Decomposition and processing of negative adjectival comparatives

Daniel Tucker, Barbara Tomaszewicz and Alexis Wellwood

Abstract Recent proposals in the semantics literature hold that the negative comparative *less* and negative adjectives like *short* in English are morphosyntactically complex, unlike their positive counterparts *more* and *tall*. For instance, the negative adjective *short* might decompose into LITTLE TALL (Rullmann 1995; Heim 2006; Büring 2007; Heim 2008). Positing a silent LITTLE as part of adjectives like *short* correctly predicts that they are semantically opposite to *tall*; we seek evidence for this decomposition in language understanding in English and Polish. Our visual verification tasks compare processing of positive and negative comparatives with *taller* and *shorter* against that of less symbolically-rich mathematical statements, $A > B$, $B < A$. We find that both language and math statements generally lead to monotonic increases in processing load along with the number of negative symbols (as predicted for language by e.g. Clark and Chase 1972). Our study is the first to examine the processing of the gradable predicates *tall* and *short* cross-linguistically, as well as in contrast to extensionally-equivalent, and putatively non-linguistic stimuli (cf. Deschamps et al. 2015 with quantificational determiners).

1 Introduction

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 2007; 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. We look for evidence of such decomposition in processing,

Daniel Tucker
Northwestern University, 2016 Sheridan Road, Evanston, IL 60208, USA, e-mail: danieltucker2017@u.northwestern.edu

Barbara Tomaszewicz
Universität zu Köln, Albertus-Magnus-Platz, 50923 Köln, Germany, e-mail: btomasze@uni-koeln.de
Uniwersytet Wrocławski, Instytut Filologii Angielskiej, Kuźnicza 22, 50-138 Wrocław, Poland

Alexis Wellwood
Northwestern University, 2016 Sheridan Road, Evanston, IL 60208, USA, e-mail: wellwood@northwestern.edu

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 (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.

We test for these effects with *taller* (positive) and *shorter* (negative). If negative 'features' are specifically linguistic, then it is possible that such an asymmetry might not be observed with the processing of mathematical statements like $A > B$ and $A < B$. Deschamps et al. (2015) tested a similar hypothesis in their study of *more/less than half* and *many/few*, contrasting processing of those expressions with that of extensionally-equivalent, quasi-algebraic inequalities. They found that the sentences with relevantly negative quantifiers in English took longer to process than the corresponding ones with positive quantifiers, but no such asymmetry was observed for the analogous math statements.

This paper 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. Like Deschamps et al. (2015), we examine the effects of polarity on processing linguistic and non-linguistic statements; unlike those authors, we 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 so can support a finer-grained picture of the underlying cognitive processes involved in these tasks.

Our investigation is broadly compatible with research conducted under the banner of the Interface Transparency Thesis, offered to precisify a representational role for formal semantics within the broader project of cognitive science (Lidz et al. 2011). The idea is that cognition, by default, carries out procedures that align with the operations specified in the semantic representation of a sentence. If a thesis like this is correct, investigations of processing will be a useful tool for understanding the nature of speakers' semantic representations in general, in addition to paving the way for tests that mediate between specific representational proposals.

In what follows, we first discuss the recent proposals for decomposition in negative adjectival comparatives in order to motivate our processing studies (Section 2.1). Next, we recall both early and recent results investigating the processing of 'implicit' negation in cognitive psychology and in linguistics (Section 2.2). Then, we present the results of a sentence-to-picture verification task in English (Section 3) and in Polish (Section 4). To preview, our results provide support for the decompositional analysis of forms like *shorter* in both languages. Section 5 concludes.

2 Background & motivation

Positive gradable adjectives like *tall* are morphemes—they are not amenable to further morphological analysis. However, Büring's (2007) 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, the antonym of an adjective like *long* is formed by two pieces, i.e. *kawo-hra*,

which Bobaljik (2012) glosses as 'long-not'. Our research brings to bear a new kind of evidence for these questions through an examination of gradable adjectives like *tall* and *short* in English and Polish, seeking a different kind of evidence for decomposition in sentence processing.

In this section, we motivate our experimental project: Section 2.1 reviews the decompositional approach in semantics, and 2.2 discusses relevant contemporary and classic literature that informs our linking hypotheses.

2.1 Morphosyntax and semantics of shorter

In the contemporary degree semantics tradition, *tall* is analyzed as involving a relation between individuals and their heights, and a sentence like (1a) 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 (1b)?

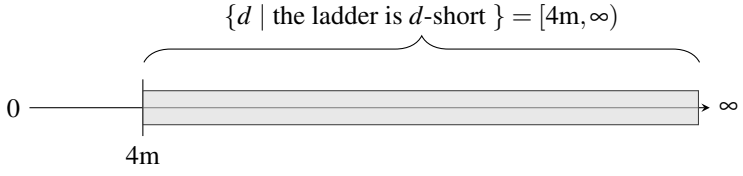
(1) a. Al is taller than Bill is.
b. Bill is shorter than Al is.

(1a) and (1b) stand in a mutual entailment relationship: competent speakers of English intuitively infer that if (1a) is true, (1b) 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, p56).

Büring's (2007) decompositional approach, in contrast, supports an analytic relationship between (1a) and (1b). His analysis begins by considering Kennedy's (2001) explanation of the oddity of (2), 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 (3), while TALL maps *the building* to a set of degrees like that in (4).[1] What Heim (2008) calls *Kennedy's constraint* is that -ER cannot compare positive and negative degrees.

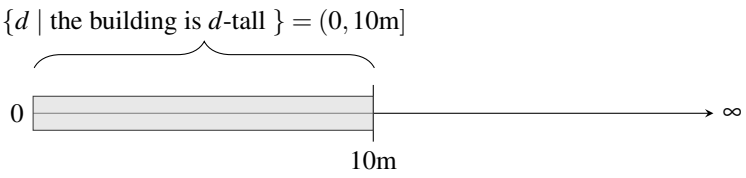
(2) ? The ladder is shorter than the building is tall. ?HEIGHT

(3)



[1] 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).

(4)



Büring points out that, as given, Kennedy's explanation for (2) incorrectly predicts that (5) 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, (5) should also be anomalous.

(5) 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 equivalent 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 (1b) to be analyzed as ambiguous between two structures, one containing the bundling [LITTLE-ER] TALL and the other -ER [LITTLE TALL]. (5) 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 (2) and (5) as follows. In principle, there could be two bracketings for (2), but either would be problematic. On the bundling -ER [LITTLE TALL] for *shorter*, (2) 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, (2) 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 (2) and (5), Büring's account extends to cases of ambiguity with *less high* and *lower* that are not evidenced by comparatives with their antonym *higher*, (6a)-(6c) (Seuren 1973, Rullmann 1995). (6a) describes a helicopter flying some degree higher than the maximal height a plane can safely fly, while both (6b) and (6c) 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.)

(6) 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 (7a) 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), (7b) cannot be read this way: (7b) only has the reading where whatever speed Polly drives, it *has to* be less than Larry's.

(7) 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 (2) and (5). 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 (8) 'feel different'; if (8a) has an instance of a covert LITTLE, and (8b) results from LITTLE morphologically exerting itself on the adjective, why does (8b) seem more difficult to understand than (8a)?[2]

(8) 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 our focus. Rather, we assume that the linguistic evidence amassing in favor of a decompositional analysis of *shorter* is strong, at least strong enough to warrant further investigation. Our interest is in the fact that decompositional proposals can be seen to make explicit predictions about sentence comprehension.

2.2 Relating language and vision

How can the decompositional approach be tested in processing? In what follows, we 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, we 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 60s and early 70s 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, we 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*. We will be more explicit about this shortly.

[2] 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 (7).

Separately from the representational assumptions, models of sentence-picture matching were designed to account for the response latencies of judgments in extremely simple tasks.[3] 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. We focus on the first type of task, since it will be most relevant for our own experiments (though see Section 3.4).

On the "Sentence-First Model" (Clark and Chase, 1972), the process of comparing a sentence with a picture proceeds in four stages, summarized in (9). 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.

(9) **"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 1969b calls this the 'principle of congruence'). However, it would be 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 $\textsc{below}(A,B)$ and visual $\textsc{above}(B,A)$ should then be judged as 'false', which would be incorrect.

Thus, according to Clark & Chase (1972, p. 478), "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.); we will attempt to remain at a fairly informal level here.

So what parameters affect the latency of a participant's judgment, and how? 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 (10). 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 & Chase 1972, p. 487). Finally, there is an overall and independent cost of t_0 for the time to plan and execute the response.

[3] 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.

(10) **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

Somewhat differently 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. We also test the processing of math expressions against expressions in natural language (English and Polish), asking whether the 'simpler' math expression leads to different effects. Our study differs in that we test comparative adjectives, provide a shorter viewing time for the picture (theirs was 2500-2800ms), and we include tests for congruence effects.[4]

3 Experiment 1: English sentence-picture matching

We 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, we expected *shorter* should take longer to process than *taller*. We contrast this processing with that of prima facie 'simpler' mathematical statements like '$A > B$' and '$A < B$'.

3.1 Design & participants

We designed a sentence-to-picture verification task in a two-2x2 design according to linguistic and non-linguistic statements. In our task, participants were presented with a statement, followed by a picture, and asked to judge whether the statement accurately described the picture. Each of our two-2x2 sub-designs corresponded to the 'language' that the statement was presented in, either English or Math.

For each of the English and Math sub-designs, we manipulated POLARITY (positive, negative) and CONGRUENCE (congruent, incongruent). As can be seen in Table 1, we 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 8 statements. The factor CONGRUENCE varied whether the statement was true of the paired picture or not, corresponding to the congruent and incongruent conditions, respectively.

[4] A further difference is that Deschamps et al. (2015) presented their linguistic statements auditorily.

	English	Math
Positive	*A is taller than B*, *B is taller than A*	$A > B$, $B > A$
Negative	*A is shorter than B*, *B is shorter than A*	$A < B$, $B < A$

Table 1: English and Math statements used in Experiment 1.

Stimuli. We created 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 1 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 160 trials.

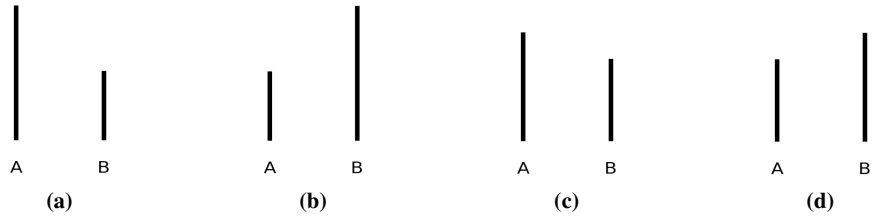


Fig. 1 Sample picture stimuli used in Experiment 1.

Procedure. The experiment was designed using jsPsych, a JavaScript library for creating behavioral experiments in a web browser (de Leeuw 2015). After consenting to participate, participants were presented with instructions for the experiment (see below). Following this, participants completed the 160 trials,[5] 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 spacebar. 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 4 blocks, each defined by one combination of linguistic/non-linguistic statements and line order (A first vs. B first). The order of presentation of the blocks and of the trials within the blocks was completely randomized.

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

> Welcome to the experiment!

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

> There are 160 trials in this experiment. Each trial will consist of a statement, an image, and your response. The statement may be in a natural or mathematical language. You will have as much time as you wish to view the 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.

Participants. We recruited 15 participants through a Human Intelligence Task (HIT) posted on Amazon's Mechanical Turk. We 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. No Mechanical Turk master workers were recruited for this study.

3.2 Predictions

We assume the decompositional analysis of English negative comparatives in line with Büring (2007), the 'simple' hypothesis about math statements, and combine these assumptions with the predictions of the Sentence-First model of Clark and Chase (1972). In what follows, we discuss the predictions for English and math statements separately, and in turn.

Linguistic stimuli (English). 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 (11). The major operand of the semantic representation is ER, which specifies a greater-than relation between two quantities. These quantities are provided by $\textsc{tall}(A)$ and $\textsc{tall}(B)$ in (11a), and by an operation over such quantities (e.g. complementation) provided by LITTLE, (11b).

(11) a. $[\![\text{A is taller than B.}]\!] = \textsc{er}(\textsc{tall}(\text{A}),\textsc{tall}(\text{B}))$

b. $[\![\text{A is shorter than B.}]\!] = \textsc{er}(\textsc{little}(\textsc{tall}(\text{A})), \textsc{little}(\textsc{tall}(\text{B})))$

In light of the early cognitive psychology literature, we expected that the added presence of LITTLE should correspond to an increase in processing load: processing (11b) requires to processing something like (11a) in addition to the contributions of the two instances of LITTLE. Such additional processing steps should correspond to an increase in RTs. Furthermore, we 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, these two 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 we 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 2.[6]

[6] 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

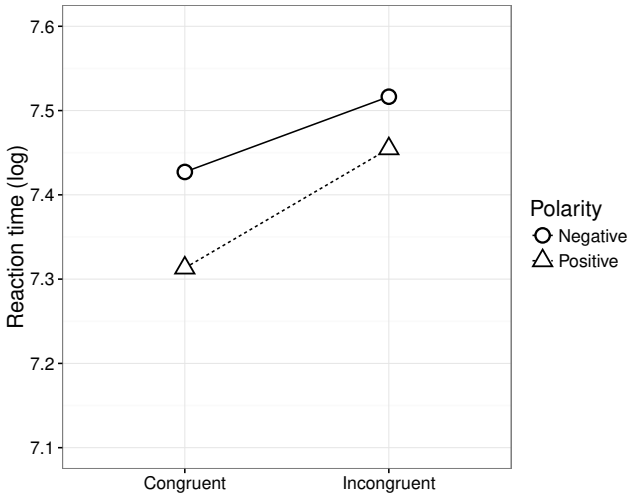


Fig. 2 Predicted main effects of polarity and congruence for natural language, given the decompositional analysis of forms like *shorter* and the Sentence-First model of Clark and Chase (1972).

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. We thus expected overall error rates to be similar in our task: broadly, higher RTs should pattern with higher error rates.

Non-linguistic stimuli (Math). Our expectations for Math statements are somewhat less clear. On the one hand, Deschamps et al. (2015) report no effect of polarity on processing quasi-algebraic inequalities, in contrast to English sentences with *many/few* and *more/less*. Such an expectation aligns with the 'simple' hypothesis that statements like $A > B$ and $A < B$ are essentially non-linguistic, and representationally transparent (i.e. non-decompositional), and so should be processed differently than linguistic statements.

However, we might expect an effect of congruence here—whether the statement matches the scene. Clark and Chase's (1972) characterization of congruence effects was that they were essentially an independent consequence of comparing two mismatching representations. In light of this, we do not expect such effects to apply only to linguistic statements. This amounts to the expectation that incongruent situations will lead to increased RTs for processing Math statements.[7]

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

[7] As noted above, congruence effects were not discussed by Deschamps et al. (2015).

3.3 Analyses & exclusions

We 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, we 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 we report are derived from comparison of the maximal model in each case, against the same model minus the relevant parameter.

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). We 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.

Of the 2400 datapoints we collected, 45 were excluded (1.9%) for either a missed response (i.e., the participant failed to respond within the 5s time window), or because the response time was greater than three standard deviations from that participant's mean RT. Each main effect reported in the next section was based on an average of 585 observations per condition, while each interaction was based on an average of 295 observations per condition.

All analyses were conducted using R's *lme4* package (Bates et al. 2014).

3.4 Results: RTs

We conducted two separate linear mixed effects model comparisons on the log-transformed RT data. The results for both English and Math are presented in Figure 3.

3.4.1 Linguistic conditions (English)

Participants took longer to evaluate sentences with *shorter* than with *taller*. This was reflected in a robust main effect of POLARITY (means: negative 6.21, positive 6.07, $\beta = -.14$, $SE = .03$, $\chi^2 = 12.56$, $p < .001$) in the predicted direction: RTs in the negative conditions were longer than in the positive conditions (means, in ms: negative 634.12ms, positive 586.70ms).

Additionally, participants took longer to reject false statements than to accept true statements. This was reflected in a marginal main effect of CONGRUENCE (means: congruent 6.32, incongruent 6.39, $\beta = -.09$, $SE = .05$, $\chi^2 = 3.34$, $p = .067$), in accord with our 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 593.10ms, incongruent 627.53ms).

Moreover, 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 ($\beta = -.13$, $SE = .07$, $\chi^2 = 3.89$, $p = .048$). RTs in the positive congruent condition were shorter than in the negative congruent condition (means: negative 6.21, positive 5.99; means, in ms: negative 636.99ms, positive 549.36ms), while there was little difference between the negative incongruent condition and the positive incongruent condition (means: negative 6.22, positive 6.15; means, in ms: negative 631.21ms, positive 623.90ms).

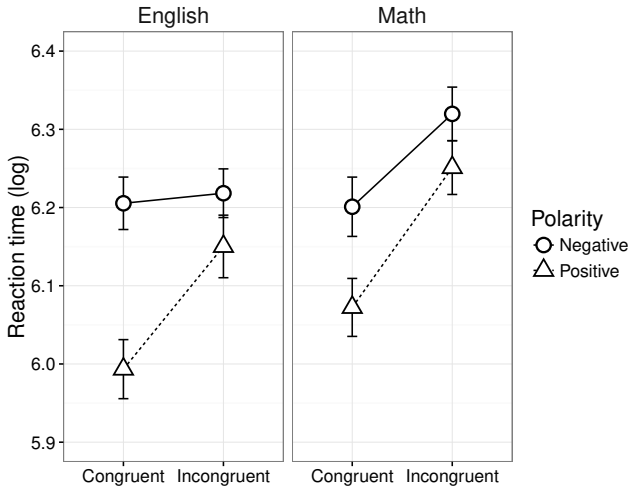


Fig. 3 Mean log RTs and SEs by POLARITY and CONGRUENCE for the linguistic (English) and non-linguistic (Math) sub-experiments of Experiment 1.

3.4.2 Non-linguistic conditions (Math)

Participants took longer to evaluate Math statements with $<$ than with $>$. This was reflected in a strong main effect of POLARITY (means: negative 6.26, positive 6.16, $\beta = -.10$, $SE = .04$, $\chi^2 = 6.40$, $p = .01$), in which reaction times in the negative conditions were substantially longer than for the positive conditions (means, in ms: negative 678.06ms, positive 610.35ms). These results stand in contrast to Deschamps et al. (2015), who report no asymmetry in the evaluation of positive and negative Math statements.

Participants also took longer to reject Math statements that didn't match the picture than to reject those that did. This was reflected in a strong main effect of CONGRUENCE (means: congruent 6.14, incongruent 6.29, $\beta = -.15$, $SE = .04$, $\chi^2 = 9.45$, $p = .002$): the incongruent conditions took longer to evaluate than the congruent conditions (means, in ms: congruent 602.63ms, incongruent 685.76ms). This congruence effect was expected as reflecting a general cost of rejecting false statements.

All of the effects of congruence were accounted for in the main effects, in contrast to our results for English. That is, there was no interaction between POLARITY and CONGRUENCE ($\beta = -.07$, $SE = .11$, $\chi^2 = .40$, $p > .5$). RTs in the positive congruent condition were faster than in the negative congruent condition (means: negative 6.20, positive 6.07; means, in ms: negative 636.63ms, positive 568.75ms). Similarly, RTs in the positive incongruent condition were faster than in the negative incongruent condition (means: negative 6.32, positive 6.25; means, in ms: negative 719.07ms, positive 652.12ms).

3.5 Results: accuracy

To assess response accuracy (a binary variable), we conducted model comparisons over mixed effects logistic regressions. The results are presented graphically in Figure 4, with accuracy plotted in terms of the percentage of correct responses summarized by participant in each condition.

3.5.1 Linguistic conditions (English)

Our participants' accuracy was not any 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 94.0%, positive 93.9%, $\beta = .02$, $SE = .31$, $\chi^2 < .01$, $p > .9$). This result is unexpected in light of the early cognitive psychology literature, which found an inverse correlation between reaction time and response accuracy.

Participants were no less accurate at rejecting false statements than at accepting true statements. That is, we found no effect of CONGRUENCE on mean response accuracy (means: congruent 94.2%, incongruent 93.7%, $\beta = .22$, $SE = .24$, $\chi^2 = .80$, $p = .4$): a statement's veracity with respect to its accompanying picture made little difference.

Analyses revealed no interaction was found between POLARITY and CONGRUENCE ($\beta = -.05$, $SE = .48$, $\chi^2 = .01$, $p > .9$); there was no difference in mean response accuracy in the negative versus positive congruent conditions (means: negative 94.2%, positive 94.2%). Such was also the case in the negative and positive incongruent conditions (means: negative 93.4%, positive 93.6%).

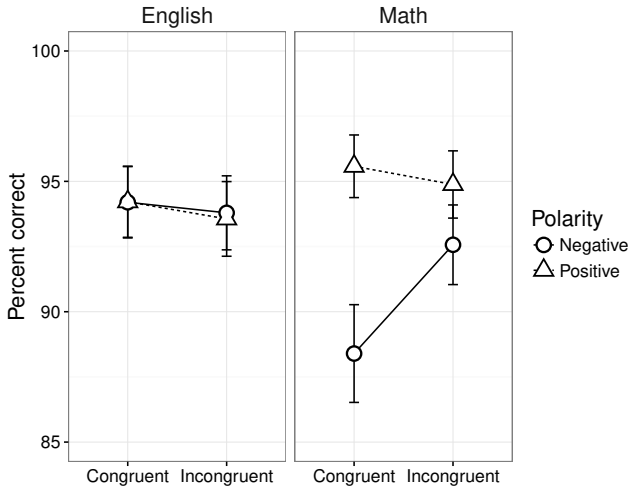


Fig. 4 Mean subject accuracy and SE by POLARITY and CONGRUENCE for the linguistic (English) and non-linguistic (Math) sub-experiments of Experiment 1.

3.5.2 Non-linguistic conditions (Math)

In contrast to the results for English, participants were less accurate at evaluating sentences with $<$ than with $>$. This was revealed in a marginal main effect of POLARITY (means: negative 90.5%, positive 95.2%, $\beta = .78$, $SE = .36$, $\chi^2 = 3.83$, $p = .05$): average response accuracy was lower for the negative conditions than for the positive conditions.

Similar to the results for English, participants were as accurate at rejecting false statements as accepting true statements. That is, we found no main effect of CONGRUENCE on mean response accuracy (means:

congruent 92.0%, incongruent 93.7%, $\beta = -.22$, $SE = .28$, $\chi^2 = .57$, $p > .5$): whether the sentence was true of the picture made little difference to average response accuracy.

Finally, no interaction was found between POLARITY and CONGRUENCE ($\beta = .78$, $SE = .50$, $\chi^2 = 2.31$, $p = .1$); accuracy was lower in the negative congruent condition than in the positive congruent condition (means: negative 88.4%, positive 95.6%); such was also the case for the negative and positive incongruent conditions (means: negative 92.6%, positive 94.9%).

3.6 Discussion

In Experiment 1, we 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 (in both English and Math). 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). We also found an interaction effect that was not observed in earlier works.

In the Math sub-experiment, we found that statements with $<$ took longer to process than statements with $>$, and that statements which were false of the accompanying picture took longer to process than statements that were true of the picture. In this sub-experiment, we found no interaction effect, suggesting that these results provided a better match to the predictions of the Sentence-First model proposed by Clark and Chase (1972). This is not what a simple hypothesis about how math statements are processed would predict, and it contrasts to the findings of Deschamps et al. (2015), who found that that processing quasi-algebraic inequalities was qualitatively different than the processing of natural language. We do not have a good explanation for why their results differ from ours.[8]

We found one major difference between the English and Math sub-experiments, which was the interaction between polarity and congruence. Evaluating true and false statements with *shorter* took roughly the same amount of time, however evaluating true statements with *taller* was much faster than evaluating false statements with *taller*. We did not find a corresponding effect in the Math sub-experiment. What could explain this difference?[9] One possibility, again considering the discussion in Clark and Chase (1972), is that our speeded task involves a different sort of processing for English than for their Math correspondents.

One line of inquiry is suggested by considering the results that those authors found testing sentences with *above* and *below* in a Picture-to-Sentence verification task, as in Figure 5. On the surface, the "Sentence-First" processing model in (9) and a "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 cue, sentence encoding could impact picture encoding.

[8] It is possible that our participants understood the math statements in terms of natural language translations like *A is greater/less than B*, which lead to the language-like effects. The quasi-algebraic expressions tested in Deschamps et al. (2015) consisted of blue and yellow squares on both sides of the $>$ and $<$ operators. Such representations might be less likely to be translated into natural language than $A > B$ and $A < B$, potentially accounting for the differences between our study and theirs.

[9] An anonymous reviewer notes that we so far have not directly compared these two sub-experiments, and so haven't shown that they are statistically different from one another. Conducting a post-hoc LMEM comparison over the combined data from the English and Math sub-experiments, we found no main effect for the contrast-coded factor LANGUAGE (English vs. Math), nor any interactions with that factor. Subsequently, in the text, we focus on the qualitative difference that can be seen in Figure 3, and which was borne out in the independent 2x2 analyses.

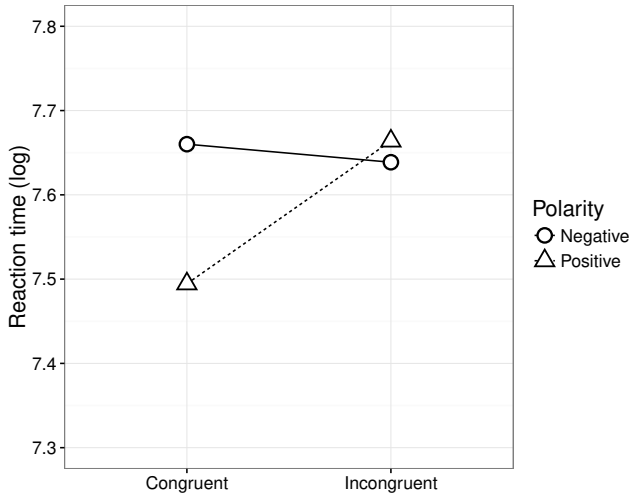


Fig. 5 Results of Clark and Chase's (1972) Picture-to-Sentence verification task with *above* (positive) and *below* (negative), modeled after the presentation in Clark et al. (1973).

That is, the Sentence-First model assumes that the representation of the sentence formed during Stage 1 impacts how the picture is encoded during Stage 2. Clark and Chase assumed that given a sentence specified with *above*, the picture will be encoded in terms of the matching ABOVE relation, and given a sentence specified with *below*, the picture will be encoded with the matching BELOW relation. Increased latencies for polarity are seen to arise due to the negative feature on *below* (*a* - Below time), and for congruence due to the mismatching subjects (*b* - Falsification time). An instance of this type of processing is shown schematically in (12).

(12) **"Sentence-First" processing for a negative incongruent trial, Clark & Chase (1972)**

a. **Stage 1**: Read *A is below B* $\Rightarrow$ BELOW(A,B) $+a$

b. **Stage 2**: See picture of A above B $\Rightarrow$ BELOW(B,A)

c. **Stage 3**: Are *A* and *B* in the same position in the relation? $\Rightarrow$ No $+b$

d. **Stage 4**: Respond with button press $+t_0$

The major surface difference between this model and the "Picture-First" model is the latter's 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.

Importantly, our Experiment 1 differs from these earlier studies in that we imposed a 200ms viewing time for the picture, a threshold more often imposed in contemporary experiments (see Section 2.2). It is possible that the 'preference' to encode visual scenes in positive terms manifests as a necessity under this kind of time pressure given that it takes approximately 200ms to initiate a regular saccade movement in response to an unexpected stimulus—with an expected stimulus, peripheral vision may be sufficient (Carpenter 1977, Allopenna et al. 1998).

Thus, there may be a way of thinking about the processing demands imposed in the present task which is relevant to predicting the differences between the English and Math sub-experiments. Suppose that the scene is always encoded positively under the 200ms time pressure, and that encoding a statement negatively always imposes its own cost (x). Assume that there is an additional 'check' imposed for matching English with the picture on whether the subject of the sentence corresponds to the first position of the (positive) relation encoded by vision (y).[10] Along with the cost of congruence (z), the sum of the processing costs would be as in (13) for one type of trial.

(13) **English negative incongruent trial**

a. language: ER(LITTLE(TALL(A)), LITTLE(TALL(B)))
b. vision: ER(TALL(A),TALL(B))
c. unmarked English form? NO. $+x$
d. subjects match? YES.
e. congruent representations? NO. $+z$
f. Respond with button press $+t_0$

If one applies this reasoning to each of the conditions of the English sub-experiment, we might predict the relative magnitudes of the effects that we found (positive congruent t_0, positive incongruent t_0+y+z, negative congruent t_0+x+y, negative incongruent t_0+x+z). In contrast, if the Math task imposes no such 'check' on whether the 'subject' of the statement corresponds to the first position of the (positively-encoded) visual representation, we might predict the relative magnitudes we found there as well (positive congruent t_0, positive incongruent t_0+z, negative congruent t_0+x, negative incongruent t_0+x+z).

Of course, this is a post-hoc analysis, and it remains unclear why checking the match for 'subject' would differ between English and Math statements in this task (apart from the fact that statements like $A > B$ might not necessitate a notion of 'subject'). However, we do observe a clear difference between English and Math, and it is possible that probing the effects of this type of task on processing with different types of statements could provide new insight into how statements are matched with pictures, and why this might differ across 'languages'.

Regardless, limiting participants to 200ms appears to have had an important effect on the task demands, at least in the case of sentence-to-picture matching with natural language. We have suggested that, in this case, participants could be relying on a bias to positively-encode a scene in order to actually perform the task under these pressures. In the next experiment, we design a very similar task, but do not impose such stringent restrictions on how long participants have to view the scene.

With respect to response accuracy, it is unclear why the error rates did not pattern with response latencies for *taller* and *shorter* as they did in work on *above* and *below*. As an anonymous reviewer suggests, this could be due to a ceiling effect in the English sub-experiment, since accuracy rates there were very high across the board. In the math sub-experiment, however, accuracy patterned with the RT data: accuracy was lower for statements with $<$ than with $>$, suggesting greater difficulty with the more 'negative' of the pair . However, why the English and Math sub-experiments should differ in this respect is a matter we leave for future research.

[10] Clark and Chase (1972) point to studies by Huttenlocher (1969) and Clark (1969b,a) for evidence that the 'theme/rheme' distinction is important in these tasks, which is reflected in the specific type of comparison operation that Clark and Chase posit for Stage 3, shown in (12).

4 Experiment 2: Polish sentence-picture matching

We test the predictions of decompositional analyses of negative comparatives in Polish, contrasting this with the processing of math statements. Polish *wyższe* and *niższe* comparatives, (14), have a very similar underlying syntactic structure to those of English (Pancheva 2006), and thus will provide an interesting test for the cross-linguistic robustness of the decompositional proposals so far offered explicitly only for English.

(14) a. A jest wyż-sze niż B
A is tall-er than B
b. A jest niż-sze niż B
A is short-er than B

4.1 Design & procedure

We conducted a sentence-to-picture verification task like Experiment 1, in a two-2x2 design according to linguistic (Polish) and non-linguistic (Math) statements. As before, participants were presented with the statement, followed by a picture, and asked to judge whether the statement matched the picture. Experiment 2 differed in that it was conducted while participants' eyes were tracked. However, because the eye-movement data is not relevant for our reaction time hypotheses, we do not investigate that data here.

As in Experiment 1, for each of the Polish and Math sub-designs we manipulated POLARITY (positive, negative) and CONGRUENCE (congruent, incongruent). The 8 statements we tested, sorted by the levels of these factors, are shown in Table 2. Participants were presented with images very much like those presented above in Figure 1, except that the two lines were spaced further apart (see below). Unlike in Experiment 1, we allowed participants up to 4 seconds to view the image; this viewing time is more consonant with that employed in the early cognitive psychology studies.

	Polish	Math
Positive	*A jest wyższe niż B*, *B jest wyższe niż A*	$A > B, B > A$
Negative	*A jest niższe niż B*, *B jest niższe niż A*	$A < B, B < A$

Table 2: Polish and Math statements used in Experiment 2.

Stimuli. 20 pictures featuring two lines marked A and B were paired with the 8 statements in Table 2 for a total of 80 pairings. This is half of the number of pairings featured in Experiment 1 in order to keep the time required to complete the experiment under 20 minutes. Each of the 8 conditions was presented with 10 of the 20 pictures in an alternating fashion (counterbalancing how many images with line lengths of 28 pixels and 42 pixels were presented, and in how many of them the line labeled A or the line labeled B was longer). As in Experiment 1, the shorter line appeared in one of two sizes (24 or 42 pixels), and differed from the longer line by one of five different length ratios (.5, .75, .833, .875, .9). The distance between the two lines was 700 pixels, substantially larger than in Experiment 1. This was in order to facilitate tracking of participants' eye-movements during picture verification, and to prevent encoding the scene solely using peripheral vision. As noted above, we only report the behavioral results in this paper.

Procedure. The experiment was run on a Windows PC connected to the SR Research EyeLink 1000 Plus eye-tracker. The participants were first presented with the printout of the instructions and the experimenter answered any questions. Participants then saw the same instructions on the screen followed by a practice session with trial structure parallel to the experimental trials but with different statements and pictures. In a trial, a statement was presented until button press, followed by a picture displayed for up to 4s. Participants pressed the left arrow key on the response box when they decided that the picture matched the sentence, and the right arrow key when they decided that it did not. Accuracy was encouraged by an auditory signal in case of a wrong response. When a response was recorded, or no response was made during the 4s second window, the picture disappeared and a new trial began. Each sentence display and each picture display was preceded by a 1s pause followed by a fixation point (during which drift-correction was performed). Trials were organized into 4 blocks, each defined by one combination of linguistic/non-linguistic statements and POLARITY. Block order and trial order within blocks was pseudo-randomized (no more than three trials of the same type consecutively). After each block participants were able to take a short break. The experiment took approximately 20 minutes to complete.

Instructions to participants. The Polish instructions given to participants are presented below translated into English. Unlike in the previous study, we explicitly indicated that participants should attempt to make their judgment as accurately (as opposed to as quickly) as possible.

> Welcome and thank you for your participation in our experiment!
>
> Your task is to read short sentences and mathematical expressions and to decide if they match the pictures. The accuracy of your responses matters. Before the experiment, there will be a practice session, where you will see some examples.
>
> We begin with the process of CALIBRATION: You will see a black point. Look at its white center. The point will be appearing in a different locations. Track the point.
>
> When calibration is successful, we begin the practice session. First, you will see a cross. Look at its center. Next, the text will appear. When you read it, press the bottom button. You will then see a point. Look at its center. Now the picture will appear. Decide whether the picture matches the text.
>
> If it does, press the LEFT button for YES.
>
> If it doesn't, press the RIGHT button for NO.
>
> Do the same for the following pairs of text and pictures.
>
> The accuracy of your responses matters. If you make a mistake, you cannot go back or repeat.
>
> Are you ready for calibration and practice session? Press the bottom button to begin.

Participants. 32 participants were recruited from the University of Wrocław, Poland. One participant was excluded due to calibration failures. Participants received a payment equivalent to 9 EUR for participation.

4.2 Predictions

We assumed the decompositional analysis as extended to Polish negative comparatives, and that simple Math statements would be processed similarly. We again test the predictions of the Sentence-First model of Clark and Chase (1972), and discuss these separately for the Polish and Math sub-experiments.

Linguistic stimuli (Polish). In Experiment 1, the pattern of RTs for English was different than that predicted by the Sentence-First model. We speculated in Section 2.2 that this was due to the 200ms time window in

which participants had to view the picture, which seems to have selectively impacted the processing of natural language. A 200ms constraint on the presentation time prevents the initiation of a saccades in response to an unexpected stimulus (Carpenter 1977, Allopenna et al. 1998), and thus may have contributed to this pattern. With a 4s window for viewing the picture, the predictions of Clark and Chase's (1972) model of Sentence-to-Picture verification should unambiguously apply. Moreover, since the Polish comparative sentences in Table 2 are compatible with the same syntactic and semantic analyses as their English counterparts in Table 1, any such effects can be attributed to decomposition. Thus, we predicted a main effect of POLARITY, with longer RTs corresponding to the processing of the two instances of LITTLE. We also predicted a main effect of CONGRUENCE—whether the statement was true of the picture. As in the earlier studies reported in the literature, we expect only additive effects of these two factors (i.e., no interaction).

Non-linguistic stimuli (Math). If the processing of Math statements in Experiment 1 was reflective of the processing of such statements regardless of the viewing time for the picture, we expected to replicate the main effects of POLARITY and CONGRUENCE. If the restricted viewing time did have an impact, the predictions here are less clear.

4.3 Analyses & exclusions

Similar to Experiment 1, all results we report reflect mixed effects model comparisons with maximal random effects structure. Out of 2480 observations collected for this experiment, 14 observations were excluded as missed responses (approximately .57% of the data). As with the previous experiment, results for RT measures are plotted and reported in log space, but also reported in milliseconds (ms) in the prose for ease of interpretation.

4.4 Results: RTs

We report the results of linear mixed effects regressions on the Polish and Math RT data. The results are presented graphically in Figure 6.

4.4.1 Linguistic conditions (Polish)

Participants took longer to process Polish negative comparatives than positive comparatives. This was reflected in a main effect of POLARITY (means: negative 7.14, positive 7.09, $\beta = -.06$, $SE = .02$, $\chi^2 = 7.48$, $p < .01$) in the predicted direction: the negative conditions took longer to evaluate overall (means, in ms: negative 1384.83ms, positive 1340.55ms).

Participants also took longer to reject false statements than to accept true statements. This was reflected in a robust main effect of CONGRUENCE (means: congruent 7.06, incongruent 7.17, $\beta = -.11$, $SE = .02$, $\chi^2 = 18.17$, $p < .01$), in accord with our predictions: a statement's truth or falsity with respect to its accompanying picture made a substantial difference to verification response latency (means, in ms: congruent 1279.27ms, incongruent 1446.11ms).

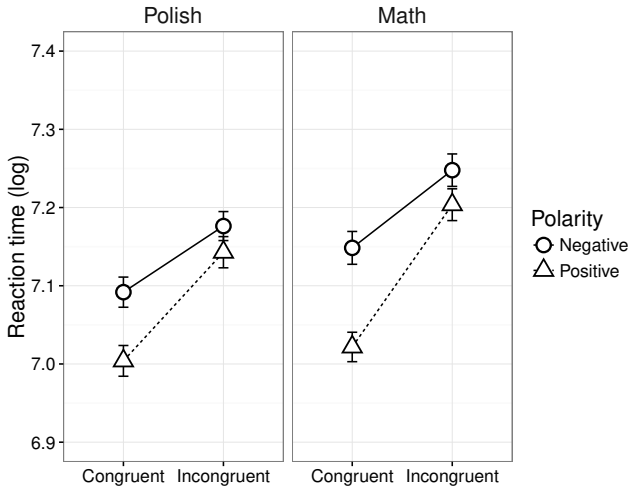


Fig. 6 Mean log RTs and SEs by POLARITY and CONGRUENCE for the linguistic (Polish) and non-linguistic (Math) sub-experiments of Experiment 2.

These two effects were only additive in our data. There was no interaction of POLARITY and CONGRUENCE ($\beta = -.08$, $SE = .09$, $\chi^2 = .77$, $p > .1$); RTs in the positive congruent condition were somewhat faster than in the negative congruent condition (means: negative 7.10, positive 7.01; means, in ms: negative 1317.54ms, positive 1241.36ms), while a smaller difference in the same direction held in the incongruent conditions (means: negative 7.18, positive 7.16; means, in ms: negative 1451.46ms, positive 1440.71ms).

4.4.2 Non-linguistic stimuli (Math)

Participants took longer to process statements with $<$ than with $>$. This was reflected in a main effect of POLARITY (means: negative 7.22, positive 7.12, $\beta = -.09$, $SE = .02$, $\chi^2 = 23.23$, $p < .01$): the negative conditions took longer to process than the positive conditions (means, in ms: negative 1524.02ms, positive 1377.45ms). This result is similar to the results for Math in our Experiment 1, and again stand in contrast to the results reported in Deschamps et al. (2015).

Again, participants took longer to reject false statements than to accept true statements. This was reflected in a main effect of CONGRUENCE (means: congruent 7.09, incongruent 7.25, $\beta = -.16$, $SE = .02$, $\chi^2 = 71.40$, $p < .01$): the incongruent conditions had longer associated RTs than the congruent conditions (means, in ms: congruent 1328.17ms, incongruent 1573.53ms). This result replicates Experiment 1, and was predicted if there is a general cost for judging statements to be false.

As in the Experiment 1 Math sub-experiment, these effects were only additive. That is, we found no interaction between these two factors ($\beta = -.05$, $SE = .09$, $\chi^2 = .32$, $p > .1$). RTs in the positive congruent condition were marginally faster than in the negative congruent condition (means: negative 7.15, positive 7.03; means, in ms: negative 1405.18ms, positive 1251.17ms); a similar pattern was observed in the incongruent conditions (means: negative 7.29, positive 7.22; means, in ms: negative 1645.22ms, positive 1501.38ms).

4.5 Results: accuracy

Similar to Experiment 1, we assessed response accuracy via mixed effects logistic regression model comparisons. The results for mean response accuracy summarized by participant by condition for both linguistic and non-linguistic statements are presented in Figure 7.

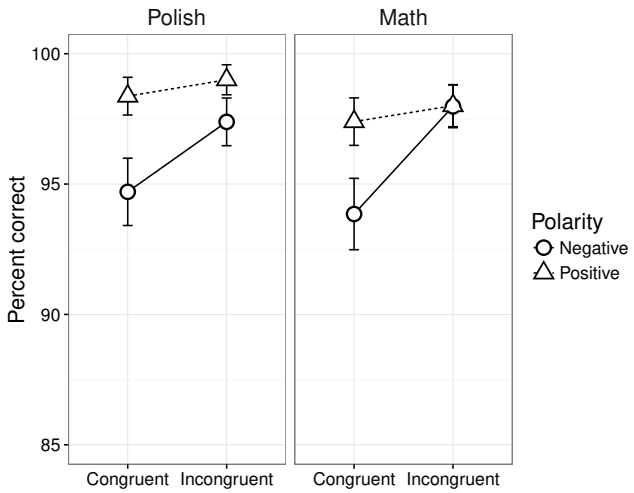


Fig. 7 Mean subject accuracy and SE by POLARITY and CONGRUENCE for the linguistic (Polish) and non-linguistic (Math) sub-experiments of Experiment 2.

4.5.1 Linguistic conditions (Polish)

Unlike in the Experiment 1 English sub-experiment, participants made more errors on the negative comparatives than the positive comparatives. This was reflected in a main effect of POLARITY (means: negative 96.1%, positive 98.7%, $\beta = 1.12$, $SE = .46$, $\chi^2 = 5.32$, $p = .02$), in which accuracy was lower in the negative conditions than in the positive conditions.

However, participants were equally accurate at rejecting false statements and accepting true statements. That is, we found no effect of CONGRUENCE here (means: congruent 96.6%, incongruent 98.2%, $\beta = -.63$, $SE = .48$, $\chi^2 = 1.64$, $p > .1$): a statement's veracity with respect to its accompanying picture made little difference to response accuracy.

No interaction was found between POLARITY and CONGRUENCE ($\beta = .28$, $SE = .95$, $\chi^2 = .09$, $p > .1$); accuracy was lower in the negative congruent conditions than in the positive congruent conditions (means: negative 94.8%, positive 98.4%). The same was observed in the negative and positive incongruent conditions (means: negative 97.4%, positive 99.0%).

4.5.2 Non-linguistic conditions (Math)

The apparent difference in POLARITY seen in Figure 7 did not reach statistical significance here, unlike in the Polish sub-experiment. This was revealed in the lack of a main effect of POLARITY (means: negative 95.8%, positive 97.7%, $\beta = .56$, $SE = .38$, $\chi^2 = 2.25$, $p > .1$): there was little difference in accuracy between the negative and positive conditions.

Similarly, participants were no more or less accurate at rejecting false statements than at accepting false statements. That is, there was no main effect of CONGRUENCE on accuracy (means: congruent 95.6%, incongruent 97.9%, $\beta = -.68$, $SE = .41$, $\chi^2 = 2.65$, $p > .1$). Whether the statement matched the picture made little difference to average response accuracy.

Finally, these two factors did not interact. No interaction was found between POLARITY and CONGRUENCE ($\beta = .80$, $SE = .88$, $\chi^2 = .83$, $p > .1$); accuracy was lower in the negative congruent condition than in the positive congruent condition (means: negative 93.8%, positive 97.4%), which was also observed in the negative and positive incongruent conditions (means: negative 97.7%, positive 98.0%).

4.6 Discussion

In Experiment 2, we found that negative Polish statements took longer to process than their positive counterparts, and rejecting a false statement took longer than accepting a true statement, regardless of whether the statement was provided in Polish or in a quasi-algebraic inequality. These results replicate the major results of Experiment 1, and support the viability of a decompositional analysis of negative comparatives in languages like Polish.

Unlike in the previous natural language sub-experiment, we found no interaction between POLARITY and CONGRUENCE in the Polish data. The effects of these factors appeared to be independent and additive, as predicted by the Sentence-First model of Clark and Chase (1972). In Experiment 2, participants were able to view the picture for up to 4 seconds, unlike the 200ms time window imposed in Experiment 1, and moreover participants were encouraged to focus on accuracy over speed. These design parameters were more consistent with the original testing conditions for the Sentence-First model, so this result is perhaps not surprising. Future research should investigate why imposing a shorter viewing window leads to a different pattern of behavior.

The results tended in the same direction in the Math sub-experiment of Experiment 2. We found main effects of both POLARITY and CONGRUENCE, with no interaction between these factors. These results provide further evidence against the simple hypothesis concerning how Math statements might be processed.

In terms of accuracy, we found a main effect of POLARITY, but no effect of CONGRUENCE, and no interaction between these factors—both for Polish and Math. With respect to negation, accuracy appeared to pattern inversely with response latency, in line with Clark and Chase (1972) and Trabasso et al. (1971). This finding differed from Experiment 1 for English, which did not show this pattern; again, this lack of difference appears to be due to the differing amounts of time participants had to view the picture.

The results of these two experiments are thus consistent with the model of the Sentence-to-Picture verification task of Clark and Chase (1972). Perhaps surprisingly, they are met both in a task using natural language (Polish) and putatively non-linguistic statements (Math). Unlike Deschamps et al's (2015) finding, it appears that putatively non-linguistic stimuli can be processed in a highly similar fashion to linguistic stimuli, perhaps suggesting some sort of translation at test.

5 General discussion

We have considered the processing of positive and negative adjectival comparatives in English and Polish, in contrast to analogous quasi-mathematical statements. We drew an explicit link between the decompositional analysis proposed by Büring (2007) and the additive effects on response latencies in simple tasks presented primarily in Clark and Chase (1972). Our results are predicted by decompositional analyses of *shorter* versus *taller*, given the linking hypotheses we have assumed. If the posited decomposed forms are reflective of speakers' semantic representations of positive and negative comparatives, and if comparing those representations to visual inputs involves additional symbolic manipulation whenever the representations mismatch, the predictions (and our results) follow straightforwardly.

One of the specific costs of processing *shorter/niższe* comparatives over *taller/wyższe* comparatives can be seen to reflect the cost of computing (shown here for A, but equally well for B) $\textsc{little}(\textsc{tall}(A))$ over $\textsc{tall}(A)$. However, this cost is reflected at the point of making the judgment, *after* participants view the scene. How can we conceptualize this pattern? If we assume that scenes are essentially represented in positive terms (contra Clark and Chase 1972), we can suppose that the cost reflects one of comparing *representations in canonical (positive) form*.

(15) a. Language: *A is shorter than B.* $\stackrel{lg}{\Rightarrow} \textsc{er}(\textsc{little}(\textsc{tall}(A)), \textsc{little}(\textsc{tall}(B)))$

b. Vision: (a line marked A is longer than a line marked B) $\stackrel{vis}{\Rightarrow} \textsc{er}(\textsc{tall}(A), \textsc{tall}(B))$

c. Language representation in canonical form? NO.
$\textsc{er}(\textsc{little}(\textsc{tall}(A)), \textsc{little}(\textsc{tall}(B))) \stackrel{lg}{\Rightarrow} \textsc{er}(\textsc{tall}(B), \textsc{tall}(A))$ '*Below* time'

d. Representations match? NO. '*Falsification* time'

Such a picture crucially involves the assumption that perceptual events are interpreted into a kind of representation that is 'written' in the same format as the semantic representations of sentences (Clark et al. 1973, Carpenter 1974). If semantic representations have a propositional format, then it must be possible to interpret the things we see into a similar format to feed later comparison operations. It has been suggested that the manipulation of such symbolic representations is reflected in the time it takes to initiate a response given visual input (Just and Carpenter 1971), as well as by the pattern and duration of eye fixations during tasks involving visual input (Just 1974). This paper has supported this pattern for temporal duration; the question of eye movements will be of particular interest for future research.

The results of our Experiment 1 for English raised some questions about this model. There, we saw that the pattern of response latencies for English were different from those for Math, and different again from the predictions of the "Sentence-First" model. In Section 3.6, we speculated that these results reflected an additional cost imposed under the time pressure, specifically involving checking whether the entities in the two positions of the linguistically- and visually-derived ER relations were the same. Incorporating such a cost into the processing model helped to explain the pattern we observed for English in that experiment, but it does not explain why it was not observed for Math under the same conditions. We leave the development of these ideas for future research.

We also observed that the predictions of the 'simple' hypothesis for how math statements would be processed was not borne out. In two experiments with different task demands, and very different populations of speakers, we consistently observed patterns of response latency and accuracy that matched the predictions of the Sentence-First model for matching natural language and pictures. A possible hypothesis about why this pattern was observed, briefly entertained in Section 3.6, is that participants may have been 'translating' the math statements into their natural language during the initial processing of the statement. If so, we

might expect that participants would spend more time on the statement screen for math statements than for sentences. To explore this idea, we conducted a post-hoc independent samples t-test on the length of time participants spent reading the statement before pressing spacebar to advance to the picture in Experiment 1.[11] This comparison was not significant ($t(1199) = .075, p = .9$). Thus, if math statements required an up-front additional cost for translation, it was not reflected in reading times.

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 we 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 the language-cognition interface. On the decompositional approach, the mapping from syntax to conceptualization is uniform, and the same, for English, Polish, and Hixkaryana (see Section 2): explicit constituents of the syntactic representation are related to explicit operations in processing. On the alternative view, this mapping is transparent in Hixkaryana, but requires a detour through the lexicon in English and Polish.

It could thus be particularly fertile to investigate links between processing and linguistic typology. It is well-known that negation is 'special' in language, and one vexing question that has yet to be resolved is *why* it is so special. We have begun to suggest a view on which negative forms are 'non-canonical', while those of other cognitive systems may be 'canonical'—at least in the mental language in which these representations are compared with those derived from other information sources. Comparing or interfacing representations across domains may require transformations of non-canonical representations into canonical ones, which is costly. If linguistic forms are furthermore required to be 'transparent' to non-linguistic cognition, then it seems that negative elements will be very costly indeed.

Two areas stand out as ripe for further investigation along these lines.

For one, Horn (1972) discusses the fact that of the four corners of the Aristotelian Square of Opposition, only three are found as distinct lexical items across languages: an existential (*some*), a universal (*all*), and a negative existential (*none*) (cf. Roelandt 2016). The fourth member of the set—a quantificational determiner equivalent in meaning to the phrasal form *not all* in English—-never appears as a lexical item. Beyond the quantificational determiners, Bobaljik (2012) notes that no language, in the over 300 languages that he surveyed, features a synthetic comparative of inferiority. This gap is exemplified in the following examples, with the unattested meanings incorporating elements of the Büring and Heim semantics for *less*.

(16) a. Mary is *smart-er* than Bill. $\Rightarrow$ $\textsc{er}(\textsc{smart}(M), \textsc{smart}(B))$ ATTESTED

b. * Bill is *smart-le* than Mary. $\Rightarrow$ $\textsc{er}(\textsc{little}(\textsc{smart}(B)), \textsc{little}(\textsc{smart}(M)))$ UNATTESTED

Why should these typological gaps exist? Perhaps they reflect constraints on 'how much meaning' can be bundled into a single morpheme (Dunbar and Wellwood 2016), or on 'how much non-transparency' is permitted at the language-cognition interface. More broadly, the requirement for a transparent interface might require that the smallest meaningful pieces are alignable in a regular way with representations and processes in non-linguistic cognition. Thus, evidence from processing could provide new insight into what those pieces are, by looking at the kinds, and amount, of information recruited during linguistic understanding.

Acknowledgements We are grateful to Rebecca Way for creating the diagrams illustrating the Kennedy semantics from section 2, and assisting with programming Experiment 1. We also extend our gratitude to Joanna Błaszczak, Andreas Brocher, Johannes Gerwien, Naomi Kamoen, and Maria Mos for their involvement with coding Experiment 2. Finally, we extend a

[11] These data were not collected for Experiment 2 due to a programming error.

special acknowledgment to Yaman Özakın for his work on creating the picture stimuli used in both experiments. The work on Experiment 2 was supported by the Polish National Science Center (NCN) grant OPUS 5 HS2 (DEC-2013/09/B/HS2/02763).

References

Allopenna, P. D., Magnuson, J. S., and Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, 38:419–439.

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.

Bates, D., Maechler, M., Bolker, B. M., and Walker, S. (2014). lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-7.

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

Carpenter, P. A. (1974). On the comprehension, storage and retrieval of comparative sentences. *Journal of Verbal Learning and Verbal Behavior*, 13(4):401–411.

Carpenter, R. H. S. (1977). *Movements of the eyes*. Pion Ltd, London, UK.

Clark, H. H. (1969a). The influence of language in solving three term series problems. *Journal of Experimental Psychology*, 82:205–215.

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

Clark, H. H. (1970). How we understand negation. Paper presented at COBRE Workshop on Cognitive Organization and Psychological Processes. 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.

de Leeuw, J. R. (2015). jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior Research Methods*, 47(1):1–12.

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

Dunbar, E. and Wellwood, A. (2016). Addressing the 'two interface' problem: The case of comparatives and superlatives. *Glossa: a journal of general linguistics*, 1(1):5.

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

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.

Horn, L. (1972). *On the semantic properties of the logical operators in English*. Indiana University Linguistics Club, Bloomington, IN.

Huttenlocher, J. (1969). Imaginal processes in reasoning. Paper presented at the XIX International Congress of Psychology. London, UK.

Just, M. A. (1974). Comprehending quantified sentences: The relation between sentence-picture and semantic memory verification. *Cognitive Psychology*, 6(2):216–236.

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

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

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.

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

Pancheva, R. (2006). Phrasal and clausal comparatives in slavic. In *Formal approaches to Slavic linguistics*, volume 14, pages 236–257. Citeseer.

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

Roelandt, K. (2016). *Most or the art of compositionality: Dutch de/het meeste at the syntax-semantics interface*. PhD thesis, Utrecht University.

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.

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

Sternberg, S. (1969). The discovery of processing stages: Extensions of donders' method. *Attention and performance II. Acta Psychologica*, 30:387–404.

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

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

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