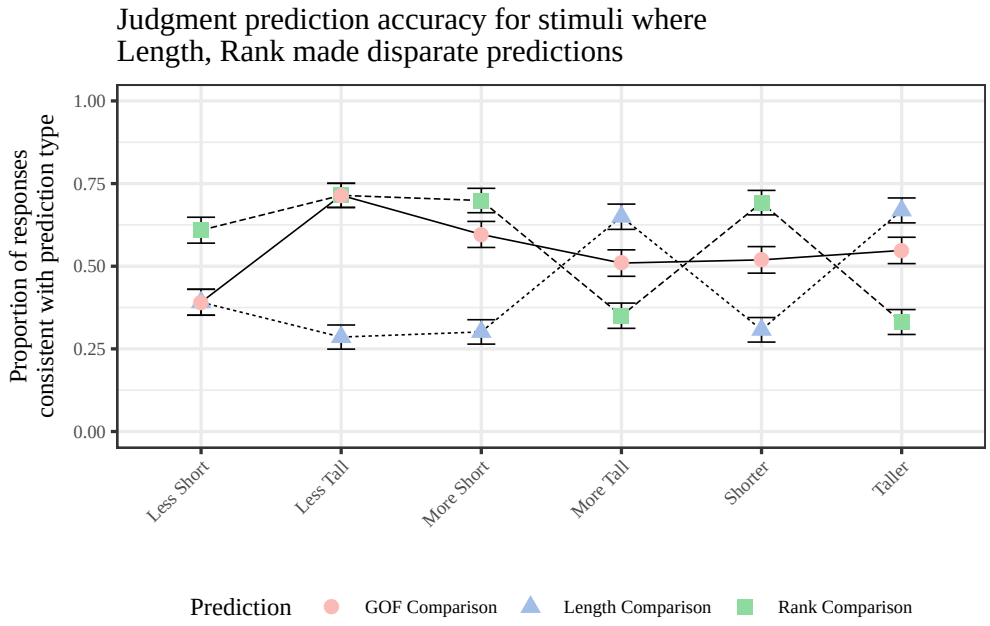


Figure 19: Stimuli for which Length Comparison and Rank Comparison made disparate predictions. Error bars represent standard error of the mean.

In Figure 19, LENGTH and RANK show complementary accuracy predictions due to the fact that the data is subsetted to include only cases where disparate true/false predictions were made. Here, RANK tends to outperform LENGTH in comparatives with analytic morphology, where EVAL is predicted to have an effect on interpretation. Looked at this way, GOF is consistently at chance for all comparatives except *less tall*, where it matches in performance with RANK.

In Figure 20, which looks at only the data for which RANK and GOF made dissimilar predictions, RANK and GOF are characterized by complementary accuracy rates. Here, GOF consistently outperforms both RANK and LENGTH prediction accuracies across the board,

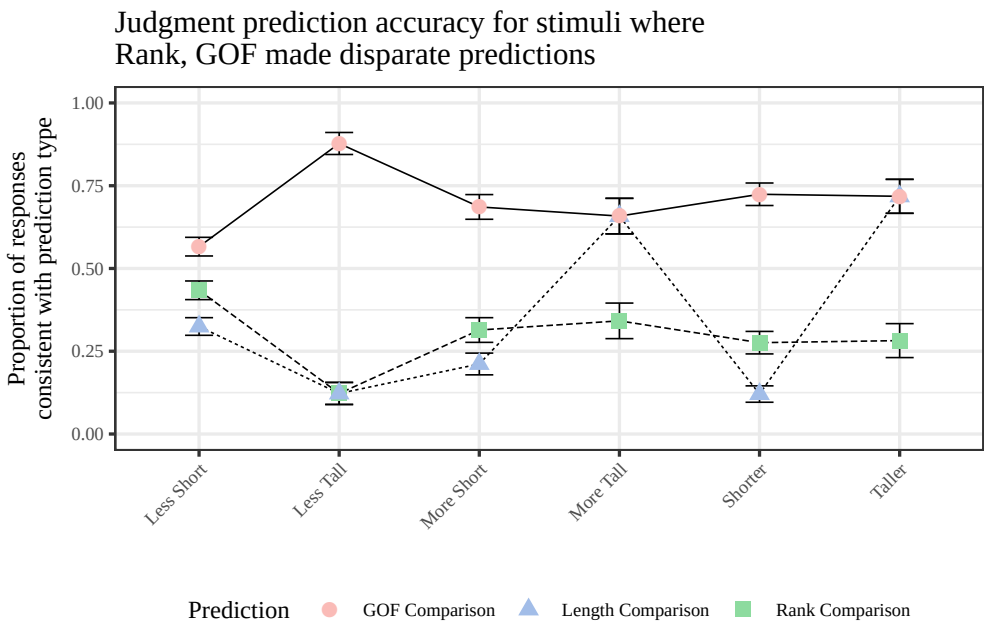


Figure 20: Stimuli for which Rank Comparison and GOF Comparison made disparate predictions. Error bars represent standard error of the mean.

except in *more tall*, where LENGTH has a similar accuracy rate.

Finally, in Figure 21, LENGTH and GOF are complementary because this subset considers only data points in which LENGTH and GOF make opposite predictions. Here again, GOF consistently makes more accurate predictions than both RANK and LENGTH, except for in the *more tall* and *taller* cases. In these exceptional cases, however, all prediction accuracy rates tended to be equally predictive, hovering around chance.

Taken together, these results suggest that GOODNESS-OF-FIT as operationalized here makes better predictions across the board than RANK, and thus may be taken as a better characterization of evaluativity. However, it is also important to note that GOODNESS-OF-FIT has generally high prediction accuracy rates not only in comparatives where evaluativity is expected, but also in synthetic comparatives, where evaluativity is not expected to apply. This unexpected observation merits additional elaboration.

It seems plausible that our participants assimilated *more tall* to *taller* (thus evaluating by LENGTH). A couple of possibilities may help explain why judgments for *shorter* did not look

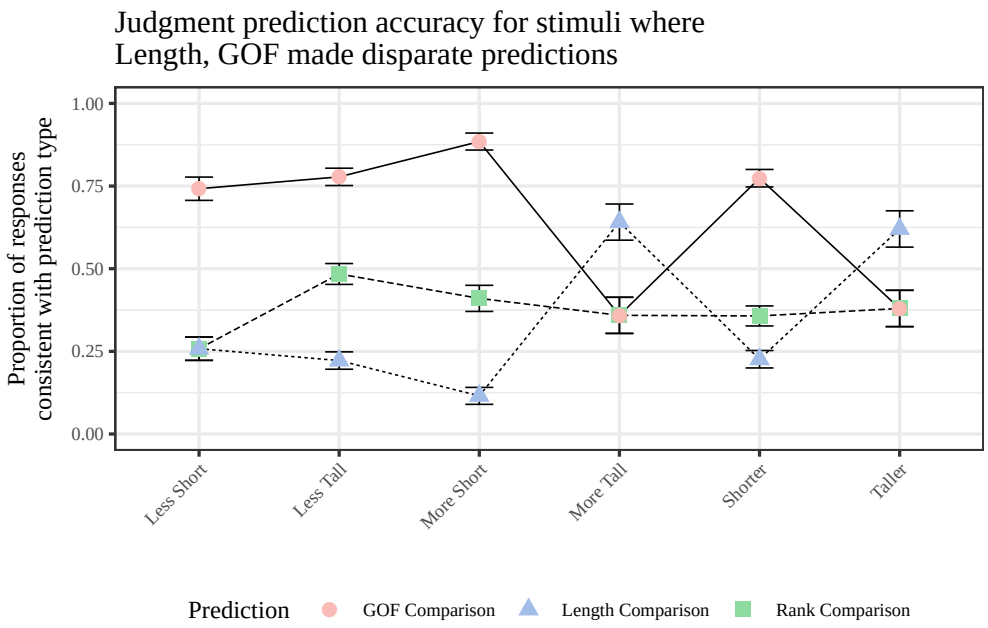


Figure 21: Stimuli for which Length Comparison and GOF Comparison made disparate predictions. Error bars represent standard error of the mean.

as expected. One possibility is that participants may have assimilated *shorter* to *more short*, hence evaluating both by GOODNESS-OF-FIT. Alternatively, and contrary to the literature I have reviewed in this chapter, it could be the case that *short* is +EVAL, and so its evaluation by GOODNESS-OF-FIT was not unexpected.

Follow-up studies testing stimuli between-subject may help address the question of whether the suspected assimilation is indeed ocurring, and if so, whether its occurrence is task-dependent. Further linguistic study of the distribution of EVAL with respect to negative adjectival comparatives may help to determine the viability of a +EVAL analysis of *shorter*. Such follow-up studies will be necessary to decide between these possibilities.

Statistical predictor importance. Figure 22 summarizes the (scaled) predictor importance values from an artificial neural network (ANN) that was fit to the *tall* and *short* baseline data from Experiment 3.[18] These values reflect the relative impact, in both magitude and

[18]These variable importance values were computed using Olden's algorithm, a method for computing weights asso- ciated with layers in an ANN. The ANN fit to each subset of the data consisted of only one layer. I elected to

direction, that each statistical predictor had on a participant's decision to categorize a picture as 'tall' or 'short'. Importance values greater than 0 reflect the likelihood of being assigned to the category in question, while values less than 0 reflect the likelihood of not being assigned to that category.

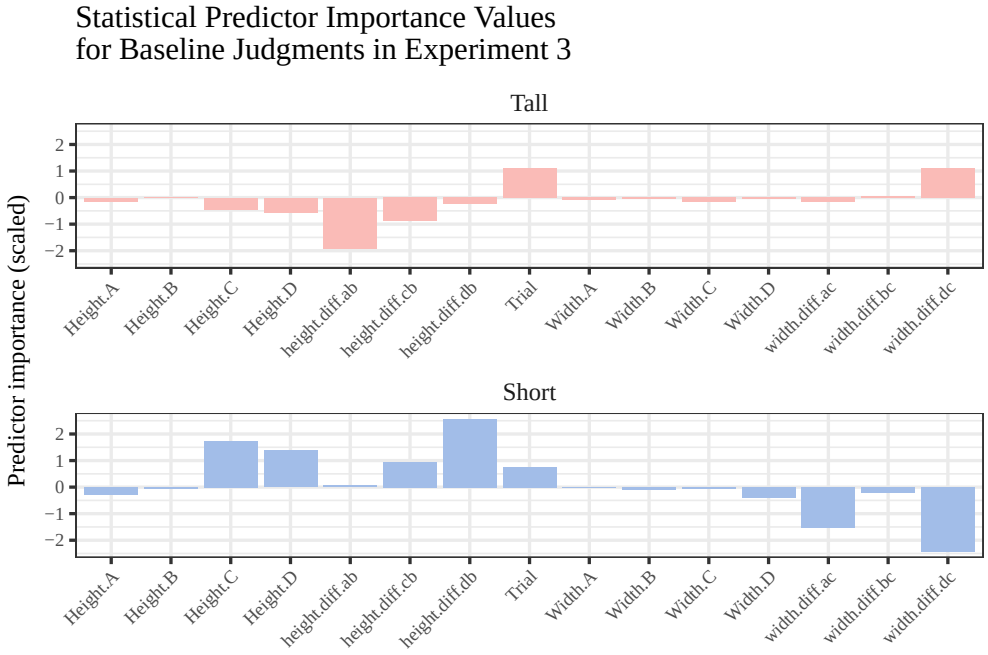


Figure 22: Importance of post-hoc statistical predictors as computed by Olden's algorithm on ANN feature weights.

I will not attempt to reason about every aspect of the predictor importance values shown in Figure 22. Instead, I will simply note that a few prominent observations. One noteworthy result here is the observation that the importance values for 'tall' are not simply the inverse of the values for 'short', and vice versa. Pre-theoretically, we might have expected such an inverse relationship to hold here given the representations I have used for these gradable adjectives throughout this dissertation. A second observation relates to differences in sensitivity to statistical properties: here, participants categorizing images as 'tall' and 'short' appear to be sensitive to different properties of the image for each adjective. Finally, there

include only a single (input) layer due to concerns with interpreting the output of Olden's algorithm when multiple layers comprise the architecture of the ANN.

appears to be a non-trivial importance of trial order (denoted 'Trial') for both adjectives. This suggests that a picture's relative standing in the trial order of the experiment influenced its categorization as 'tall' or 'short'—something that accords with my intuitions about how comparison classes might be constructed online.

This after-thought on what statistical properties of the picture stimuli may have had an impact on participants' decisions to bin them as 'tall' or 'short' serves to prime a future discussion on what a follow-up to Experiment 3 might investigate. I will take up this discussion again briefly in the concluding chapter that follows.

4.4 General discussion

In this chapter, I introduced the interpretational and distributional differences between commensurating and categorizing comparatives, and discussed accounts concerning silent morphosyntactic elements proposed to capture these differences. I then presented the results of two analyses—a post-hoc analysis revisiting Experiment 2b, and Experiment 3. In the post-hoc analysis, I found evidence suggesting that the addition of a predictor corresponding with Rett's (2015) EVAL resulted in a better-fitting model to the RT data obtained from Experiment 2b (initially presented in Chapter 3).

The results of Experiment 3 suggest that 'goodness-of-fit' as operationalized here may capture aspects of the behavioral data beyond its expected distributional scope. The concept underlying this predictor was an intuition concerning how evaluativity might be borne out behaviorally. Specifically, the assumption was that the evaluation of a statement like *Box B is tall* would involve evaluation of *tallness* in the same operational manner as the evaluation of a categorizing comparative like *more tall*, but crucially not in the same manner as a commensurating comparative like *taller*.

Taken together, these results suggest that goodness-of-fit may capture some subset of the psychological realia associated with evaluativity, but further research will be required to unpack precisely what realia these results correspond to. Moreover, there also exists the

possibility that participants differ with respect to which verification strategies they use, and consequently, goodness-of-fit may in reality subsume a number of different strategies. These possibilities are discussed in greater detail in the conclusion.

5 Conclusion

Throughout this dissertation, I have suggested that we can fruitfully think of semantic description as instantiating recipes for language understanding. Operating under the assumption that the mapping between linguistic structure and mental representation is transparent (Lidz et al. 2011), I operationalized decompositional hypotheses concerning the morphosyntactic structure of comparatives and investigated how those hypotheses might relate to behavioral signatures of processing during language comprehension. In this way, my experiments were used to leverage processing evidence towards deciding among different theories of the representation of gradable adjectives and comparative operators. In particular, I explored how the structure of morphosyntactic representations interact with visual representations.

I began by considering a distinction between atomic and decompositional theories of antonymic relations between gradable adjectives (e.g. *tall/short*). Under an atomic view, the entries for both polar pairs *tall* and *short* might natively have the primitive semantics in (55), where $H(x)$ is x's maximal degree of height, and the essentially trivial morphophonological spell-out rules in (56).

(55) a. $[\![\text{TALL}]\!] = \lambda x.(0, H(x))$

b. $[\![\text{SHORT}]\!] = \lambda x.(H(x), \infty)$

(56) a. TALL $>$ *tall*

b. SHORT $>$ *short*

Under a decompositional view, the entry in (55a) and spell-out rule in (56a) can remain the same, but there is no entry corresponding to *short* in the lexicon, and no morphophonological rule like (56b). Instead, *short* is built syntactically via the composition of TALL and some type of negative operator, e.g. LITTLE in (57a), and realized morphophonologically by a rule like (57b).

(57) a. $[\![\text{LITTLE}]\!] = \lambda A.\neg A$

b. LITTLE TALL $>$ *short*

Adopting ITT, I hypothesized an explicit relationship between pieces of morphosyntax and representations and operations in nonlinguistic cognition. With this linking hypothesis in hand, I predicted that a representation in which *short* decomposes into LITTLE TALL (Büring (2007b), cf. Heim (2008)) should yield an RT component corresponding to implicit negation (LITTLE) not necessarily implied by an atomic representation for *short*. I conducted an experiment in which participants judged the truth/falsity of sentences containing the comparatives *shorter* and *taller* against pictures of lines with varying heights. I found that participants took longer to verify sentences with *shorter* than *taller*; additionally, they took longer to verify sentences that were false of the images they were paired with than those that were true of the associated images.

These results suggest that a decompositional view of negative gradable adjectives, wherein *short* decomposes into LITTLE TALL, is tenable given the behavioral results obtained in Chapter 2. More generally, these results suggest that explicitly linking morphosyntactic units to cognitive operations can offer insight into language understanding. and may serve as a promising heuristic for adjudicating between competing representational theories. Building on this initial insight, in Chapter 3 I investigated whether processing evidence can be leveraged toward a similar decompositional analysis of comparatives with *less*. Heim's (2008) decompositional analysis of *less*-comparatives contrasts two variants of LITTLE: a scopally-fixed variant—the same as Büring's LITTLE—whose spellout yields *short* as in (59b), and a scopally-mobile variant (58b), whose spellout yields *less* as in (59b).

(58) a. $[\![\text{LITTLE}]\!] = \lambda A.\neg A$

b. $[\![\text{LITTLE*}]\!] = \lambda d.\lambda A.d \notin A$

(59) a. TALL LITTLE* $>$ *less*

b. LITTLE TALL $>$ *short*

To test these hypotheses, I designed two sentence verification tasks with synthetic and analytic comparatives (both *-er*/*more* and *less*) to investigate possible processing effects of the implicit negation underlying *less*. In both experiments, the RT evidence pointed to a

robust asymmetry between *less* and *more*. In the latter experiment, this effect was coupled with an additive effect of response latency associated with both *less* and *short*. Both results were taken to suggest that a decompositional account as summarized in (58) and (59) is promising.

Chapter 3 left open an important question concerning the assumptions I had so far made about a putative interpretational equivalence between synthetic and analytic comparatives. That is, following only Büring and Heim, there was no reason for me to expect a difference between how participants evaluated sentences with comparative forms like *taller* and those with forms like *more tall*. Yet, there is a small literature suggesting that a semantic difference attends the morphosyntactic difference between these forms (e.g. Embick 2007; Morzycki 2011; Giannakidou and Yoon 2011; Wellwood 2014), which has most recently been linked to the semantics of a morpheme called EVAL (e.g. Rett 2015; Moracchini 2018; (60)).

(60) $[\![\text{EVAL}]\!] = \lambda D \lambda d . D(d) \wedge d > s$

(61) a. ER TALL $>$ *taller*

b. ER EVAL TALL $>$ *more tall*

Thus, Chapter 4 extended my psycholinguistic investigation of decomposition to include consideration of evaluativity. I had two goals. The first was to see, via a post-hoc analysis on the data presented in Chapter 3, whether including a modeling predictor corresponding to EVAL resulted in a better-fitting model. This initial exploratory phase offered promising prospects. The second goal was to find out how people understood EVAL. I designed an experiment looking at the consistency of people's responses to analytic comparatives with different understandings of evaluativity proposed in the literature (e.g. Bale (2006, 2008)). To this space of extant semantic hypotheses, I also added my own—dubbed 'goodness-of-fit'—which tracks the likelihood of labeling an object as *tall*, *short*, or *wide* in a given context. This novel measure was intended to operationalize the notion that EVAL involves comparison to a contextually salient standard within a class of objects, as on previous approaches.

To weigh the predictions made by a 'goodness-of-fit' hypothesis against two other hypotheses–length comparison and rank comparison—I designed an experiment in which participants evaluated the truth/falsity of subcomparative statements (e.g. *Box B is taller than Box C is wide*). Analysis of the judgment data revealed that whenever goodness-of-fit and another hypothesis made disparate predictions, my goodness-of-fit measure consistently outperformed the prediction accuracy of that alternative hypothesis. These results suggested that, as operationalized and encoded, my goodness-of-fit metric captured some important aspect of evaluative interpretation. However, it leaves open the question of precisely what this metric captures concerning participants' evaluation procedures. A brief post-hoc investigation was conducted in which I fit a series of deep neural networks to the judgment data in hopes of revealing variable importance biases introduced by the best fitting model. These model weights revealed a number of interesting differences between, e.g., sensitivity to different statistical properties of the stimuli that distinguished evaluation of comparatives with *short* vs. *tall*.

A number of veins of inquiry stand out as ripe for further investigation. First, and of direct consequence to work done in Chapter 4, precisely what was captured by my 'goodness-of-fit' metric merits further probing in a follow-up investigation. In such an investigation, one could imagine explicitly, systematically varying the statistical properties of the picture stimuli, and designing an experimental procedure which attempts to zero-in on the hypothesis, or class of hypotheses, most consistent with participants' judgments. Such an experiment could proceed by exposing each participant to a wide distribution of items in an initial block, and in subsequent blocks, exposing the participant to items selected as a function of performance in the initial block.

In this adaptive experiment design, predictions would need to be attached (tagged) for each possible item so that the posterior probabilities of predictions could be recalculated following each successive participant response. To be more top-down about this putative experiment design, one could imagine generating a set of well-defined predictions on the basis of hypotheses already proposed in the literature (e.g. Bale 2006, 2008), in addition to

more fine-grained 'goodness-of-fit' hypotheses that can be made more precise by examining the variable importance weights of neural network models. Furthermore, one could imagine fitting a convolutional neural network to the images themselves, rather than to the derived statistical predictors, and looking under the hood to see when specific neurons fired during the process of categorizing images as *tall*, *short*, etc. Ultimately, any number of increasingly sophisticated models could potentially be fit to the data to better match the model's output to the output of the psycholinguistic processes implicated in the task under consideration—the crucial question would be precisely which—if any—psychological realia these models tap into.

To clarify: I am not suggesting that the meaning of an utterance is interchangeable with the algorithm used to compute the meaning of an utterance. I can, e.g. ask whether a certain line is longer than another without explicitly running any relevant verification procedures in my head. Additionally, I am not suggesting that we think about formal semantics as giving us a description of an independent level of semantic representation. Instead, throughout this dissertation I have looked at specific (and in some cases, competing) semantic descriptions for details about how morphosyntactic units (e.g. names of lines, gradable adjectives) map onto visual units (e.g. representations of lines, operations over those units). These are therefore descriptions of mappings between two levels of representation, and not a description of an independent level of representation.

Outside the scope of this dissertation was the question of the timecourse involved in computing the meaning of sentences. However, investigating the step-by-step timecourse of the evaluation of sentences stands out as something ripe for future investigation. Specifically, eye-tracking, mouse-tracking, and reading time methods could be leveraged to better understand not only what computations are involved, but also what information is being conscripted to arrive at truth judgments. Such methods could also be used to better understand possible individual variation in verification procedures. That is, nothin I have said rules out the possibility that some speakers may lexicalize an expression like *short* differently, and correspondingly, that their verification procedures might differ. As I will suggest below,

however, certain strong views on the nature of the interface between linguistic and non-linguistic cognition would not expect such variation. However, exploring this in subsequent studies will be critical.

Beyond variation at the individual level, we might also expect cross-linguistic variation, as the same thought can be expressed using very different morphophonology. How seriously should we take morphophonological variation as indicative of underlying morphosyntactic variation—i.e. should morphophonological variation be taken to indicate a different morphosyntax and therefore different verification procedures? If so, what might we expect to be universal vs. language-dependent? The strong hypothesis I have offered in this dissertation assumes that there is a universal set of morphosyntactic primitives (at least for functional vocabulary), and that the mappings between these primitives and extralinguistic cognition should not vary, and so the default verification procedures should not vary, either. Thus, the variation that we do see across languages should be characterizable as a matter of how morphosyntactic structure is bundled on the surface (i.e. in the morphophonology).

To summarize, this dissertation suggests that semantic descriptions can be fruitfully mined for predictions about how people represent and utilize sentential meanings, and how behavioral data can be leveraged, in turn, to decide between extensionally-equivalent, yet representationally-distinct, semantic descriptions. Adopting the assumption that the interface between linguistic and non-linguistic cognition is transparent, I found evidence for decompositional analyses of negative adjectival comparatives (cf. Büring 2007a,b), *less*-comparatives (cf. Heim 2008), and evaluativity in English categorizing comparatives (cf. Rett 2015; Moracchini 2018). In addition, by close comparison of hypotheses concerning how a hypothetical morpheme like EVAL is understood, I found evidence that 'goodness-of-fit' better characterized participants' responses across the board.

Taken together, this work points to the potential tenability of a strong understanding of the transparency thesis, whereby each symbol proposed in the logical representation of the meaning of a sentence corresponds to an operation or class of operations invoked during linguistic understanding. Investigating this hypothesis further can be used to gain deeper

insight into the nature of the interface between linguistic and nonlinguistic representation, and thus into how language is understood and acquired.

References

Bale, A. (2006). *The universal scale and the semantics of comparison*. PhD thesis, Montreal, Quebec.

Bale, A. (2008). A universal scale of comparison. *Linguistics and Philosophy*, 3(1):1–55.

Bale, A. C. (2011). Scales and comparison classes. *Natural Language Semantics*, 19(2):169–190.

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68:255–278.

Bartsch, R. and Vennemann, T. (1972). *Semantic structures: A study in the relation between semantics and syntax*. Athenaum, Frankfurt am Main.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4.

Beck, S. (2013). Lucinda driving too fast again: The scalar properties of ambiguous *than*-clauses. *Journal of Semantics*, 30:1–63.

Bobaljik, J. D. (2012). *Universals in Comparative Morphology: Suppletion, Superlatives, and the Structure of Words*. MIT Press, Boston, MA.

Bresnan, J. (1973). Syntax of the comparative clause construction in English. *Linguistic Inquiry*, 4(3):275–343.

Büring, D. (2007a). Cross-polar nomalies. In Friedman, T. and Gibson, M., editors, *Proceedings of Semantics and Linguistic Theory 17*, pages 37–52, Ithaca, NY. Cornell University.

Büring, D. (2007b). *More* or *less*. In *Paper presented at the Chicago Linguistic Society meeting*, Chicago, IL. CLS.

Champely, S. (2018). *pwr: Basic functions for power analysis*. R package version 1.2-2.

Church, A. (1936). An unsolvable problem of elementary number theory. *American Journal of Mathematics*, 58(2):345–363.

Clark, H. H. (1969). Linguistic processes in deductive reasoning. *Psychological Review*, 76:387–404.

Clark, H. H. (1970). How we understand negation. Huntington Beach, CA.

Clark, H. H., Carpenter, P. A., and Just, M. A. (1973). On the meeting of semantics and perception. In Chase, W., editor, *Visual Information Processing*, pages 311–381. Academic Press, New York, NY.

Clark, H. H. and Chase, W. G. (1972). On the process of comparing sentences against pictures. *Cognitive Psychology*, 3:472–517.

Cresswell, M. J. (1976). The semantics of degree. In Partee, B. H., editor, *Montague Grammar*, pages 261–292. Academic Press, New York, NY.

Deschamps, I., Agmon, G., Lewenstein, Y., and Grodzinsky, Y. (2015). The processing of polar quantifiers, and numerosity perception. *Cognition*, 143:115–128.

di Sciullo, M. and Williams, E. (1987). *On the definition of word*. MIT Press, Cambridge, MA.

Embick, D. (2007). Blocking effects in analytic/synthetic alternations. *Natural Language and Linguistic Theory*, 25(1):1–37.

Embick, D. and Noyer, R. (2007). Distributed morphology and the syntax/morphology interface. In Ramchand, G. and Reiss, C., editors, *Oxford Handbook of Linguistic Interfaces*, chapter 9, pages 289–324. Oxford University Press, Oxford, UK.

Gelman, A. and Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, UK.

Giannakidou, A. and Yoon, S. (2011). The subjective mode of comparison: Metalinguistic comparatives in Greek and Korean. *Natural Language and Linguistic Theory*, 29:621–655.

Hackl, M. (2009). On the grammar and processing of proportional quantifers: *most* versus *more than half*. *Natural Language Semantics*, 17:63–98.

Heim, I. (1985). Notes on comparatives and related matters. Unpublished manuscript, University of Texas, Austin.

Heim, I. (2001). Degree Operators and Scope. In Fery, C. and Sternefeld, W., editors, *Audiatur Vox Sapientiae. A Festschrift for Arnim von Stechow*, pages 214–239. Akademie Verlag, Berlin.

Heim, I. (2006). *Little*. In Gibson, M. and Howell, J., editors, *Proceedings of Semantics and Linguistic Theory 16*, pages 35–58, Ithaca, NY. Cornell University.

Heim, I. (2008). Decomposing antonyms? In Gronn, A., editor, *Proceedings of Sinn und Bedeutung 12*, pages 212–225, Oslo. ILOS.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Blackwell, Malden, MA.

Just, M. A. and Carpenter, P. A. (1971). Comprehension of negation with quantification. *Journal of Verbal Learning and Verbal Behavior*, 10:244–253.

Just, M. A. and Clark, H. H. (1973). Drawing inferences from the presuppositions and implications of affirmative and negative sentences. *Journal of Verbal Learning and Verbal Behavior*, 12(1):21–31.

Kennedy, C. (1999). *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*. Garland, New York, NY.

Kennedy, C. (2001). Polar opposition and the ontology of 'degrees'. *Linguistics and Philosophy*, 24:33–70.

Klima, E. S. (1964). Negation in English. In Fodor, J. A. and Katz, J. J., editors, *The Structure of Language: Readings in the Philosophy of Language*, pages 246–323. Prentice Hall.

Lakoff, G. (1970). Linguistics and natural logic. *Synthese*, 22:151–271.

Lidz, J., Halberda, J., Pietroski, P., and Hunter, T. (2011). Interface transparency and the psychosemantics of *most*. *Natural Language Semantics*, 6(3):227–256.

Marantz, A. (1997). No escape from syntax: Don't try morphological analysis in the privacy of your own lexicon. In *University of Pennsylvania Working Papers in Linguistics*, pages 201–225, Ithaca, NY. Cornell University.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Freeman, New York, NY.

May, R. (1977). *The Grammar of Quantification*. PhD thesis, Massachusetts Institute of Technology, Cambridge, MA.

McCawley, J. D. (1988). *The Syntactic Phenomena of English*. Chicago University Press, Chicago, IL.

Moracchini, S. (2018). Evaluativity and structural competition. In *Proceedings of Semantics and Linguistic Theory 28*, Cambridge, MA. MIT.

Morzycki, M. (2011). Metalinguistic comparison in an alternative semantics for precision. *Natural Language Semantics*, 19:39–86.

Peacocke, C. (1986). Explanation in computational psychology: Language, perception and level 1.5. *Mind and Language*, 1(101–123).

Pietroski, P. (2010). Concepts, meanings and truth: First nature, second nature and hard work. *Mind and Language*, 25(3):247–278.

Pietroski, P., Lidz, J., Hunter, T., and Halberda, J. (2009). The meaning of *most*: Semantics, numerosity, and psychology. *Mind & Language*, 24:554–585.

Rett, J. (2015). *The Semantics of Evaluativity*. Oxford University Press, Oxford, UK.

Roberts, S. and Sternberg, S. (1993). The meaning of additive reaction-time effects: Tests of three alternatives. In Meyer, D. E. and Kornblum, S., editors, *Attention and performance 14: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, pages 611–653. MIT Press, Cambridge, MA.

Rullmann, H. (1995). *Maximality in the semantics of* wh-*constructions*. PhD thesis, University of Massachusetts, Amherst, MA.

Seuren, P. A. M. (1973). The comparative. In Kiefer, F. and Ruwet, N., editors, *Generative Grammar in Europe*, pages 528–564. D. Reidel Publishing Company, Dordrecht, NL.

Solt, S. (2015). Q-adjectives and the semantics of quantity. *Journal of Semantics*, 32(221–273).

Solt, S. and Gotzner, N. (2012). Experimenting with degree. In *Proceedings of Semantics and Linguistic Theory 22*, Cambridge, MA. MIT.

Stafford, T. and Gurney, K. N. (2011). Additive factors do not imply discrete processing stages: a worked example using models of the stroop task. *Frontiers in psychology*, 2:287–287.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica, Attention and Performance II*, 30:276–315.

Szabolcsi, A. (2012). Compositionality without word boundaries: *(the) more* and *(the) most*. In *Proceedings of Semantics and Linguistic Theory 22*, pages 1–25, Ithaca, NY. CLC Publications.

Trabasso, T., Rollins, H., and Shaughnessy, E. (1971). Storage and verification stages in processing concepts. *Cognitive Psychology*, 2:239–289.

Van Zandt, T. and Ratcliff, R. (1995). Statistical mimicking of reaction time data: Single-process models, parameter variability, and mixtures. *Psychonomic Bulletin & Review*, 2:20–54.

von Stechow, A. (1984). My reaction to Cresswells, Hellans, Hoeksemas, and Seurens comments. *Journal of Semantics*, 3:183–199.

Wellwood, A. (2012). Back to basics: *more* is always *much-er*. In Chemla, E., Homer, V., and Winterstein, G., editors, *Proceedings of Sinn und Bedeutung 17*, Paris. ENS.

Wellwood, A. (2014). *Measuring predicates*. PhD thesis, University of Maryland, College Park, MD.

Wellwood, A. (2015). On the semantics of comparison across categories. *Linguistics and Philosophy*, 38(1):67–101.