Extremely costly intensifiers are stronger than quite costly ones

Erin D. Bennett (erindb@stanford.edu), **Noah D. Goodman** (ngoodman@stanford.edu)
Department of Psychology, Stanford University.

Abstract

We show that the wide range in strengths of intensifying degree adverbs (e.g. *very* and *extremely*) can be partly explained by pragmatic inference based on differing cost, rather than differing semantics. The pragmatic theory predicts a linear relationship between the meaning of intensifiers and their length and log-frequency. We first test this prediction in three studies, using two different dependent measures, finding that higher utterance cost (i.e. higher word length or surprisal) does predict stronger meanings. In two additional studies we confirm that the relationship between length and meaning is present even for novel words. We discuss the implications for adverbial meaning and the more general question of how extensive non-arbitrary form-meaning association may be in language.

Keywords: intensifiers; degree adverbs; scalar adjectives; pragmatics; m-implicature

Introduction

How do different words get their meanings? For instance, why is an "extremely good paper" better than a "quite good paper"? The traditional answer (De Saussure, 1916) is that different meanings have been arbitrarily and conventionally assigned to the different word forms. This view has been challenged by a number of examples in which word meaning appears to be non-arbitrarily related to properties of the word. In some cases, the phonetic form of a word is systematically related to its meaning, for example rounded vowels and voiced consonants tend to refer to round objects (Khler, 1970; Ramachandran & Hubbard, 2001; Holland & Wertheimer, 1964; Davis, 1961). In other cases, orthographic form is diagnostic of meaning, for example, speakers of Hebrew who have never seen Chinese characters are nonetheless above chance at matching them to their corresponding Hebrew words (Koriat & Levy, 1979). Similarly, the length of words predicts aspects of their meanings: across languages longer words refer to more complex meanings (Lewis, 2016). Open questions remain about the systematic factors that can influence meaning and the source of these effects.

In this paper, we explore adjectival intensifiers[1], like *extremely* and *quite*, as a case study in which to empirically explore the relationship of meaning to factors like word form and distribution of usage. Intensifiers form a good case study because they are amenable to simple quantitative measures of meaning: Many adjectives correspond to concrete numeric scales, and intensifier's strength can be measured as the numeric extent to which it shifts the interpretation of such a scalar adjective. Intesifiers are of interest because theoretical considerations, which we lay out below, suggest a relationship between intensifier meaning and their communicative cost (i.e. frequency and length). This account of intensifier meaning adds to a growing body of literature exploring how principles of recursive, rational communication shape language interpretation (e.g. Grice, 1975; Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; Franke, 2011; Russell, 2012; Kao, Wu, Bergen, & Goodman, 2014; Bergen, Levy, & Goodman, 2014).

In the next section, we discuss a minimal semantics for intensifiers, building off of previous work on scalar adjectives. We show how pragmatic effects predict systematic variation in the meanings of intensifiers: the meanings of intensifiers are expected to be influenced by their form (in length) and their distribution (frequency) of usage. We formalize this semantics in our Appendix, and derive the prediction that the interpreted strength of an intensified phrase should be linearly related to communicative cost (i.e. length and frequency) of that phrase. The impact of word length is reminiscent of the results of Lewis (2016), who studied noun categories. While word frequency is known to have major effects on sentence processing (e.g. Levy, 2008), the prediction that frequency should affect meaning is more surprising.

We confirm, in our first series of studies (Studies 1a, 1b, and 2), that English intensifiers in adjective phrases are indeed interpreted as much stronger for less frequent intensifiers. This holds in quantitative judgments of meaning and in forced comparisons, and across a number of adjectival dimensions. With the more sensitive dependent measure of Study 2, we also find an additional effect of length above and beyond surprisal. In our second set of studies (Studies 3 and 4), we replicate this finding, and extend it to novel intensifiers, showing that length is a significant predictor of the strength of an intensifier's meaning even in the absence of any conventional meaning. We conclude with a discussion of different interpretations of these phenomena and future directions.

The semantics of intensifying degree adverbs

Our paper focuses on intensifying degree adverbs applied to scalar adjectives.[2] Scalar adjectives have been described as

[1] Intensifiers are adverbs that modify scalar adjectives so that the interpretation of the intensified adjective phrase is more extreme than the interpretation of the bare adjective phrase. The word "intensifier" is often used to denote the full range of degree adverbs, be they "amplifiers", or "downtoners" (Quirk, Greenbaum, Leech, & Svartvik, 1985). The "intensifiers" we are looking at in this paper are, according to this typology, "amplifiers" because they increase (rather than decrease) the threshold associated with a gradable predicate. This typology also distinguishes between two different kinds of amplifiers: those that increase an adjective maximally (e.g. *completely* and *utterly*) and those that merely increase (e.g. *greatly* and *terribly*). We do not make this distinction. The word "intensifier" is sometimes used for a completely different linguistic phenomenon, where a reflexive is used for emphasis, e.g. "The king himself gave the command," which we do not analyze in this paper.

[2] Some of these intensifiers can also apply to verbal and nominal predicates, and different restrictions apply for different intensifiers, e.g. *I truly like carrots* is an acceptable utterance, whereas *I very like*

having a threshold semantics (Kennedy, 2007), where, for example, *expensive* means "having a price greater than θ" and θ is a semantic variable inferred from context (e.g., $100). Above the threshold degree θ, the adjective is true of an object, and below, the adjective is false. Lassiter and Goodman (2013) build on the Rational Speech Acts (RSA) framework (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013) to give a formal, probabilistic model of how this threshold might be established by pragmatic inference that takes into account statistical background knowledge (such as the distribution of prices for objects). We return to this model below and present a full model in the Appendix.

Previous researchers have proposed that adjective phrases modified by intensifiers have the same semantics as unmodified adjective phrases, except with new, higher thresholds (Kennedy & McNally, 2005; Klein, 1980; Wheeler, 1972). That is, some threshold, inferred from context, exists above which objects are *expensive* and below which they are not, and the intensifier *very* determines a new, higher threshold for the adjective phrase *very expensive*. These researchers suggest that the intensified thresholds are determined by first collecting the set of objects in the comparison class for which the bare adjective is true, and then using that as the comparison class to infer a new threshold, i.e. *very expensive laptop* means "expensive for an expensive laptop". This analysis results in the expected intensification of adjectives ("expensive for an expensive laptop" has a higher threshold for being true than simply "expensive for a laptop") and is appropriately sensitive to different domains (e.g. the absolute difference in price between thresholds for *expensive* and *very expensive* is much higher in the context of "That space station is very expensive," than in the context of "That coffee is very expensive."). However, this proposal does not distinguish between the graded strengths of different intensifiers, for example, *very expensive* and *phenomenally expensive*.

Intuition suggests that different intensifiers do have different strengths (e.g. *outrageously* seems stronger than *quite*), and we provide further evidence of this in our studies, where participants interpret and compare different intensifiers. It could be that the degree of strength of different intensifiers is conventionally specified by the lexicon. But the semantics must then specify how these entries affect the very flexible threshold of the relevant adjective. In addition, the multitude of intensifiers (Bolinger, 1972) and their apparent productivity[3] suggest a more parsimonious solution would be welcome. That is, having a lexically determined meaning for each different intensifier might overlook the similarity among words of this class. In the account that follows, we build minimally on existing models of adjective interpretation and rational communication to articulate a model of intensified adjective phrase interpretation.

carrots is not. See Bolinger (1972) for a discussion.

[3]For example, *altitidinously expensive* is not in common usage, but one can easily interpret *altitidinously* as a novel intensifier.

Intensification as an M-implicature

We explore the idea that an adjective phrase with an intensifying degree adverb derives much of its meaning from a M(arkedness)-implicature (Levinson, 2000): more marked (costly to utter) versions of an adjective phrase will be interpreted as implicating higher values (e.g. in case of the adjective *expensive*, higher prices). Given two possible utterances a speaker could say to communicate the same meaning, a speaker will usually choose the less costly utterance. If the speaker instead chooses a more costly utterance (e.g. "I got the car to start" as opposed to "I started the car"), they may be doing so in order to communicate something more distinct, intense, or unusual (e.g. "I got the car to start, but it was unusually difficult"). In other words, the marked form corresponds to the marked meaning. If scalar adjectives include a free threshold variable inferred from context, then the speaker's use of a longer, intensified adjective phrase could lead the listener to infer that the threshold for this adjective phrase is unusually extreme relative to other, less costly phrases that the speaker could have used.

To realize such an M-implicature, we suggest extending Lassiter and Goodman (2013)'s probabilistic model of scalar adjective interpretation slightly. We assume that each time a scalar adjective is used, in each phrase, it introduces a free threshold variable—a new token threshold is inferred for each access of the lexical entry of the adjective. The set of thresholds, for the actual sentence and all alternative sentences, is then established by a pragmatic inference that takes into account the differing costs of the sentences. The intensifiers themselves do not contribute to the semantics but increase the cost of the utterance, thus affecting pragmatic inferences. This model is described in detail in the Appendix. As in previous RSA models that include utterances with similar semantics but different costs (Bergen, Goodman, & Levy, 2012; Bergen et al., 2014), we find an M-implicature, such that more costly intensifiers result in stronger adjective phrases. As illustrated in the Appendix this relationship is expected to be approximately linear, resulting in a straightforward quantitative hypothesis that we evaluate against empirical data in our studies.

We view this model as an illustrative caricature of intensifier meaning: In this model intensifiers contribute *nothing* to the literal, compositional semantics. Yet, pragmatic interpretation yields a spectrum of effective meanings for the intensifiers, determined by their relative usage costs. This predicts an empirically testable systematic variation in meaning as a function of cost. It is very likely that the meaning of individual intensifiers includes idiosyncratic, conventional aspects in addition to these systematic factors. This would be expected to show up as residual variation not predicted by cost, but would not nullify the hypothesized relationship between cost and meaning. This account applies straightforwardly only to intensifying degree adverbs; "de-intensifying" adverbs that effectively lower the threshold will require further work to explain.

Factors affecting utterance cost

We have identified an intensifier's cost as a potentially critical determiner of its interpreted meaning. To connect this prediction to empirical facts, we still must specify (at least a subset of) the factors we expect to impact cost. The most natural notion of cost is the effort a speaker incurs to produce an utterance. This could include cognitive effort to access lexical items from memory, articulatory effort to produce the sound forms, and other such direct costs. Speakers might also seek to minimize comprehension cost for their listeners, resulting in other contributions to cost. For the purposes of this paper, we restrict ourselves to the most obvious contributors to production cost and use proxies that are straightforward to quantify: length (longer utterances are more costly)[4] and frequency (rarer intensifiers are harder to retrieve from memory in production and therefore more costly). In a number of different tasks, lexical frequency affects difficulty in an approximately logarithmic way. For instance, word recognition time (McCusker, 1977) and reading time in context (Smith & Levy, 2013) are both logarithmic in frequency. We thus use the log-frequency (whose negative is also called *surprisal*) as the quantitative contribution to cost.

Our model predicts a linear contribution of longer and higher surprisal intensifiers to the meaning of an adjective phrase (see the Appendix for more detail). This leaves open the relative importance of length and surprisal (as well as other factors that might enter into cost), which we explore in our studies. For interpreting the results of these studies, we use mixed linear models, since further quantitative comparison between the pragmatic account and the data would simply be overfitting.

Utterance cost predicts intensifier strength

The proposal detailed above predicts an association between measures of cost and strength of interpretations. In our first series of studies, detailed in this section, we tested whether our measures of communicative cost can in fact predict intensifier strength.

We used two measures of intensifier strength. Our first measure of intensifier strength (used in Studies 1a and 1b) was asking participants for a numeric interpretation of intensified adjective phrases. Our second measure (used in Study 2) was asking participants to rank the strength of adjective phrases that differed only in their intensifier. The first measure allowed us to compare our full set of intensifiers to one another on a numeric scale. The second allowed us to test our hypothesis on a wider range of adjectives at once, some of which (e.g. *beautiful*) correspond to more abstract, non-numeric scales.

[4] We measure length in number of syllables, although length in characters (which might be a more relevant source of utterance cost in a written format) has similar predictive power to syllable length in all of our analyses.

Study 1a

In Study 1a, we tested the qualitative prediction that as the communicative cost of an intensifier increases, so will the numeric interpretation of the adjective phrase it is part of. We tested this prediction by eliciting free-response price estimates from participants for phrases such as *very expensive watch* and determining whether the prices participants responded with were correlated with independent measures of communicative cost.

Methods

Participants 30 participants with US IP addresses were recruited through Amazons Mechanical Turk and paid $0.40 for their participation. 1 participant was excluded from the analysis for admitting that they did not think they followed the instructions in a post-experiment survey and another for not being a native speaker of English.

Items The sentences in the study included intensifiers paired with the adjective *expensive* (Figure 1). There were three categories of objects (*laptop*, *watch*, and *coffee maker*) and 40 intensifiers (see Table 1).

The intensifiers in our study were collected from word lists online and searching thesauri for more intensifiers. We chose intensifiers that have a wide range of frequencies and excluded intensifiers that are either more commonly used to signal affect than to signal degree (e.g. *depressingly expensive* might indicate a degree, but it mainly indicates affect) or are ambiguous between other parts of speech (e.g. *super* can be used as an intensifier, as in "super expensive", but it can also be used as an exclamation, as in "Super!").

We chose the domain of price for Study 1a and used only the adjective *expensive*. Because price constitutes a quantitative scale with standard units (dollars for our US participants), this allowed us to quantitatively measure the relative strengths of different intensifiers.

Procedure[5] We asked participants to estimate the prices of different objects based on different descriptions of those objects.

Each participant gave price judgments for every intensifier-category pairing in a randomized order (different for different participants), for a total of 120 price judgments per participant.

The only allowable characters in responses were the digits 0-9 and (optionally) one decimal point (.) followed by two digits. All other responses were immediately rejected. Participants were prevented from continuing until they provided a valid numeric response for each trial.

[5] A demo of Study 1a can be found at `http://cocolab.stanford.edu/links_for_papers/bennett2017extremely/experiments/Study1a/`.

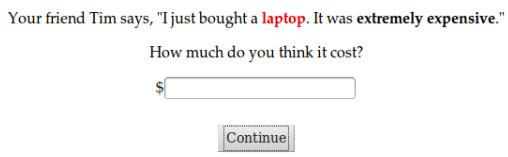

Figure 1: Screenshot from Study 1a target question.

Corpus Methods Table 1 shows word frequency and length in syllables for the intensifiers used in Study 1a. The frequencies were collected from the Google Web 1T 5-grams database (Brants & Franz, 2006).[6] In the analysis below we use word length and word surprisal (negative log-frequency) as proxies for a word's cost, as motivated above. The syllable lengths of our intensifiers and the surprisals were correlated ($r = 0.26$). This correlation makes it somewhat difficult to determine the effect of one measure of cost above and beyond the other. In our first series of studies, we focus on the primary effect of surprisal, since we have more range in surprisal across intensifiers than in length and since we manipulate length in syllables for our next series of studies. However, we model both measures of communicative cost in our analyses: We include both predictors in regressions and report likelihood-ratio tests between the full model and simpler models.

Analysis Prices that participants give obviously vary systematically with the object (expensive laptops tend to be more expensive than expensive coffee makers). Responses are also likely sensitive to variation across participants due to their different beliefs about likely prices. Because we have few objects, we are unable to model the variation due to object, but because we have many intensifiers (they are fully crossed with objects), normalizing is fairly effective at converting all objects to the same scale. We are not theoretically interested in the variation due to objects, and so in order to compare intensifiers across these objects, we first normalized log-transformed prices within participant and object. We used the logarithm of participants' price estimates because of evidence that people's representation of numbers, including prices, is logarithmic (Fechner, 1860).[7]

We ran a linear mixed effects regression to predict scaled price estimates. We included centered fixed effects of length and surprisal. To model random effects due to participant, we only included random slopes, since the normalization gives each participant an intercept of 0. We also included a random intercept for intensifier to model any idiosyncratic meaning there might be to a particular intensifier beyond communicative cost. The predicted mean y_{ij} for the i^{th} intensifier and the j^{th} participant under this model is shown in Equation 1, where f_i represents the surprisal of the i^{th} intensifier and l_i represents its length.

$$y_{ij} = \beta_0 + U_{0i} + (U_{1j} + \beta_1) f_i + (U_{2j} + \beta_2) l_i \qquad (1)$$

As previously noted, surprisal and length in syllables are correlated in this set of intensifiers ($r = 0.26$). We focus on surprisal as our primary effect, but we are also interested in the independent contribution of length in syllables. To address the independent effects of these variables, we ran model comparisons using log likelihood, leaving out information about one of the predictors. We also ran two additional mixed effects reggressions: one in which the suprisal is first residualized against syllables (using ordinary linear regression), and one where length is residualized against surprisal.

Results If the meaning of an intensifier is stronger for higher cost intensifiers, we would expect to find that as surprisal increases and length in syllables increases, the prices participants give will also increase. We find that this is the

Table 1: Intensifiers from Study 1a, number of occurences in Google Web 1T 5grams corpus, and length in syllables.

ngram	frequency	length
surpassingly	11156	4
colossally	11167	4
terrifically	62292	4
frightfully	65389	3
astoundingly	73041	4
phenomenally	120769	5
uncommonly	135747	4
outrageously	240010	4
fantastically	250989	4
mightily	252135	3
supremely	296134	3
insanely	359644	3
strikingly	480417	3
acutely	493931	3
awfully	651519	3
decidedly	817806	4
excessively	877280	4
extraordinarily	900456	6
exceedingly	977435	4
intensely	1084765	3
markedly	1213704	3
amazingly	1384225	4
radically	1414254	3
unusually	1583939	4
remarkably	1902493	4
terribly	1906059	3
exceptionally	2054231	5
desperately	2139968	3
utterly	2507480	3
notably	3141835	3
incredibly	4416030	4
seriously	12570333	4
truly	19778608	2
significantly	19939125	5
totally	20950052	3
extremely	21862963	3
particularly	41066217	5
quite	55269390	1
especially	55397873	4
very	292897993	2

[6] We also ran the same analyses on frequency information collected from the Google Books American Ngrams Corpus (Michel et al., 2011) and found similar results.

[7] I.e. the perceptual distance between two prices the same dollar amount apart is more for small numbers (e.g. $3 and $6) and less for large numbers (e.g. $1,543 and $1,546).

case for surprisal, but do not show a significant effect of syllable length beyond the effect of surprisal. Our results are shown in Figure 2, in a way that highlights the surprisal predictor. We present full regression output in Appendix B.

We find a significant effect of surprisal ($b = 0.106, t(38.9) = 3.411, p = 0.0015$) such that less frequent words tend to be associated with higher price estimates. In this regression, we did not a significant effect of syllable length ($p = 0.0936$), above and beyond surprisal.

Because surprisal and syllable-length are correlated, in addition to the analysis reported here we also used likelihood ratio tests to compare the full model to models with only one of the two predictors. These tests show that length in syllables alone account for the data less well than the full model ($\chi^2(8) = 17.055, p < 0.0005$) and that surprisal alone also accounts for the data less well than the full model ($\chi^2(8) = 13.697, p < 0.0005$), suggesting that both might separately contribute to the cost of the utterance. When length was residualized with respect to suprisal, both the residuals ($b = 0.106, t(38.9) = 3.411, p = 0.002$) and length ($b = 0.202, t(40.9) = 2.693, p = 0.010$) were significant predictors of scaled ratings. However, when surprisal was residualized with respect to length, surprisal ($b = 0.12, t(39.9) = 3.985, p < 0.0005$) but not the residuals ($p = 0.0936$) was a significant predictor of scaled ratings. This discrepency is likely a due to the fact that, since length in syllables is a discrete value taking one of 6 values, whereas surprisal is continuous, length can be more informatively predicted from surprisal (different surprisals can map onto approximately the same length) than surprisal can be predicted from length (the same length cannot map onto different surprisals).

Overall, we confirmed our main prediction that intensifiers that are less frequent (and therefore are more costly to communicate) also tend to be interpreted as having stronger meanings, at least when used to modify the adjective *expensive*. We found inconclusive results for our secondary prediction that length (another factor in communicative cost) has any effect beyond that of surprisal. This ambiguity motivated a more sensitive dependent measure in Studies 2 and 4.

Study 1b

Our initial selection of intensifiers was somewhat haphazard, being chosen partly to give the best (intuitive) chance of observing an effect. To show that there was no implicit bias induced by this selection, we next replicated using a revised, and more systematic, set of intensifiers. We additionally took this opportunity to increase the sample size, guided by a power analysis based on the results of Study 1a.

Methods

Participants We wanted enough participants for a power level of 0.8 for the principal effect of surprisal. Power analyses for mixed effects models with continuous predictors are not analytically straightforward, and so we approximated the number of participants necessary for our desired power by

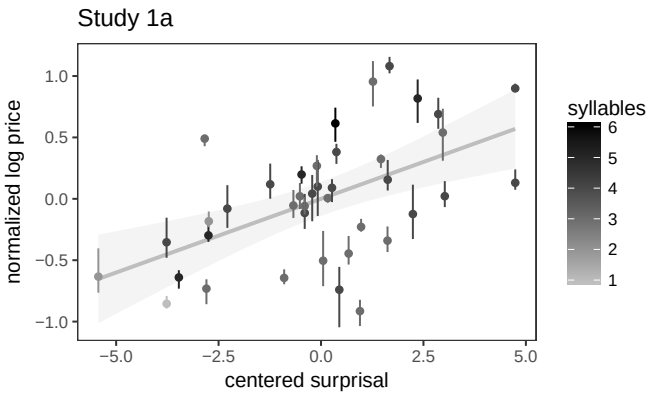

Figure 2: Results of Study 1a. As surprisal increases, participants' scaled free response prices increase.

bootstrapping. We used the data from our original Study 1a to simulate alternative possible datasets with different numbers of intensifiers and participants. For each of 100 iterations, we sampled with replacement from Study 1a a set of P participants and a set of I intensifiers, varying P and I. We created a resampled dataset where we combined the resampled participants with the resampled intensifiers and collected the corresponding responses for each pair. For each resampled dataset, we ran the regression model from Study 1a. We computed the proportion of runs in which the surprisal term was significant and interpreted this as the power of such a study.

We found that our original Study 1a was somewhat underpowered by this metric (statistical power was near 0.70 for the surprisal regressor, our principal effect of interest). For the replication, we determined that with the 71 intensifiers we collected from grammars of English (described below), 50 was the minimum number of subjects for power at the 0.8 level for the surprisal term. We doubled this amount for a goal of 100 participants for the replication. (Despite having sufficient power for the surprisal term, our bootstrapped power analysis suggested that even with 100 participants, the power for the length term in the model is only 0.29. We address this instead, with a direct manipulation of length, in experiments 4 and 5.) Due to a slight error in data collection, we actually recruited 108 participants.

Following this power analysis, 108 participants with US IP addresses were recruited through Amazons Mechanical Turk and paid $1.00 for their participation. 1 participant was excluded from the analysis for admitting that they did not think they followed the instructions in a post-experiment survey.

Items For Study 1b, we again used *expensive* as the adjective for the intensifiers to modify, and we again collected responses as prices in dollars. We used the same objects (*laptop*, *watch*, and *coffee maker*) as in Study 1a.

As we explain above, in Study 1a and the remaining studies detailed in this paper, our choice of the set of intensifiers to include was somewhat arbitrary and collected after formation

of our hypothesis. To address the concern that this might bias our results, we followed a more systematic procedure for generating a list of intensifiers for our replication Study 1b. We sought out English grammars that listed intensifiers. Since no single source contained the number of intensifiers desired for sufficient power in Study 1b, our process for collecting intensifiers in this replication was to combine word lists from multiple grammars of English. We first found 12 grammars of English that mentioned one of the following terms: "intensifiers", "adverbs of degree", or "amplifiers" (Aarts, Chalker, & Weiner, 2014; Douglas & Broussard, 2000; Declerck, 1991; Garner, 2016; Givn, 1993; Greenbaum, 1996; Huddleston & Pullum, 2002; Huddleston, 1984; Nelson, 2010; Quirk, 1972; Quirk & others, 1990; Van Gelderen, 2010). Most of these grammars mentioned examples of such words, and many contained lists of them. The average length of such a list or collection was approximately 21 words. We collected an aggregate list of all words that occurred in an "intensifiers", "adverbs of degree" or "amplifiers" list in at least one grammar. Some "downtoners"/"diminishers" were mixed into some of these lists (e.g. *slightly*, *barely*). Some other intensifiers cannot occur felicitously as simple pre-modifiers (e.g. *a lot*, *indeed*) and therefore would not fit in the same syntactic frame as the other intensifiers. Other intensifiers were marked as occurring exclusively in British English (e.g. *bloody*, *jolly*), and since our participants were restricted to US IP addresses, we did not include those words in our study. In addition, some lists included comparative degree adverbs like *more*. We excluded downtowners, intensifiers that do not pre-modify, comparatives, and exclusively British English intensifiers. This resulted in a total of 71 unique intensifiers. Of these, only 19 had been in Study 1a. 21 words that appeared in our previous experiment did not appear in a list in any of the English grammars, including *insanely*, *wildly*, *exceptionally*, and *frightfully*. Surprisal and length in syllables were even more correlated in this new set of intensifiers ($r = 0.63$). The full list of intensifiers included in Study 1b is in Table 2.

Because replicating Study 1a perfectly with this new set of intensifiers would require each participant to answer 213 very similar questions and would likely take at least 30 minutes (the higher end of task lengths on Amazons Mechanical Turk), we opted for a "replication design" for our replication Study 1b (following Judd, Westfall, & Kenny, 2017). We randomly spit the full set of intensifiers into 3 replication sets and varied the set of intensifiers between participants. Within each replication set, the subset of intensifiers was fully crossed with objects. Our final simulations for the bootstrapped power analysis detailed above included this replication design in computing power.

Procedure[8] The procedure for Study 1b was identical to that of Study 1a.

[8]A demo of Study 1b can be found at `http://cocolab.stanford.edu/links_for_papers/bennett2017extremely/experiments/Study1b/`

Analysis As in Study 1a, we normalized log-transformed prices within participant and object and then ran a linear mixed effects regression with centered fixed effects of length and surprisal, a random slope for participant, and a random intercept for intensifier.

Results Our results are shown in Figure 3, with full output in Appendix C. As in Study 1a, we found a significant main effect of surprisal ($b = 0.107, t(70.1) = 3.255, p = 0.0017$) such that less frequent words tend to be associated with higher price estimates. We again did not find a significant main effect of syllable length ($p = 0.3106$), above and beyond surprisal.

In likelihood ratio tests, we found that length alone accounts for the data less well than the full model ($\chi^2(8) = 30.318, p < 0.0005$) and that surprisal alone also accounts for the data less well than the full model ($\chi^2(8) = 11.571, p = 0.009$). When residualizing length with respect to surprisal, we found a significant effect of surprisal ($b = 0.128, t(74.9) = 4.96, p < 0.0005$) but no significant effect of length residuals ($p = 0.3106$). When residualizing surprisal with respect to length, we again found significant effects of both length ($b = 0.166, t(74.1) = 3.93, p = 0.0002$) and surprisal residuals ($b = 0.107, t(70.1) = 3.255, p = 0.0017$).

This replicates the results of Study 1a, and extends them to a larger, more systematic set of intensifiers.

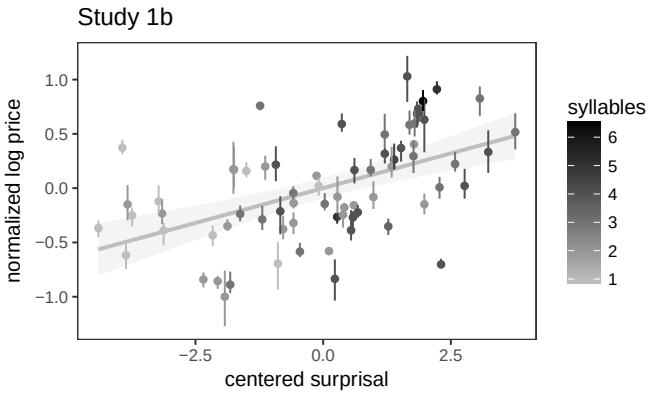

Figure 3: Results of Study 1b. As surprisal increases, participants' scaled free response prices increase.

Study 2

For Study 2 we used a forced-ranking dependent measure, which provides a more sensitive measure of the differences in degrees between similar adjective phrases. This sensitivity makes Study 2 a better test of the independent effects of length and surprisal. This depenent measure, because it does not depend on a numeric scale, also allows us to consider additional adjectival scales.

The M-implicature account described above implies that there is no semantic interaction between the intensifier and the adjective it is applied to. Instead an intensifier should

contribute similar cost, and therefore meaning, to the different adjective phrases in which it occurs[9]. To explore this, we extend our results to four additional adjectival scales.

Methods

Participants 30 participants with US IP addresses were recruited through Amazon's Mechanical Turk and paid $0.40 for participation. 2 participants were excluded from the analysis for admitting that they did not think they followed the instructions in a post-experiment survey.

Items The full set of intensifiers in Study 2 is identical to that of Study 1a. For Study 2, we used a ranking dependent measure: asking participants to sort a set of adjective phrases according to strength. Because arranging these phrases required participants to be aware of the full set of adjective phrases and access all of them on the same computer screen (which might vary in size for different participants), not all of our 40 intensifiers could effectively be presented at once. To make the task easier for participants and to extend to more adjective scales, we divided the full set of intensifiers into 4 smaller sets, maintaining a range of syllable lengths and surprisal across the intensifiers in each smaller set.

With the ranking dependent measure, we no longer need the adjectives to correspond to a standard numeric scale. We therefore included 4 adjectives: *old*, *expensive*, *beautiful*, and *tall*.

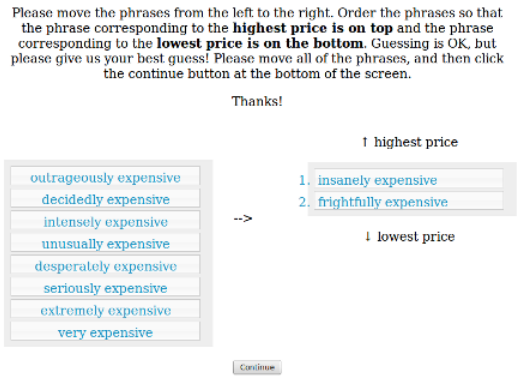

Figure 4: Screenshot from Study 2 target question.

Procedure[10] For each trial of Study 2, we asked participants to order a set of 10 adjective phrases according to strength of meaning. Adjective phrases were randomly ordered in a list on the left. Participants were asked to move the adjective phrases from the left to the right side of the screen, reordering the phrases from the "lowest" to the "highest" degree (Figure 4). Each adjective phrase in a trial contained the same adjective, modified with different intensifiers. Each participant saw 4 intensifier sets, each paired with one of the four adjectives. The pairings of intensifier-sets to adjectives varied between participants. Participants completed 4 such trials, ranking intensifiers for all 4 adjectives and all 4 intensifier-sets.

Analysis We ran a rank-ordered logit model (Beggs, Cardell, & Hausman, 1981; Hausman & Ruud, 1987) with alternative-specific variables for surprisal and for length in syllables. This model did not include an intercept. To test whether we find an interaction between intensifier strength and the adjective it modifies, we ran a second rank-ordered logit model with additional alternative-specific variables for the interaction between the adjective being modified and both surprisal and length.

Results Our results for Study 2 are shown in Figure 5, with full output in Appendix D. In our first model, with surprisal and length as the only predictors in question, we again found a significant effect of surprisal ($b = 0.164, t = 10.128, p < 0.0005$). We also find a significant effect of syllable length ($b = 0.243, t = 5.888, p < 0.0005$).

In likelihood ratio tests, we again found that length alone accounts for the data less well than the full model ($\chi^2(1) = 108.085, p < 0.0005$) and that surprisal alone also accounts for the data less well than the full model ($\chi^2(1) = 34.694, p < 0.0005$). When residualizing length with respect to surprisal, we again found a significant effect of surprisal ($b = 0.19, t = 11.663, p < 0.0005$). In addition, with this more sensitive ranking measure, we found a significant effect of length residuals ($b = 0.243, t = 5.888, p < 0.0005$). When residualizing surprisal with respect to length, we found significant effects of length ($b = 0.35, t = 8.459, p < 0.0005$) and surprisal residuals ($b = 0.164, t = 10.128, p < 0.0005$).

In the second model, with adjective interactions, we again found significant effects of surprisal ($b = 0.074, t = 2.482, p = 0.0131$) and length ($b = 0.318, t = 3.658, p = 0.0003$), and also found a significant interaction between surprisal and the adjective being modified (estimates for *tall* and *expensive* were higher than for *beautiful*, while *old* was not significantly different from *beautiful*). This suggests that context-specific surprisal (e.g. bigram frequencies) might affect the utterance cost, or that factors of utterance cost might have different effects for different adjectives.

In other words, we again found that participants assign stronger interpretations to intensifiers with higher surprisals and/or higher syllable lengths, extending now across four different adjectival scales. In addition, we found interactions between modifier and adjective (or perhaps scale).

[9] If the bigram frequency of the modified adjective phrase (*very expensive*) deviated from that expected based on independent word frequencies our frequency-based cost account would predict an interactive effect on meaning. This would likely be a relatively small effect, and the relevant bigrams were too sparse in our corpora to pursue.

[10] A demo of Study 2 can be found at `http://cocolab.stanford.edu/links_for_papers/bennett2017extremely/experiments/Study2/`.

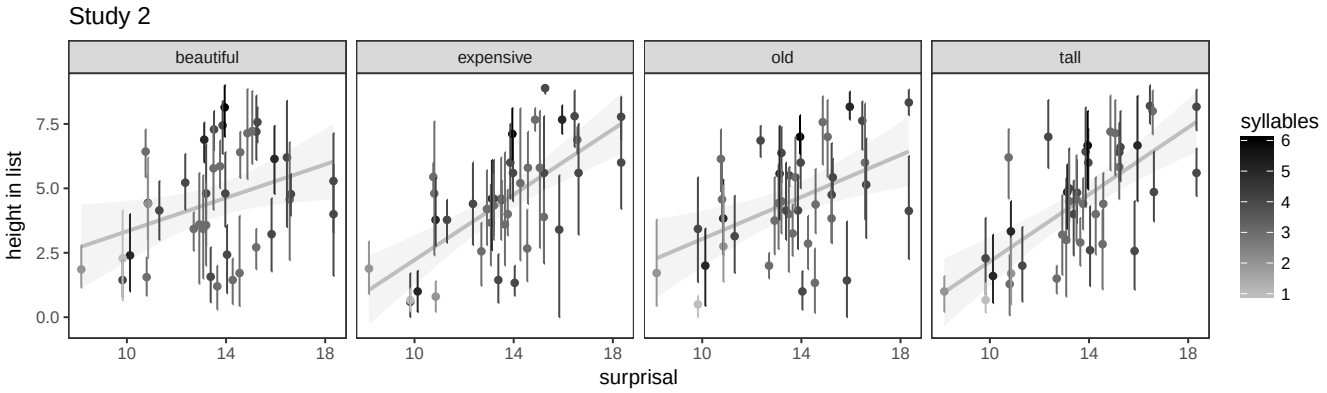

Figure 5: Results of Study 2. As surprisal and length in syllables increase, participants' rankings increased.

Discussion

In Studies 1a and 1b, we found evidence that an intensifier's surprisal is a significant predictor of strength of its meaning. In Study 2, we used a more sensitive dependent measure and showed that length in syllables is also a significant predictor of strength of meaning. These studies taken together provide evidence that intensifier meanings depend systematically on the length and frequency of their word forms.

Since this is a correlational study, such a relationship does not confirm that an intensifier's cost *causes* it to have a given meaning.

In particular, we can explain the correlation with rarity if the strength of an intensifier's meaning causes it to be rarely used. That is, more extreme intensifiers naturally refer to less probable events and properties in the world, and therefore might be used less frequently[11]. Syllable length in turn can depend on the frequency, simplicity, or predictability of a word (Zipf, 1935; Lewis, 2016; Mahowald, Fedorenko, Piantadosi, & Gibson, 2013), either because words that are frequently used get shortened over time (Kanwal, Smith, Culbertson, & Kirby, 2016) or perhaps because words that refer to simpler or more common concepts enter the lexicon sooner (when more short word forms remain to be assigned meanings). It is therefore possible that neither of these measures of cost have causal influence on the meanings of intensifiers within a particular communicative act.

To more directly address the question of whether utterance cost *causes* people to interpret an intensifier as stronger, we ran Studies 3 and 4, where we directly manipulated one of our measures of cost—length—in novel intensifiers which have no conventional meaning associated to them.

[11] However, note that the frequency with which things occur in the world does not map directly on to how often those things are talked about. Although it seems reasonable to suspect that word frequencies reflect the probabilities of the real-world concepts they describe, it might also be the case that improbable things are more likely to be commented on, and so to a certain extent the frequencies of words that describe rare concepts will be inflated.

Novel intensifier length increases strength of interpretation

Although the meanings of our existing English intensifiers could have influenced their lengths and frequencies over time, novel intensifiers have no meaning already associated with them. Therefore, if we found a relationship between the length of a novel intensifier and its interpreted meaning, we would have evidence that length can causally influence meaning. In the following two experiments, we directly manipulate the lengths of novel intensifiers and show that longer novel intensifiers are interpreted as having stronger meanings.

Study 3

In Study 3 we show that longer novel intensifiers are interpreted as having stronger meanings, using our dependent measure from Study 1a.

Method

Participants 30 participants with US IP addresses were recruited through Amazon's Mechanical Turk and paid $0.80 for their participation. 2 participants were excluded from the analysis for admitting that they did not think they followed the instructions in a post-expeirment survey and 1 for being a non-native English speaker.

Items In Study 1a, we rescaled prices within participants and objects. In order to do the same normalization for novel intensifiers, we chose 9 filler intensifiers to include in Study 3. We chose a set of filler intensifiers to cover a wide range of surprisals and syllable lengths: *colossally*, *phenomenally*, *mightily*, *extraordinarily*, *amazingly*, *terribly*, *notably*, *significantly*, *quite*. Other than covering a range of prices, the particular choice of fillers should not affect our analysis of the novel intensifiers. Each novel intensifier was presented in the same context of 9 filler intensifiers. Eliciting ratings for existing English intensifiers along with novel intensifiers allowed us to again rescale and normalize responses within

participants and items, placing all responses for the novel intensifiers on the same scale. It also allowed us to somewhat obscure our use of fabricated words, and thus decrease task demand.

We varied the novel intensifier between participants from a set of 6 novel intensifiers, three of which were relatively short (*lopusly*, *ratumly*, and *bugornly*) and three of which shared the same "root" but were two CVCV syllables longer (*fepolopusly*, *gaburatumly*, and *tupabugornly*). These items were taken from previous studies on complexity bias (Lewis, 2016) and modified by adding a final *-ly* suffix.

Procedure[12] The procedure for Study 3 was identical to that of Study 1a, except that we included only a subset of the intensifiers from Study 1a and each participant also saw one novel intensifier, randomly mixed in with the rest.

Participants again estimated prices for objects of three different categories paired with all of the intensifiers. The order of the questions was randomized between and within participants.

Analysis In Study 3, to study the effect of the length of a novel intensifier on its interpretation, we ran a linear mixed effects model on only the novel intensifiers, with length ("long"=1 or "short"=-1) as a fixed effect, a random slope for participant, and random intercepts for the three different "roots".

We included as filler a subset of the intensifiers we tested in Study 1a, and so we again replicated our findings from Study 1a. As in Study 1a, we ran a linear mixed effects regression with centered fixed effects of length and surprisal and a random slope for participant, and a random intercept for intensifier. To see the effect of the two measures of communicative cost separately, we compared the full model to a model without that measure as a regressor using a likelihood ratio test and ran versions of the model where we first residualized the measures with respect to one another.

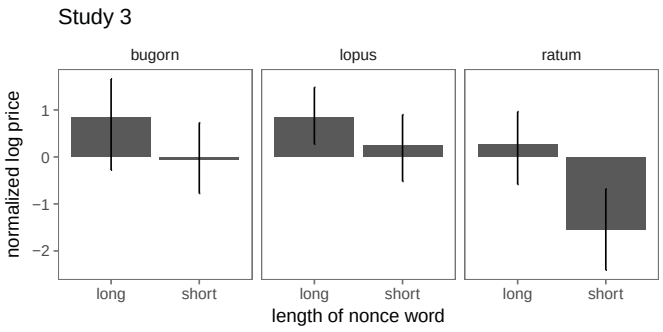

Figure 6: In Study 3, we found a significant effect of length for all novel intensifiers.

[12]A demo of Study 3 can be found at `http://cocolab.stanford.edu/links_for_papers/bennett2017extremely/experiments/Study3/`.

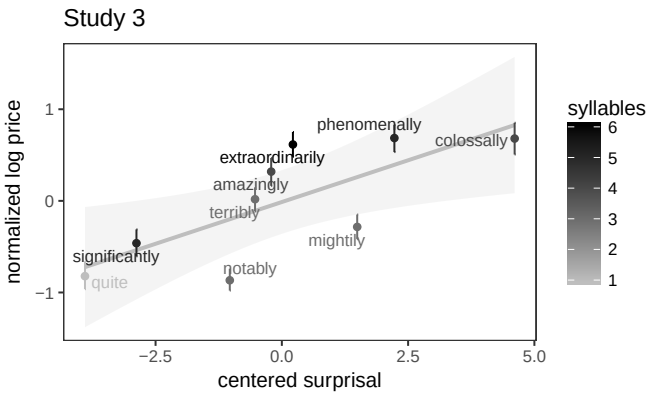

Figure 7: In Study 3, we replicated our findings from Study 1a.

Results In our analysis of novel intensifiers, we found a significant effect of length condition ($b = 0.538$, $t = 2.757$, $p = 0.0002$), indicating that people interpret a longer intensifier as stronger, even for novel intensifiers with no conventional meaning (Figure 6).

Replicating our findings from Study 1a, we found significant main effects of surprisal ($b = 0.14, t(7.2) = 2.49, p = 0.0407$) but no significant effect of syllable length ($p = 0.0846$) (Figure 7). In likelihood ratio test, we found that the full model was a better fit than the model with surprisal only ($\chi^2(8) = 8.296, p = 0.0403$) and the model with length only ($\chi^2(8) = 39.194, p < 0.0005$). Residualizing each regressor with respect to one another, we again found qualitatively similar effects to Studies 1a and 1b, with a significant effect of surprisal when residualized by length ($b = 0.14, t(7.2) = 2.49, p = 0.0407$) and no significant effect of length when residualized by surprisal ($p = 0.085$).

We observe that, numerically, average responses for *ratumly* were much lower than all the other intensifiers used in Experiment 3. In a post-hoc regression with fixed effects for the novel adverb roots (*ratum/other*=2/-1, *lopus/bugorn*=1/-1), we found a significant difference between *ratum* and the other roots ($b = -0.56, t(23) = -2.60, p = 0.0160$) and no significant difference between the remaining roots ($p = 0.759$). The very low rankings for the root *ratum* suggest possible additional effects of form that we have not captured with length in syllables alone.

Full results for Study 3 are presented in Appendix E.

Study 4

In Study 4 we again show that longer novel intensifiers are interpreted as having stronger meanings, this time for an additional adjective, using our dependent measure from Study 2.

Methods

Participants 60 participants with US IP addresses were recruited through Amazon's Mechanical Turk and paid $0.16

for their participation. 3 participants were excluded from the analysis for admitting that they did not think they followed the instructions in a post-experiment survey.

Items In Study 4, each participant saw exactly one of two adjectives (*expensive* or *tall*, varied between participants) with the set of intensifiers from Study 3. This set included 9 context/filler words and one novel intensifier, which we varied between participants.

Procedure[13] Except for the narrower set of items, noted above, Study 4 was identical to Study 2.

As in Study 2, adjective phrases for each intensifier-adjective pairing were initialized on the left in a random order.

Analysis For the novel intensifiers, we ran a cumulative logit model on the rankings (relative to the filler intensifiers) that participants gave to the novel intensifier.

As in Study 3, we also ran such a model with regressors for the word stem (or "root").

With our filler intensifiers for Study 4 we again ran a ranked order logit model (re-ranking to ignore novel intensifiers) to confirm effects of syllable length and surprisal.

Results In the cumulative logit model, we found a significant effect of length condition ($b = 0.7039, z = -2.79, p = 0.0054$).

When we included a regressor for the root of the novel intensifier, we did not find a significant difference between participants' relative rankings for *ratum* and other roots ($p = 0.110$) or between the other two roots ($p = 0.304$).

Rankings for novel intensifiers had much higher variance than rankings for attested intensifiers, since we had many fewer rankings for the novel intensifiers (which varied between participants) than for the attested ones (which every participant saw once). The novel intensifier *ratumly* was again on average ranked as less strong than any other intensifier, but the highest-ranked novel intensifiers (*gaburatumly* and *tupabugornly*) were on average ranked below the highest-ranked attested intensifiers (*colossally*, *phenomenally*, *extraordinarily*, and *amazingly*).

For the filler items, we replicated our findings from Study 2, showing significant effects of both surprisal ($b = 0.447, t = 13.879, p < 0.0005$) and syllable length ($b = 0.581, t = 10.738, p < 0.0005$) on the order in the list that participants chose for the intensifiers (see Figure 8). In likelihood ratio tests, we again found that length alone ($\chi^2(1) = 231.574, p < 0.0005$) or surprisal alone ($\chi^2(1) = 130.33, p < 0.0005$) accounted for the data less well than the full model. As in Study 2, residualized surprisal was significant ($b = 0.447, t = 13.879, p < 0.0005$) and residualized length was also significant ($b = 0.581, t = 10.738, p < 0.0005$). This replication is perhaps unsurprising, since we chose intensifiers to cover the full range of surprisal and syllable lengths.

Full results for Study 4 are presented in Appendix F.

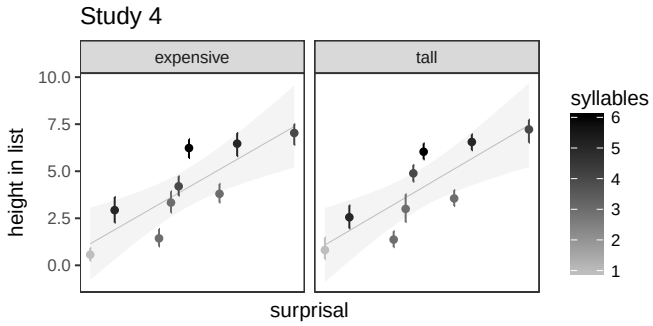

Figure 8: In Study 4, we replicated our finding from Experiment 2: longer and less frequent intensifiers are ranked higher than shorter and more frequent ones.

Discussion

Overall, in Studies 3 and 4 we found that word length in syllables is a significant predictor of interpretation strength for novel intensifiers. These novel intensifiers have no established meaning, so the relationship between their length and strength cannot be a direct consequence of the lexicon becoming more efficient over time. This result is consistent with the hypothesis that participants are inferring the meanings of the novel intensifiers pragmatically, as in the M-implicature account sketched above.

Alternatively, it could be that participants have learned a general relationship between length and meaning of intensifiers in English, and are utilizing this meta-linguistic knowledge to interpret the new words they encounter.

This meta-linguistic hypothesis is less parsimonious than the pragmatic hypothesis, since the pragmatic hypothesis relies only on mechanisms (M-implicature) that we know to be involved in other examples of language understanding (e.g. as in the "I got the car to start" example above). It builds on previous formal models of adjective meaning and informal models of intensifier meaning, generalizing to a wide range of intensifiers.

Whether due to a process of M-implicature or meta-linguistic knowledge, these results demonstrate that the relationship between word cost and meaning is not merely a static, gradual result of language evolution—interpreted meaning of intensifiers depends on length in an active, dynamic way.

General Discussion

Motivated by a recent probabilistic model of scalar adjectives (Lassiter & Goodman, 2013), we argued that adjectival intensifiers could gain aspects of their meaning through a systematic pragmatic inference, even in the absence of conventional literal meaning. Our M-implicature model (presented

[13] A demo of Study 4 can be found at `http://cocolab.stanford.edu/links_for_papers/bennett2017extremely/experiments/Study4/`.

in further detail in our Appendix) predicted a linear relationship between the intensity of an intensifier and its communicative cost, measured here in terms of length and negative log-frequency, and we explore this linear relationship using regression models.

In five experiments we provided evidence that intensifier meanings do depend systematically on the length and frequency of occurance of those word forms and that this relationship holds even for novel words.

While it is unlikely that this accounts for all intensifier meaning, our results do suggest that a major portion of meaning comes not from arbitrary, conventional association of signal to sign (De Saussure, 1916), but from features of the word's form and distribution, together with inferential processes of listeners.

For the semantics of adverbial modifiers, we have shown how pragmatic mechanisms could be central in establishing flexible contributions to sentence meaning. We have extended previous proposals that degree adverbs transform or create new threshold variables, providing a concrete mechanism for interpreting an arbitrary degree adverb in an arbitrary context.

This mechanism for linking a parsimonious semantics to interpretation via pragmatic inference follows naturally and straightforwardly from an understanding of rational social agents engaging in communication. With the minimal assumption that each use of a scalar adjective has a different threshold, we have been able to extend a model of adjective interpretation to account for a major source of intensifier meaning. This model adds to an abundance of linguistic phenomena — e.g. aspects of metaphor (Kao, Bergen, & Goodman, 2014), prosody (Bergen & Goodman, 2015), politeness (Yoon, Tessler, Goodman, & Frank, 2017), and presupposition projection (Qing, Goodman, & Lassiter, 2016) — which can be explained within the basic RSA framework, each with minimal, compatible assumptions. The precise specification of this model is laid out in more detail in the Appendix.

Our proposal and our experiments suggest that even a very minimal semantics for intensifying degree adverbs can be plausible and productive. We have implemented and described one version of such a semantics, but other versions might exist with different methods of generating the threshold variable for an adjective phrase.

We use surprisal and length in syllables as measures of communicative cost, but there may be other systematic features of intensifiers that contribute to strength. In particular, many intensifiers seem to be derived from adverbs having to do with emotion, and the valence and/or arousal of these root emotions might influence the strength of an intensifier or its affinity to co-occur with some adjective types rather than others. This might be especially true for intensifiers that are still making the change from manner adverb to intensifier (e.g. terribly once only carried the qualitative meaning of "bad and frightening", but now almost exclusively means simply "a lot").

As an illustration of the variance explained by our surprisal measure of communicative cost, we first averaged across participants and adjectives to get an average scaled response for each intensifier. We then computed the squared correlation between those values and the surprisal predictor. We find that surprisal can explain around 30% of the variance for attested intensifiers ($R^2 = 0.293$ for Study 1a, $R^2 = 0.273$ for Study 1b, and $R^2 = 0.358$ for Study 2).[14].

In investigating our proposal of how intensifiers get their meanings, we have collected many judgements about the relative strengths of intensifiers across several adjective scales and dependent measures. As a summary of these judgements, we have aggregated across the different dependent measures in our studies and displayed them in a common scale in Figure 9. For each study, we computed the average value of the dependent measure (normalized log-transformed price for Studies 1a, 1b and 3 and height in list for Studies 2 and 4) for each intensifier in that study. We then z-scored within the study, so that our different operationalizations of "strength" in the different study designs could be reasonably compared together. Given this rescaling, we can compare the strengths of these intensifiers to one another, including the novel intensifiers from Studies 3 and 4. We see some consistency across experiments and show that our choices of novel intensifiers spanned the full range of natural intensifier strengths.

For the broader question of form-meaning mapping, we have suggested a source of non-arbitrary association based on both properties of the word form and of its distribution. The effect of a word's distribution on its interpretation has potentially interesting implications for language change. If the distribution of a particular grammatical category of word (e.g. intensifiers) influences its meaning and the meaning of a word in turn influences its distribution, this would result in an unstable lexicon for this grammatical category. This suggests a mechanism by which overused words might become stale, and would predict the rapid creation of new, unusual intensifiers. This process indeed seems to be evident in the history of English (Bolinger, 1972).

We have argued that form-to-meaning mapping can come about through the inferences known to support pragmatic interpretation. Seen another way, the basic assumption that people are actively trying to communicate with each other—each reasoning about what the interlocutor means—*requires* non-arbitrary relationships between a variety of factors and effective meaning. Some systematic aspects of meaning follow directly from the principles of language understanding.

Acknowledgments

This work was supported by a James S. McDonnell Foundation Scholar Award to NDG and ONR grant N00014-13-1-0788.

[14] For intuition this means that one nat difference in surprisal predicts 0.128 normalized units of cost difference. In the case of laptops, for example, one nat of surprisal would predict approximately a 9% increase in the interpreted price.

References

Aarts, B., Chalker, S., & Weiner, E. (2014). *The Oxford dictionary of English grammar*. Oxford University Press.

Beggs, S., Cardell, S., & Hausman, J. (1981). Assessing the potential demand for electric cars. *Journal of econometrics*, *17*(1), 1–19.

Bergen, L., Goodman, N., & Levy, R. (2012). That's what she (could have) said: How alternative utterances affect language use. In *CogSci.*

Bergen, L., & Goodman, N. D. (2015). The strategic use of noise in pragmatic reasoning. *Topics in Cognitive Science*, *7*(2), 336–350.

Bergen, L., Levy, R., & Goodman, N. D. (2014). Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*.

Bolinger, D. (1972). *Degree words* (Vol. 53). Walter de Gruyter.

Brants, T., & Franz, A. (2006). Web 1t 5-gram Version 1.

Cliff, N. (1959). Adverbs as multipliers. *Psychological Review*, *66*(1), 27.

Davis, R. (1961, August). The Fitness of Names to Drawings. a Cross-Cultural Study in Tanganyika. *British Journal of Psychology*, *52*(3), 259–268. doi: 10.1111/j.2044-8295.1961.tb00788.x

Declerck, R. (1991). *Comprehensive Descriptive Grammar of English, A*. Kaitakusha.

De Saussure, F. (1916). Nature of the linguistic sign. *Course in general linguistics*, 65–70.

Douglas, D., & Broussard, K. M. (2000). Longman grammar of spoken and written English. *TESOL Quarterly*, *34*(4), 787–788.

Fechner, G. (1860). Elements of psychophysics. Translation, H. *Adler. Brietkoph & Hrtel*, *1*.

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

Franke, M. (2011). Quantity implicatures, exhaustive interpretation, and rational conversation. *Semantics and Pragmatics*, *4*, 1–1.

Garner, B. A. (2016). The Chicago Guide to Grammar, Usage, and Punctuation. *SPRING BOOKS 2016*, *16*, 3.

Givn, T. (1993). *English grammar: A function-based introduction* (Vol. 2). John Benjamins Publishing.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Greenbaum, S. (1996). *The Oxford English Grammar* (Vol. 652). Oxford University Press Oxford.

Grice, H. P. (1975). Logic and conversation. *1975*, 41–58.

Hausman, J. A., & Ruud, P. A. (1987, January). Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics*, *34*(1), 83–104. doi: 10.1016/0304-4076(87)90068-6

Holland, M. K., & Wertheimer, M. (1964). Some physiognomic aspects of naming, or, maluma and takete revisited. *Perceptual and Motor Skills*, *19*(1), 111–117.

Huddleston, R. (1984). *Introduction to the Grammar of English*. Cambridge University Press.

Huddleston, R., & Pullum, G. K. (2002). The Cambridge grammar of the English language. *Cambridge*, *14*, 199–212.

Judd, C. M., Westfall, J., & Kenny, D. A. (2017). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology*, *68*, 601–625.

Kanwal, J., Smith, K., Culbertson, J., & Kirby, S. (2016). The evolution of Zipfs law of abbreviation. In *The Evolution of Language: Proceedings of the 11th International Conference (EVOLANG11).*

Kao, J. T., Bergen, L., & Goodman, N. D. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the Thirty-Sixth Annual Conference of the Cognitive Science Society.*

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

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

Kennedy, C., & McNally, L. (2005). Scale structure, degree modification, and the semantics of gradable predicates. *Language*, 345–381.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and philosophy*, *4*(1), 1–45.

Koriat, A., & Levy, I. (1979). Figural symbolism in Chinese ideographs. *Journal of Psycholinguistic Research*, *8*(4), 353–365.

Khler, W. (1970). *Gestalt psychology: An introduction to new concepts in modern psychology*. WW Norton & Company.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and linguistic theory* (Vol. 23, pp. 587–610).

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. MIT Press.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lewis, M. L. (2016). *Conceptual complexity and the evolution of the lexicon*. Unpublished doctoral dissertation, Stanford University.

Mahowald, K., Fedorenko, E., Piantadosi, S. T., & Gibson, E. (2013, February). Info/information theory: Speakers choose shorter words in predictive contexts. *Cognition*, *126*(2), 313–318. doi: 10.1016/j.cognition.2012.09.010

McCusker, L. (1977). Some determinants of word recognition: Frequency. In *24th annual convention of the southwestern psychological association, fort worth, tx.*

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., ... others (2011). Quantitative analysis of culture using millions of digitized books. *science*,

331(6014), 176–182.
Nelson, G. (2010). *English: an essential grammar*. Routledge.
Qing, C., Goodman, N. D., & Lassiter, D. (2016). A rational speech-act model of projective content. In *Proceedings of the Thirty-Eighth Annual Conference of the Cognitive Science Society.*
Quirk, R. (1972). *A grammar of contemporary English*. Longman Group.
Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. London and New York: Longman.
Quirk, R., & others. (1990). *A student's grammar of the English language*. Pearson Education India.
Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesiaa window into perception, thought and language. *Journal of consciousness studies*, *8*(12), 3–34.
Russell, B. (2012). *Probabilistic reasoning and the computation of scalar implicatures*. Unpublished doctoral dissertation, Citeseer.
Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*, *128*(3), 302–319.
Sutton, R. S., & Barto, A. G. (2011). *Reinforcement learning: An introduction*. Cambridge Univ Press.
Van Gelderen, E. (2010). *An Introduction to the Grammar of English: Revised Edition*. John Benjamins Publishing.
Wheeler, S. C. (1972). Attributives and their modifiers. *Nos*, 310–334.
Yoon, E. J., Tessler, M. H., Goodman, N. D., & Frank, M. C. (2017). "I won't lie, it wasn't amazing": Modeling polite indirect speech. In *Proceedings of the Thirty-Ninth Annual Conference of the Cognitive Science Society.*
Zipf, G. K. (1935). The psycho-biology of language.

Appendix A

Lassiter and Goodman (2013)'s model of scalar adjectives belongs to the family of Rational Speech Act (RSA) models in which speaker and listener communicate by reasoning about each other's goals and inferences (Frank & Goodman, 2012; Goodman & Stuhlmüller, 2013; for related models, see also Franke, 2011; Russell, 2012). These models have been shown to account for a number of key phenomena in pragmatics. The adjective model accounts for uncertainty about the adjectival threshold by including a semantic variable, which the pragmatic listener infers at the same time that she infers the speaker's intended meaning.

RSA models begin with a literal listener, which captures the semantic denotation of sentences. We assume adjective phrases with the same scale and polarity have the same denotation. For example, *expensive*, *very expensive* and *phenomenally expensive* all denote: $\lambda x.\text{price}(x) > \theta_i$. However, every adjective phrase has its own threshold variable θ_i,[15] together notated $\vec{\theta}$, allowing their meanings to differ. Given an utterance u_i (e.g. an *expensive laptop* or a *very expensive laptop*) and a set of thresholds, a literal listener L_0 will use Bayesian inference to update his prior beliefs $P(d)$ about the degree d (e.g. the laptop's price) given that the degree is greater than the threshold for that utterance.

$$P_{L_0}(d|u_i,\theta_i) \propto P(d)\cdot\delta_{d>\theta_i}$$

A speaker with the goal of communicating some actual degree d assigns a utility $\mathbb{U}(u_i|d)$ to each utterance such that he prefers utterances which will inform the literal listener, but avoids utterance cost, $C(u_i)$:

$$\mathbb{U}(u_i|d,\vec{\theta}) = \ln\left(P_{L_0}(d|u_i,\theta_i)\right) - C(u_i)$$

Given a set of alternative utterances (e.g. the speaker might be choosing between saying *very expensive* as opposed to *expensive* or *extremely expensive*, or saying nothing at all), the speaker S_1 will choose utterances according to a softmax decision rule (Sutton & Barto, 2011) with optimality parameter λ, so that:

$$P_{S_1}(u_i|d,\vec{\theta}) \propto e^{\lambda\mathbb{U}(u_i|d,\vec{\theta})}$$

A pragmatic listener L_1 uses the prior probability, $P(d)$, of different degrees, along with knowledge of the cost of each utterance, in order to guess both the thresholds for each utterance and which degree the speaker intended to communicate[16]:

$$P_{L_1}(d,\vec{\theta}|u_i) \propto P(d)\cdot P_{S_1}(u_i|d,\vec{\theta})$$

We simulated such a model with three alternative adjective phrases (i.e. three intensifiers) with costs of 1, 5, and 10. We also included a null utterance, with trivial meaning (always true) and cost of 0. The prior distribution of degrees along this adjective's scale (which we will discuss as "prices" for concreteness and consistency with our Experiment 1) was a gaussian peaked at 0. We used an optimality parameter of $\lambda = 5$ in our simulation.

Though the literal semantics are identical (but permitting different threshold parameters), the different phrases received different interpretations: the more costly intensifiers corresponded to less probable, more extreme prices (Figure 10). This can be seen as an M-implicature: more costly intensifiers are assigned stronger, less probable, meanings. The model therefore predicts an association between intensifier meaning and utterance cost (see Bergen et al. (2014) for other M-implicature models within the RSA framework).

[15] Other versions of this model could easily be imagined in which the threshold for an adjective phrase is determined by the basic threshold for the adjective and some transformation on that threshold caused by the intensifier. For example, (Cliff, 1959) argue that intensifying adverbs operate on adjectives meanings through scalar multiplication. If, as in multiplication, the transformation is regular, with a single parameter needing to be inferred for each intensifier, and if the values of these parameters are inferred for each adjective phrase, then such a model would be functionally equivalent to the one we describe here.

[16] We assume a uniform prior on thresholds θ_i.

To assess the quantitative relationship between cost and meaning, we ran a second simulation, identical as the first except using 6 different utterance costs (or "intensifiers"). The quantitative form predicted by the model is a approximately linear (Figure 11). It is this simple prediction that we test in the main text.

Appendix B

Below, we present results of linear mixed-effects models of scaled responses in Study 1a as a function of length in syllables and surprisal. Model 1 includes raw predictors. In Model 2, we replace the length predictor by the residuals when length is modeled as a linear function of surprisal. In Model 3, we replace the surprisal predictor by the residuals when surprisal is modeled as a linear function of length. Finally, we present results of likelihood ratio tests when each of the predictors is removed from the full model.

Study 1a Regression Results

Model 1: Colinear Predictors

	Estimate	se	t	p
Length	0.133	0.077	1.72	0.094
Surprisal	0.106	0.031	3.41	0.002

Model 2: Length, residualized by Surprisal

	Estimate	se	t	p
$\text{Length}_{\text{resid}}$	0.133	0.077	1.72	0.094
Surprisal	0.120	0.030	3.98	< 0.001

Model 3: Surprisal, residualized by Length

	Estimate	se	t	p
Length	0.202	0.075	2.69	0.010
$\text{Surprisal}_{\text{resid}}$	0.106	0.031	3.41	0.002

Model Comparison: Predictors removed from Model 1

	χ^2	p
Length	13.70	0.003
Surprisal	17.05	0.001

Appendix C

Below, we present results of linear mixed-effects models of scaled responses in Study 1b as a function of length in syllables and surprisal. Model 1 includes raw predictors. In Model 2, we replace the length predictor by the residuals when length is modeled as a linear function of surprisal. In Model 3, we replace the surprisal predictor by the residuals when surprisal is modeled as a linear function of length. Finally, we present results of likelihood ratio tests when each of the predictors is removed from the full model.

Study 1b Regression Results

Model 1: Colinear Predictors

	Estimate	se	t	p
Length	0.055	0.054	1.02	0.311
Surprisal	0.107	0.033	3.26	0.002

Model 2: Length, residualized by Surprisal

	Estimate	se	t	p
$\text{Length}_{\text{resid}}$	0.055	0.054	1.02	0.311
Surprisal	0.128	0.026	4.96	< 0.001

Model 3: Surprisal, residualized by Length

	Estimate	se	t	p
Length	0.166	0.042	3.93	< 0.001
$\text{Surprisal}_{\text{resid}}$	0.107	0.033	3.26	0.002

Model Comparison: Predictors removed from Model 1

	χ^2	p
Length	11.57	0.009
Surprisal	30.32	< 0.001

Appendix D

Below, we present results of linear mixed-effects models of rankings in Study 2 as a function of length in syllables and surprisal. Model 1 includes raw predictors. In Model 2, we replace the length predictor by the residuals when length is modeled as a linear function of surprisal. In Model 3, we replace the surprisal predictor by the residuals when surprisal is modeled as a linear function of length. In Model 4, we add interactions for adjectives. Finally, we present results of likelihood ratio tests when each of the length and surprisal predictors are removed from Model 1.

Study 2 Regression Results

Model 1: Colinear Predictors

	Estimate	se	t	p
Surprisal	0.164	0.016	10.13	< 0.001
Length	0.243	0.041	5.89	< 0.001

Model 2: Length, residualized by Surprisal

	Estimate	se	t	p
Surprisal	0.190	0.016	11.66	< 0.001
$\text{Length}_{\text{resid}}$	0.243	0.041	5.89	< 0.001

Model 3: Surprisal, residualized by Length

	Estimate	se	t	p
Length	0.350	0.041	8.46	< 0.001
$\text{Surprisal}_{\text{resid}}$	0.164	0.016	10.13	< 0.001

Model 4: With Adjective Interaction

	Estimate	se	t	p
Surprisal	0.074	0.030	2.48	0.013
Length	0.318	0.087	3.66	< 0.001
Surprisal:Adj				
Adj=Tall	0.188	0.045	4.14	< 0.001
Adj=Expensive	0.170	0.047	3.66	< 0.001
Adj=Old	0.037	0.043	0.85	0.393
Length:Adj				
Adj=Tall	-0.103	0.121	-0.86	0.393
Adj=Expensive	-0.100	0.119	-0.84	0.400
Adj=Old	-0.092	0.119	-0.78	0.438

Model Comparison: Predictors removed from Model 1

	χ^2	p
Length	34.69	< 0.001
Surprisal	108.08	< 0.001

Appendix E

Below, we present results of regressions predicting scaled responses for novel intensifiers in Study 3.

Study 3 (novel intensifiers) Regression Results				
Model 1: Short/Long Novel Adverbs				
	Estimate	se	*t*	*p*
(Intercept)	0.110	0.366	0.299	0.768
Length	0.538	0.195	2.757	0.011
Model 2: Length and Root Type				
	Estimate	se	*t*	*p*
(Intercept)	0.100	0.197	0.510	0.615
Length	0.540	0.195	2.772	0.011
Ratum/other	-0.371	0.143	-2.601	0.016
Lopus/Bugorn	0.073	0.235	0.311	0.759

Below, we present results of linear mixed-effects models of scaled responses in Study 3's replication portion as a function of length in syllables and surprisal. Model 1 includes raw predictors. In Model 2, we replace the length predictor by the residuals when length is modeled as a linear function of surprisal. In Model 3, we replace the surprisal predictor by the residuals when surprisal is modeled as a linear function of length. Finally, we present results of likelihood ratio tests when each of the predictors is removed from the full model.

Study 3 (replication portion) Regression Results				
Model 1: Colinear Predictors				
	Estimate	se	*t*	*p*
Length	0.194	0.095	2.05	0.085
Surprisal	0.140	0.056	2.49	0.041
Model 2: Length, residualized by Surprisal				
	Estimate	se	*t*	*p*
$\text{Length}_{\text{resid}}$	0.194	0.095	2.05	0.085
Surprisal	0.182	0.052	3.47	0.010
Model 3: Surprisal, residualized by Length				
	Estimate	se	*t*	*p*
Length	0.286	0.088	3.24	0.016
$\text{Surprisal}_{\text{resid}}$	0.140	0.056	2.49	0.041
Model Comparison: Predictors removed from Model 1				
	χ^2	*p*		
Length	8.30	0.040		
Surprisal	39.19	< 0.001		

Appendix F

Below, we present results of regression predicting rankings for novel intensifiers in Study 4.

Study 4 (novel intensifiers) Regression Results				
Model 1: Short/Long Novel Adverbs				
	Estimate	se	*z*	*p*
Length	0.70	0.253	2.79	0.005
Model 2: Length and Root Type				
	Estimate	se	*z*	*p*
Length	0.757	0.263	2.88	0.004
Ratum/other	-0.047	0.172	-0.27	0.786
Lopus/Bugorn	-0.518	0.324	-1.60	0.110

Below, we present results of linear mixed-effects models of rankings in Study 4's replication portion as a function of length in syllables and surprisal. Model 1 includes raw predictors. In Model 2, we replace the length predictor by the residuals when length is modeled as a linear function of surprisal. In Model 3, we replace the surprisal predictor by the residuals when surprisal is modeled as a linear function of length. Finally, we present results of likelihood ratio tests when each of the predictors is removed from the full model.

Study 4 (replication portion) Regression Results				
Model 1: Colinear Predictors				
	Estimate	se	*t*	*p*
Surprisal	0.447	0.032	13.88	< 0.001
Length	0.581	0.054	10.74	< 0.001
Model 2: Length, residualized by Surprisal				
	Estimate	se	*t*	*p*
Surprisal	0.573	0.037	15.54	< 0.001
$\text{Length}_{\text{resid}}$	0.581	0.054	10.74	< 0.001
Model 3: Surprisal, residualized by Length				
	Estimate	se	*t*	*p*
Length	0.875	0.063	13.94	< 0.001
$\text{Surprisal}_{\text{resid}}$	0.447	0.032	13.88	< 0.001
Model Comparison: Predictors removed from Model 1				
	χ^2	*p*		
Length	130.33	< 0.001		
Surprisal	231.57	< 0.001		

Table 2: Intensifiers from Study1b, number of occurences in Google Web 1T 5grams corpus, and length in syllables.

intensifier	freq	syll
dreadfully	147917	3
fantastically	250989	4
supremely	296134	3
suspiciously	398581	4
strikingly	480417	3
noticeably	632679	4
awfully	651519	3
unbelievably	686210	5
downright	876782	2
excessively	877280	4
extraordinarily	900456	6
exceedingly	977435	4
tremendously	989532	4
enormously	1011751	4
immensely	1061341	3
hugely	1074430	2
intensely	1084765	3
profoundly	1172521	3
infinitely	1226005	4
amazingly	1384225	4
unusually	1583939	4
outright	1662351	2
wonderfully	1776763	3
remarkably	1902493	4
terribly	1906059	3
sharply	2377367	2
utterly	2507480	3
positively	3225521	4
extensively	3447083	4
mighty	3492518	2
surprisingly	3554188	4
altogether	3683374	4
purely	4201779	2
wholly	4308225	2
incredibly	4416030	4
badly	4808245	2
considerably	4834700	5
sufficiently	5059075	4
good and	5671809	2
thoroughly	6167601	3
damn	6930185	1
deeply	7242890	2
perfectly	10031907	3
greatly	11337773	2
largely	11379702	2
very much	11415215	3
strongly	13931652	2
entirely	14720396	4
plain	15433319	1
absolutely	16064235	4
truly	19778608	2
totally	20950052	3
extremely	21862963	3
dead	28609410	1
completely	32310795	3
highly	36460329	2
extra	36838459	2
easily	39241261	3
fully	41415591	2
pretty	43623658	2
simply	50172762	2
quite	55269390	1
rather	66341863	2
real	144660526	1
really	148918637	2
too	159399185	1
way	268084494	1
very	292897993	2
well	301853777	1
most	324420476	1
so	518878130	1

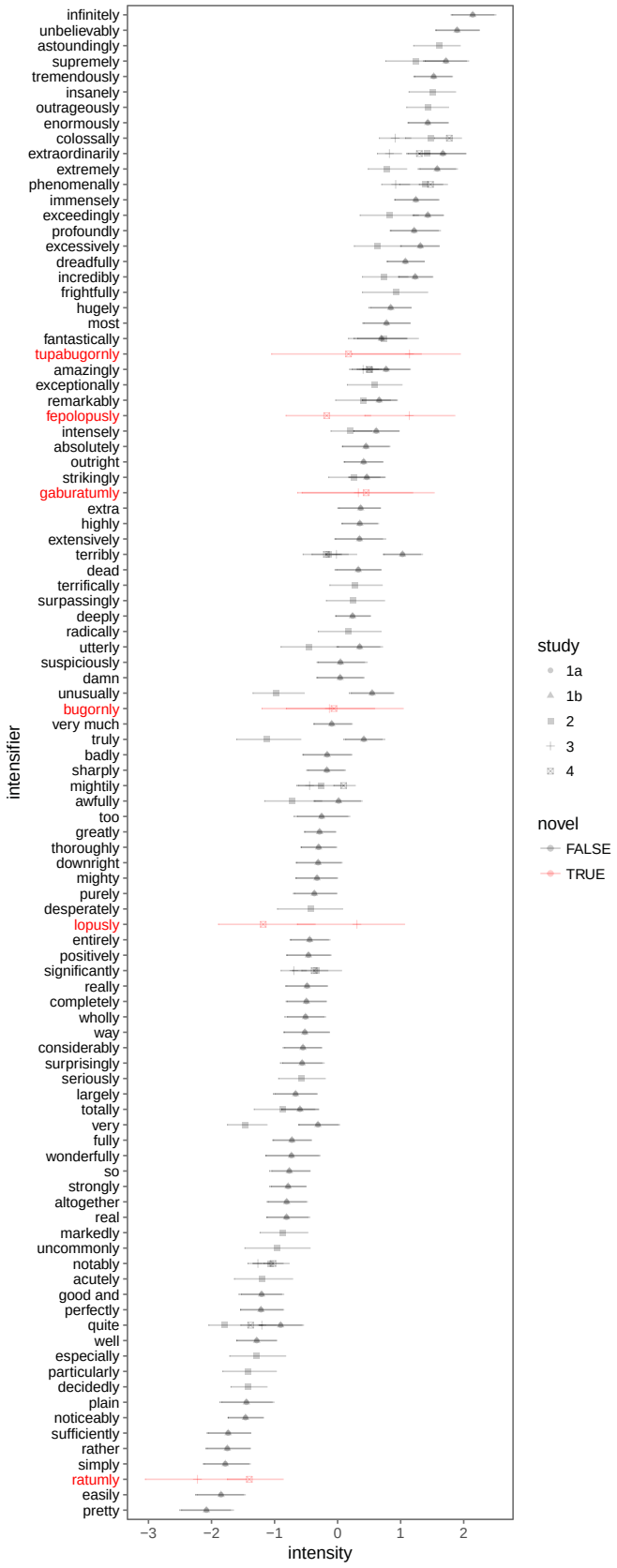

Figure 9: Intensities (each dependent measure, z-scored) of all intensifiers across all 5 studies. Novel intensifiers are shown in red, standard English intensifiers in black.

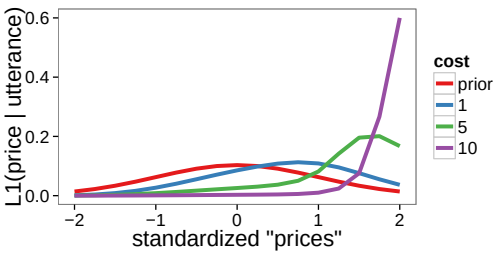

Figure 10: Modeling intensifiers as M-implicature: more costly intensifiers correspond to more extreme meanings.

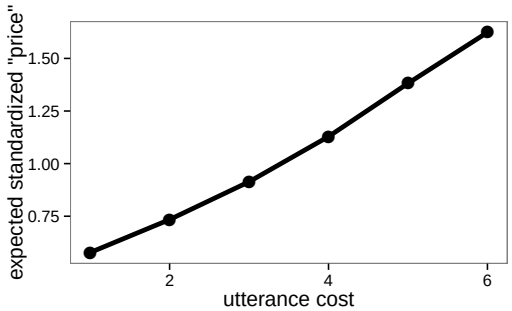

Figure 11: Model prediction of expected price as cost of intensifier increases, based on intensifiers evenly spaced in cost. The relationship is approximately linear.