Extremely costly intensifiers are stronger than quite costly ones

Erin 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 test this prediction in two 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 (Köhler, 1947; 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, Sugarman, & Frank, 2014). 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 both because they are amenable to simple quantitative measures of meaning (such as the numeric extent to which they shift the interpretation of a scalar adjective) and because theoretical considerations, which we lay out below, suggest a relationship between their meaning and their usage cost (i.e. their frequency and length).

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. The impact of word length is reminiscent of the results of Lewis et al. (2014), who studied noun categories. While word frequency is known to have major effects on sentence processing (Levy, 2008, e.g.), the prediction that frequency should affect meaning is more surprising.

We confirm, in our first two experiments, that English intensifiers in adjective phrases are indeed interpreted as much stronger for both longer and less frequent intensifiers. This holds in quantitative judgments of meaning and in forced comparisons, and across a number of adjectival dimensions. In our second two experiments, 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 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.

Previous researchers have proposed that adjective phrases

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

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 *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 experiments, where participants interperet 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.

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 by assigning a separate threshold to each intensified (or bare) adjective phrase. That is, 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 heard and all alternative sentences, is then established by a pragmatic inference that takes into account the differing costs of the sentences. 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), 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 below.

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 not nullify the hypothesized relationship between cost and meaning.

Factors affecting utterance cost

We have identified the 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

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

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

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 adjectival phrase. This leaves open the the relative importance of length and surprisal (as well as other factors that might enter into cost), which can be explored via regression models.

Utterance cost predicts intensifier strength

The proposal detailed above predicts an association between measures of cost and strength of interpretations. In our first two experiments, we tested whether our measures of cost can in fact predict the intensity of scalar adjective interpretation.

Experiment 1

In Experiment 1, we test the qualitative prediction that as cost of an utterance increases, so will the interpreted meaning of the intensifier. We tested this prediction by eliciting free response price estimates from people for phrases such as *very expensive watch* and determining whether these prices are correlated with our independent measures of utterance cost.

Methods[5] 30 participants with US IP addresses were recruited through Amazon's 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.

We asked participants to estimate the prices of different objects based on different descriptions of those objects. The descriptions 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). 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 explamation, as in "Super!"). Each particpant 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. We chose the domain of price and used only the adjective *expensive* because price constitutes a quantitative scale with standard units (dollars for our US participants) on which to measure the different intensifers.

Corpus Methods Table 1 shows word frequency and length in syllables for the intensifiers used in the experiment. 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, but not strongly so ($r = 0.26$).

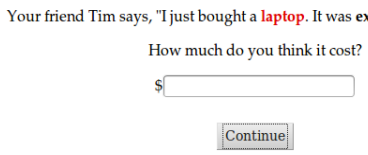

Figure 1: Screenshot from Experiment 1 target question.

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 case.

We ran a linear mixed effects regression with centered fixed effects of syllables and surprisal and random intercepts and slopes for both participant and object. We used the logarithm of participants' price estimates as the dependent variable, because of evidence that people's representation of numbers, including prices, is logarithmic (Dehaene, 2003, e.g.).[7]

Our results are shown in Figure 2, in a way that highlights the surprisal predictor. Both measures of cost play a role in predicting participants' price estimates. We found a significant main effect of surprisal ($\beta = 0.05, SE = 0.01, t(3) = 5.72, p = 0.009$) such that less frequent words tend to be associated with higher price estimates. We also found a significant main effect of syllable length ($\beta = 0.06, SE = 0.02, t(3) = 3.54, p = 0.036$), above and beyond surprisal, such that longer words predict stronger meanings.[8]

Thus intensifiers that are less frequent and longer (and therefore are more costly to utter) also tend to be interpreted as having stronger meanings, at least when used to modify *expensive*. Furthermore, the relationship appears to be linear in surprisal and length, as predicted. This is consistent with the M-implicature proposal introduced above.

[5] The full experiment can be found at `http://cocolab.stanford.edu/cogsci2015/intensifiers/Experiment1`

[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).

[8] Because surprisal and syllable-length are correlated, in addition to the analysis reported here we also computed length-corrected surprisals by regressing the raw surprisals onto word length in syllables using linear regression with a single main effect of word length. The residuals of this model, length-corrected surprisals, were used for the surprisal predictor in similar analyses to the ones we report, and we consistently found similar results. This suggests more strongly that surprisal and syllable length *both* independently contribute to participants' estimates.

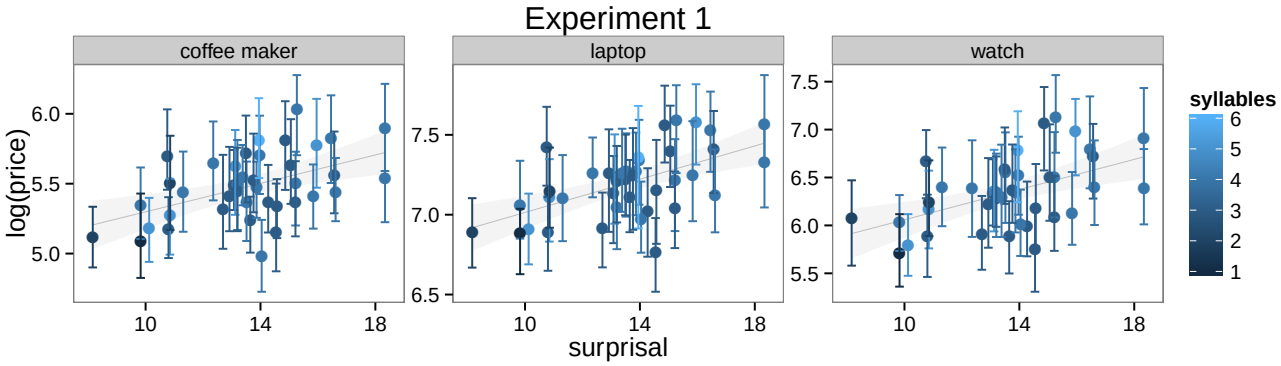

Figure 2: Results of Experiment 1. As surprisal and length in syllables increase, participants' free response prices increase.

Experiment 2

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 adjectival phrases in which it occurs[9]. To explore this issue, extend our results to additional adjectival scales. However, most scales are not so easily quantifiable as price; we require a different dependent measure in order to probe them. For Experiment 2 we used a forced-ranking dependent measure, which allows us to consider additional adjectival scales. This dependent measure has the added benefit of providing a more sensitive measure of the differences in degrees between similar adjectival phrases.

Methods[10] 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.

We asked participants to order (by clicking and dragging) various adjective phrases with the same adjective but different intensifiers according to strength of meaning. 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. We divided the 40 intensifiers from Experiment 1 into four lists of 10 intensifiers. Each list was randomly paired with one of four adjectives (*old*, *expensive*, *beautiful*, and *tall*). For each adjective-list pairing, participants were shown every combination of the 10 intensifiers with one adjective. 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 3). Each participant completed four such trials, seeing all four lists and all four adjectives. The pairings between list and adjective were randomized between participants. The division of the intensifiers into lists of 10 was constant, to simplify data analysis, so that the same 10 intensifiers were always shown together.

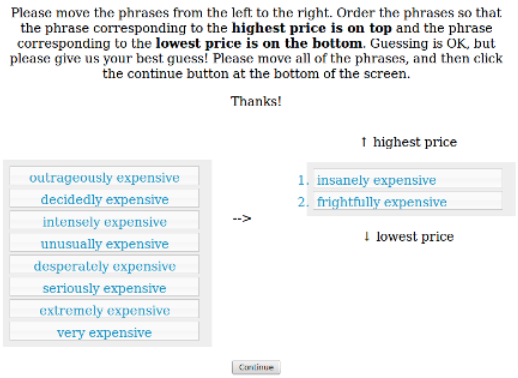

Figure 3: Screenshot from Experiment 2 target question.

Results Our results for Experiment 2 are shown in Figure 4. 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 and no intercept. As in Experiment 1, we found strong effects of surprisal ($\beta = 0.16, SE = 0.02, t = 10.13, p < 0.001$) and syllable length ($\beta = 0.24, SE = 0.04, t = 5.89, p < 0.001$). 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 syllables. We again found significant effects of surprisal ($p = 0.013$) and syllables ($p < 0.001$), and also found a significant interaction between surprisal and the adjective being modified (estimates for *tall* and *expensive* were significantly higher than for *beau-*

[9] If the bigram frequency of the modified adjective ("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 be a relatively small effect, and the relevant bigrams were too sparse in our corpora to pursue.

[10] The full experiment can be found at `http://cocolab.stanford.edu/cogsci2015/intensifiers/Experiment2`

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

ngram	frequency	syllables
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

tiful, while *old* was not significantly different from *beautiful*). This suggests that context-specific surprisal 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.

Discussion

These experiments provide evidence that intensifier meanings depend systematically on the length and frequency of distribution of their word forms. While it is unlikely that this accounts for all intensifier meaning, it does suggest that a major portion of meaning comes not from arbitrary, conventional association of signal to sign (de Saussure, 1916), but systematically from features of the word's form and distribution.

Since this is a correlational study, such a relationship does not confirm that an intensifier's cost *causes* it to have a given meaning. Rarity in particular might be correlated with strength of meaning merely because more extreme meanings refer to less probable things in the world, are therefore talked about less, and therefore the words with those meanings will necessarily be rarer. 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. Syllable length in turn can depend on the frequency, simplicity, or predictability of a word (Zipf, 1935; Lewis et al., 2014; Piantadosi, Tily, & Gibson, 2011), either because words that are frequently used get shortened over time (Lewis & Frank, 2015) or perhaps because words that refer to simpler or more common concepts enter the lexicon sooner (when more shorter word forms remain unassigned to meanings). It is therefore possible that these measures of cost have no 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 Experiments 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.

Cost effects for novel intensifiers

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.

Experiment 3[11]

We first replicate our findings from Experiment 1 when we use novel intensifiers rather than existing ones.

Method 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 Enlgish speaker.

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

We varied the novel intensifier between participants from a set of 6 novel intensifiers, three of which were relatively short

[11]The full experiment can be found at `http://cocolab.stanford.edu/cogsci2015/intensifiers/Experiment3`

[12] We chose this subset of 9 intensifiers to get a wide range of surprisals and syllable lengths (*colossally*, *phenomenally*, *mightily*, *extraordinarily*, *amazingly*, *terribly*, *notably*, *significantly*, *quite*)

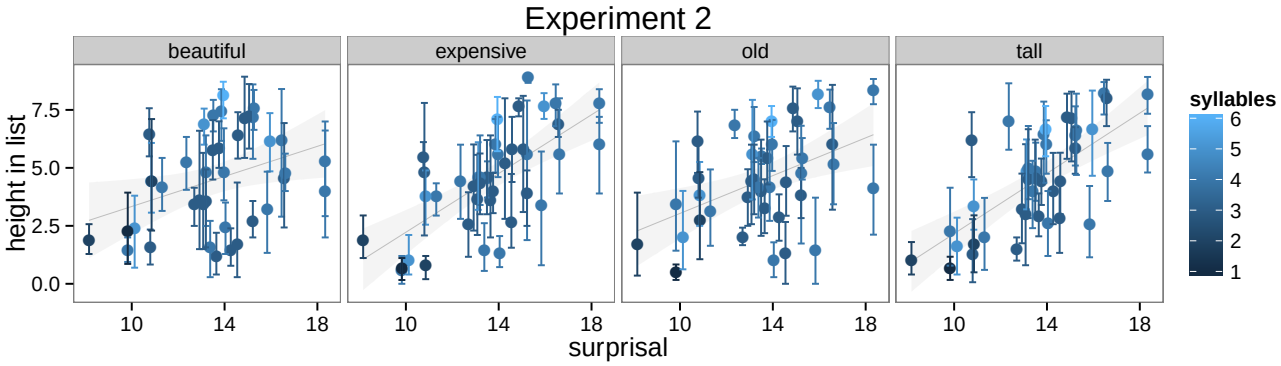

Figure 4: Results of Experiment 2. As surprisal and length in syllables increase, participants' rankings increased.

(*lopusly*, *ratumly*, and *bugornly*) and three of which shared the same "root" but were two CVCV syllables longer (*fepolopusly*, *gaburatumly*, and *tupabugornly*).

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.

Results In Experiment 3, we included as filler a subset of the intensifiers we tested in Experiment 1, and so we first confirmed our findings from Experiment 1. As in Experiment 1, we ran a linear mixed effects regression with fixed effects of syllables and surprisal, and random intercepts and slopes for both participant and object, and we used the logarithm of participants' price estimates as the dependent variable. Replicating our findings from Experiment 1, we found significant main effects of surprisal ($\beta = 0.09$, $SE = 0.03$, $t = 3.55$, $p = 0.026$) and syllable length ($\beta = 0.11$, $SE = 0.02$, $t = 4.92$, $p < 0.001$) (Figure 5).

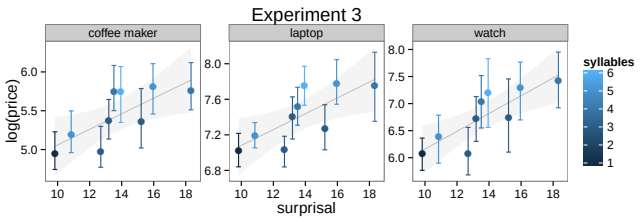

Figure 5: In Experiment 3, we replicated our findings from Experiment 1.

We then ran a linear mixed effects model on only the novel intensifiers, with length ("long" or "short") as a fixed effect, random intercepts and slopes for objects, and random intercepts for the three different "roots". We found a significant effect of length condition (β("short") $= -1.46$, $SE = 0.39$, $t = -3.74$, $p < 0.001$), indicating that people use the length of an intensifier in order to interpret its meaning, even for novel intensifiers with no conventional meaning (Firgure 6). In a post-hoc regression with a fixed effect for novel adverb root, we found a significant effect of root on response ($p = 0.016$), suggesting possible additional effects of form that we have not captured with length in syllables alone. Average responses for *ratumly* were lowest out of all the intensifiers used in Experiment 3, and average responses for *tupabugornly* were highest. The rest of the novel intensifiers had average ratings within the range of the attested intensifiers.

Experiment 4

In Experiment 4, we replicate our findings from Experiment 2, this time with novel intensifiers ranked relative to standard English ones.

Method[13] 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.

Experiment 4 was identical to Experiment 2, except that each participant saw exactly one of two adjectives (*expensive* or *tall*, varied between participants) and only the set of intensifiers from Experiment 3. This set included one novel intensifier, which we varied between participants. As in Experiment 2, adjective phrases for each intensifier-adjective pairing were initialized in a random order.

Results With our filler intensifiers for Experiment 4, again using a ranked order logit model (re-ranking to ignore novel intensifiers), we replicated our findings from Experiment 2 of significant effects of both surprisal ($\beta = 0.45$, $SE = 0.03$, $t = 13.88$, $p < 0.001$) and syllable length ($\beta = 0.58$, $SE = 0.05$, $t = 10.74$, $p < 0.001$) on the order in the list that participants chose for the intensifiers.

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 and found a significant

[13]The full experiment can be found at `http://cocolab.stanford.edu/cogsci2015/intensifiers/Experiment4`

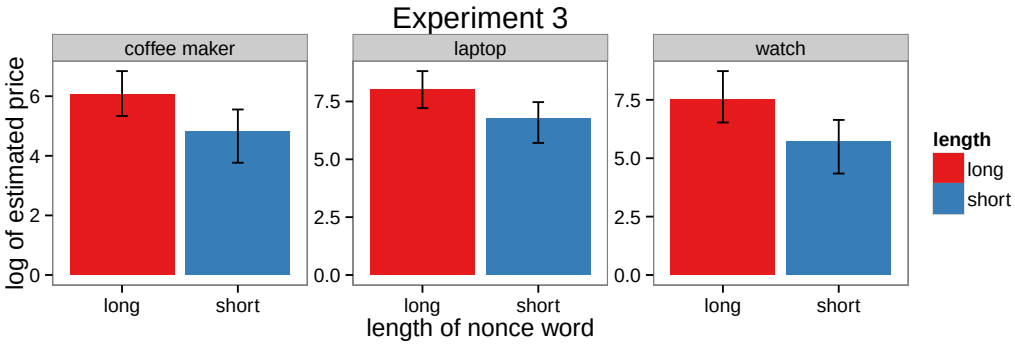

Figure 6: In Experiment 3, we found a significant effect of length for all novel intensifiers.

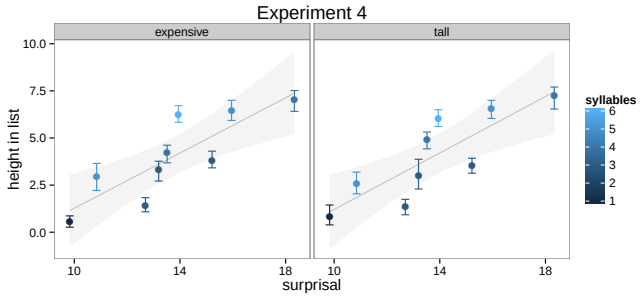

Figure 7: In Experiment 4, we replicated our finding from Experiment 2: longer and less frequent intensifiers are ranked higher than shorter and more frequent ones.

effect of length condition (β("short") $= -1.41$, $SE = 0.51$, $t = -2.79$, $p = 0.005$). When we included a regressor for the root of the novel intensifier, we did not find a significant effect of root on participants' relative rankings (for root *lopus*, $p = 0.110$; for root *ratum* $p > .250$). 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*).

Discussion

Overall, in Experiments 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). Either way, these results demonstrate that the relationship between word cost and meaning is not a static 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 model predicted a linear relationship between the intensity of an intensifier and its cost, measured here in terms of length and negative log-frequency. In four experiments we provided evidence that intensifier meanings do depend systematically on the length and frequency of distribution of those word forms and that this relationship holds even for novel words. While it is unlikely that this accounts for all intensifier meaning, it does 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. 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. In addition, the intensifiers we have looked at may have other source(s) of meaning in addition to the measures of utterance cost that we have explored. This may be because we have not exhausted the sources of utterance cost. It could also be that in addition to the relationship we have described between utterance cost and strength, conventional contributions to meaning are associated with certain adverbs. 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").

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). While we have described some evidence for this distributional source of meaning, further work will be necessary to separate the influence of the form of a word from the influence of its distribution. A fuller understanding of these factors would also enable us to explore word-types that support a similar relationships between distribution and meaning.

We have shown 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.

References

Beggs, S., Cardell, S., & Hausman, J. (1981). Specifying and testing econometric models for rank-ordered data. *Journal of econometrics*, *17*(1), 1–19.

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

Bolinger, D. (1972). *Degree words*. Paris: Mouton.

Brants, T., & Franz, A. (2006). *Web 1T 5-gram Version 1*. Philadelphia: Linguistic Data Consortium.

Davis, R. (1961). The fitness of names to drawings: a cross-cultural study in tanganyika. *British Journal of Psychology*.

Dehaene, S. (2003). The neural basis of the Weber-Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences*, *7*(4), 145–147.

de Saussure, F. (1916). *Nature of the linguistic sign.*

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

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

Hausman, J. A., & Ruud, P. A. (1987). Specifying and testing econometric models for rank-ordered data. *Journal of econometrics*, *34*(1), 83–104.

Holland, M., & Wertheimer, M. (1964). Some physiognomic aspects of naming, or *maluma* and *takete* revisited. *Perceptual and Motor Skills*.

Kennedy, C. (2007). Vagueness and grammar: The semantics of relative and absolute gradable adjectives. *Linguistics and Philosophy*.

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

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

Köhler, W. (1947). *Gestalt psychology* (Second ed.). Liveright.

Koriat, A., & Levy, I. (1979). Figural symbolism in chinese ideographs. *Journal of Psycholinguistic Research*.

Lassiter, D., & Goodman, N. D. (2013). Context, scale structure, and statistics in the interpretation of positive-form adjectives. In *Semantics and Linguistic Theory (SALT) 23*.

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

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*.

Lewis, M., & Frank, M. C. (2015). Conceptual complexity and the evolution of the lexicon..

Lewis, M., Sugarman, E., & Frank, M. C. (2014). *The structure of the lexicon reflects principles of communication.*

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*.

Piantadosi, S. T., Tily, H., & Gibson, E. (2011). Word lengths are optimized for efficient communication. *Proceedings of the National Academy of Sciences*, *108*(9).

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). A comprehensive grammar of the english language.

Ramachandran, V. S., & Hubbard, E. M. (2001). Synaesthesia – a window into perception, thought and language. *Journal of Consciousness Studies*.

Smith, N. J., & Levy, R. (2013). The effect of word predictability on reading time is logarithmic. *Cognition*.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. MIT press.

Wheeler, S. C. (1972). Attributives and their modifiers. *Noûs*.

Zipf, G. K. (1935). The psycho-biology of language.

Appendix

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). 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 adjectival 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,[14] 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, 1998) 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[15]:

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

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 8). 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.

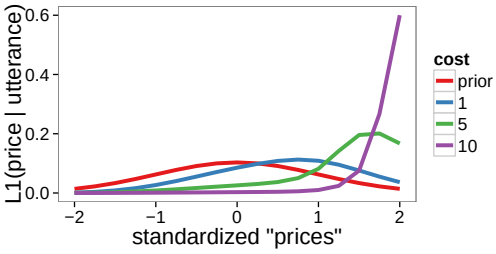

Figure 8: Modeling intensifiers as M-implicature: more costly intensifiers correspond to more extreme meanings.

Top 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 9). It is this simple prediction that we test in the main text.

[14] 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 (e.g. multiplication, addition, etc.) caused by the intensifier. If the transformation is mostly 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.

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

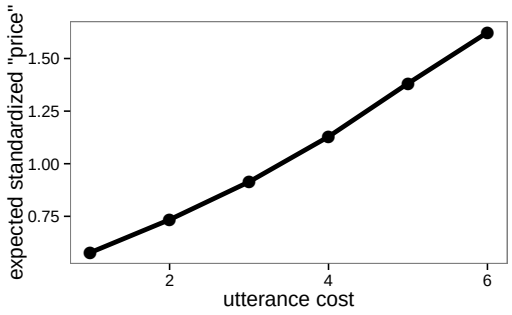

Figure 9: Model prediction of expected price as cost of intensifier increases, based on intensifiers evenly spaced in cost. The relationship is approximately linear.