Extremely costly intensifiers are stronger than quite costly ones: a case of non-arbitrary word meanings.

Erin Bennett (erindb@stanford.edu)
Department of Psychology, 450 Serra Mall , Stanford, CA 94305

Noah Goodman (ngoodman@stanford.edu)
Department of Psychology, 450 Serra Mall , Stanford, CA 94305

Abstract

abstract

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

Introduction

is there a nice saussure quote to summarize the arbitrariness position?

Where do words get their meanings? For instance, why is an “extremely good paper” better than a “quite good paper”? The traditional answer (?, ?) is that different meanings have been arbitrarily and conventionally assigned to the different word forms. In this paper we explore adjectival intensifiers, like “extremely” and “quite”, as a case study in which to empirically explore the relationship of meaning to form. Based on a previous theory of the semantics of scalar adjectives we hypothesize that the interpreted meanings of intensifiers are (at least partly) non-arbitrary, and instead are determined by their production (or comprehension) cost. We show in three experiments that the meanings of english intensifiers are predictable from their costs, and are sensitive to manipulation of cost. These results are consistent with the small but growing literature arguing that word meanings are not fully arbitrary, but instead are constrained by non-semantic features of the word.

a paragraph here unpacking the question of word-meaning mapping. start with what saussure said about signs. this suggests that meaning is purely the result of historical accident. certainly a lot of meaning for a lot of words must be arbitrary – without convention there is no language. mention the recent work on modeling language evolution. however, there is some evidence that meaning is not entirely arbitrary: kiki-boba stuff, and maybe molly’s results. to resolve this tension we will need more, and more quantitative case-studies where we can explore the extent of non-arbitrary factors.

next a paragraph talking about adjectives and intensifiers. scalar adjs probably have a threshold semantics (cite kennedy) and the threshold is set in a pragmatic way that depends on context. lassiter and goodman give a formal model of setting the threshold. given this story about adjectives, what could intensifiers do? it’s possible to posit a complex semantic mechanism by which they grab and alter the threshold which has been set for the adjective (as if they weren’t there). however a more parsimonious hypothesis is that they don’t have any semantic effect–instead their “meaning” is a result of their cost and the impact this has on pragmatic inference. possibly stick in the model result (without model details) to show that this really would be expected to happen?

perhaps a short paragraph drawing out the implication of this story: asking whether the word-form to lexical-semantics mapping is arbitrary is not the right question, because pragmatics is inextricably part of the process of interpretation, and pragmatics is sensitive to additional factors which are not purely phonetic or lexical semantic.

note: can we explain why bad can be used to mean very good?

finally a very brief paragraph giving the roadmap of our experiments.

Intensifiers, like “extremely” and “quite”, are adverbs that modify scalar adjectives to change the degree of the resulting adjective phrase. Scalar adjectives have been modeled as having implicit thresholds that need to be inferred from the context, and intensifiers seem to raise that threshold, e.g. the threshold above which people are “extremely tall” is higher than the threshold above which people are "tall". Some intensifiers seem to change this threshold more than others, and the extent to which they do might be influenced by online or conventionalized M-implicature.

Experiment 1

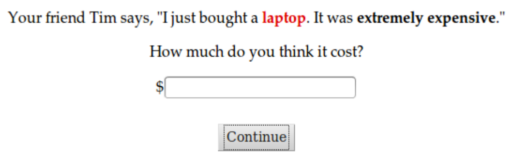
To explore the hypothesis that the interpretations of intensifiers are a function of their cost, we first wanted to see whether two possible ways of measuring the cost of a word, frequency (rarer words are probably more costly) and syllable length, were related to the interpretations of intensifiers.

l to say
we think
uency
sylla-
number
ld be
cts of
t". point
hat this
es open
relative
ortance,
they
bine, and
t other
rs might
r into

Method[1]

40 participants with US IP addresses participated in our Experiment 1 on Amazon's Mechanical Turk.

We asked participants to give us judgements of prices based on a person's description of an object that included an intensifier (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 definitely 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 adjective, as in "super hero"). Each particpant gave price judgements for every intensifier-category pairing in randomized order, for a total of 120 price judgements. We chose the domain of price and used only the adjective "expensive", because price gave a quantitative scale on which to measure the different intensifers and because we thought participants would have similar enough experience with the distributions over prices for these objects.



Your friend Tim says, "I just bought a laptop. It was extremely expensive."

How much do you think it cost?

$

Figure 1: Screenshot from Experiment 1 target question.

Corpus Methods In order to measure the cost associated with different intensifiers, we collected their length in syllables and their frequencies (Table 1). The frequencies were collected from the Google Web 1T 5-grams database (?, ?)[2] The syllable lenths of our intensifiers and the surprisals were correlated, but not strongly so (r = 0.2648144).

what sur-
l is and
we care
re using

Results and Discussion

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

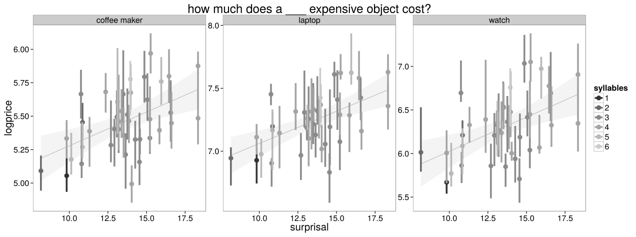


Figure 2: Results of Experiment 1. As surprisal and length in syllables increase, participants' free response prices increased.

In a linear mixed effects regression with centered fixed effects of syllables and surprisal and their interaction and random intercepts and slopes for syllables and surprisal for both participant and object, we found significant main effects of surprisal (estimate=0.054, p=0.012) and syllable length (estimate=0.093, p=0.0041) as well as a significant interaction (estimate=0.019, p=0.00018).

The interaction suggests that the function from surprisal and frequencies to cost might be multiplicative.

So intensifiers that are more surprising and longer (and therefore are more costly to utter) also tend to be interpreted as having stronger meanings.

make a big deal

Experiment 2

In Experiment 2, we extend our finding from Experiment 1 to other adjectival scales (in addition to "expensive"). We use a ranking dependent measure which is more appropriate to non-quantitative scales and which we expect to be more sensitive to small differences in meaning.

Method[3]

30 participants with US IP addresses participated in our Experiment 2 on Amazon's Mechanical Turk.

introduce the idea of the ranking measure first.

Because arranging all 40 intensifiers on a computer screen would be difficult for participants, we divided the 40 intensifiers from Experiment 1 into four lists of 10 intensifiers

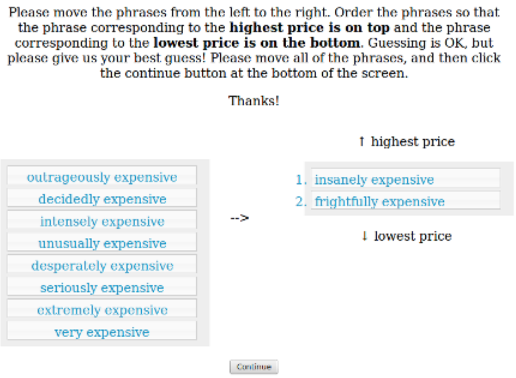
[1] The full experiment can be found at `http://web.stanford.edu/~erindb/degree-adverbs/experiments/exp5_2014-12-01/exp5.html`

[2] We also ran the same analyses on frequency information collected from the Google Books American Ngrams Corpus (?, ?) as well, and found similar results.

In addition, we did the same using the bigram frequencies of "*[intensifer]* expensive" rather than the unigram frequencies of the intensifiers alone. These data were much more sparse. For bigrams, we found no significant effects of surprisal using the books database and a negative effect using the web database.

[3] The full experiment can be found at `http://web.stanford.edu/~erindb/degree-adverbs/experiments/exp4/exp4.html`

each (Table 2). 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 and the one adjective on the left side of the screen. They 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 did four trials of this process, 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, i.e. the same 10 intensifiers were always shown together to ease data analysis.


Please move the phrases from the left to the right. Order the phrases so that the phrase corresponding to the **highest price is on top** and the phrase corresponding to the **lowest price is on the bottom**. Guessing is OK, but please give us your best guess! Please move all of the phrases, and then click the continue button at the bottom of the screen.

Thanks!

↑ highest price

outrageously expensive
decidedly expensive
intensely expensive
unusually expensive
desperately expensive
seriously expensive
extremely expensive
very expensive

-->

1. insanely expensive
2. frightfully expensive

↓ lowest price

Figure 3: Screenshot from Experiment 2 target question.

Results and Discussion

Within each intensifier list, we ran a regression using centered syllable length and surprisal to predict the ranking that participants gave the adjective phrase (the highest ranked adjective phrase in a trial got a ranking of 10, the lowest ranked adjective phrase got a ranking of 1). See Figure 4. For the two lists with a large enough range of syllable lengths, we fully replicated our results from Experiment 1. For the two lists with smaller syllable ranges, we replicated the main effect of surprisal, but found different effects of syllable length. Results were very similar across the four different adjectives.

can we do a combined regression where list is a factor (and/or random effect)? rank-within-list would still be the DV. this would allow a common estimate of effect of surprisal across the expt?

For the lists A and B, which each contained five different lengths of syllables, we found significant main effects of surprisal (list A: estimate=0.45, p=1.1e-7; list B: estimate=0.45, p=1.9e-15) and syllable length (list A: estimate=0.87, p=0.0051; list B: estimate=1.3, p=2e-16) and a significant interaction (list A: estimate=0.25, p=0.034; list B: estimate=0.20, p=1.0e-6), as in Experiment 1. For lists C and D, which had only three different syllable lengths, we found main effects of surprisal (list C: estimate=0.36, p=3.3e-5; list D: estimate=0.46, p=9.6e-6), but no positive effect of syllable length and no positive interaction. For list C, there was no main effect of syllable length (p=0.49) and a negative interaction (estimate=-0.52, p=0.0045). For list D, there was a negative main effect of syllable length (estimate=-1.4, p=4.5e-6) and a negative interaction (estimate=-0.23, p=0.024).

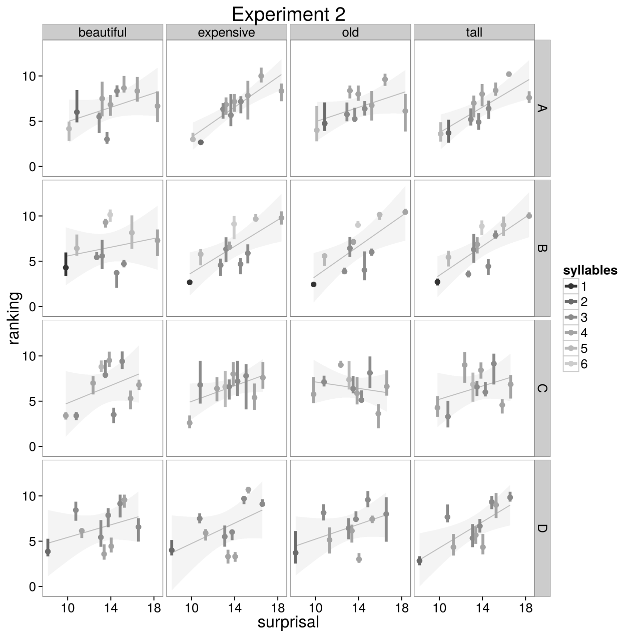


Figure 4: Results of Experiment 2. As surprisal and length in syllables increase, participants' rankings increased.

Overall, we again found that participants assign stronger interpretations to intensifiers with lower frequencies and higher syllable lengths.

more about this?

The relationship between frequency and interpretation might be causal, and the causal direction might be that the rarity of the word causes it to be costly to use and therefore to correspond to a stronger meaning, as in our hypothesis. However, the causal direction could also be the opposite. Perhaps the fact that an intensifier has a stronger meaning (which it may have gotten completely arbitrarily) causes it to be used only in extreme and unusual circumstances. Since these circumstances rarely occur, the strong intensifier will rarely be said[4]. This seems possible, but it would not account for syllable length contributing to intensifier meaning above and beyond surprisal, which was the case overall in Experiments 1 and 2. We explore this issue further in the next experiment.

[4] This assumes that people talk about things about as frequently as they happen, which might not be the case... Isn't someone here working on how representative the internet is of what actually happens, and super rare things have an inflated presence on the web? Which is kind of evidence that people talk about extreme things more than they actually happen.

Experiment 3

To test the direction of influence between intensifier frequency and meaning, we expose participants to an imaginary dialect and manipulate the frequency with which intensifiers occur. If people use the frequency of an intensifier in order to interpret it, and the rarity of an intensifier causes its interpreation to be stronger, then changing and intensifier's frequency should change its interpretation.

Method[5]

20 participants with US IP addresses participated in our Experiment 3 on Amazon's Mechanical Turk.

We trained participants on a dialect that used one of two short intensifiers, “truly” and “very” much more frequently than in standard English. The speaker of this dialect, Jim, was a character in a comic who lived “across the country” in a town with “a distinct way of speaking”. We showed participants a 9-panel comic in which Jim told his visiting cousin about a big storm that had knocked down a tree into his kitchen and about a friend's child who had taken part of the tree home with him (Figure 5). Jim said 294 words in the training story, 22 of which were the target intensifier (either “truly” or “very”, varied between participants).

After the training story, participants were immediately shown a final panel, where Jim described a coffee maker he recently purchased, but part of his utterance was missing (Figure 6). Participants were asked to give a price judgement for each of three different possible utterances: the two intensifiers, and the bare “expensive” form. One of these intensifiers was the target intensifier which occured in the training story, and one intensifier was the control intensifier which did not occur in the story. So for each of the two intensifiers, some participants gave ratings for it as a target intensifier and some participants gave ratings for it as a control intensifier.

Results and Discussion

We calculated the difference score between the each of the intensifiers and the bare adjective “expensive”. We compared this difference score for each of the intensifiers when the intensifier was the target intensifier (highly frequent) and when it was the control (normal English frequency, but no occurances in the training story).

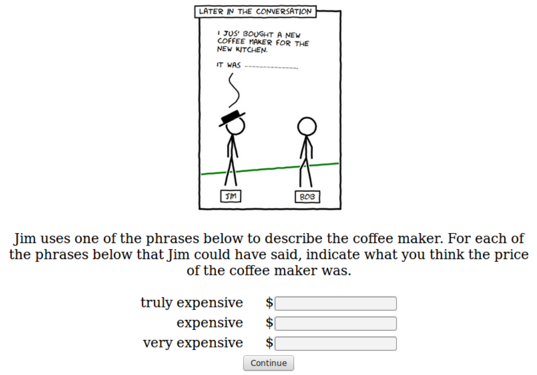
If infrequency causes an intensifier to be stronger, then we would expect participants would infer that the word is more frequent in this dialect and consequently less strong. The difference score for the target intensifier would then be lower than for the control intensifier.

We found that when participants believed the speaker's use of a word was much higher, they believed the meaning the speaker intended to convey with the word was lower (Fig 7). The difference between “*[intensifier]* expensive” and “expensive” was less for the target intensifier than for the control intensifier. In a linear gregression with word type as

analysis_files_for_writeup/images/story1_truly_circles.pr
analysis_files_for_writeup/images/story2_truly_circles.pr
analysis_files_for_writeup/images/story3_truly_circles.pr

Figure 5: Full training story comic for Experiment 3, target intensifier “truly” is repeated 22 times, control target “very”

[5] The full experiment can be found at `http://web.stanford.edu/~erindb/degree-adverbs/experiments/exp8/exp8.html`


LATER IN THE CONVERSATION
I JUS' BOUGHT A NEW COFFEE MAKER FOR THE NEW KITCHEN.
IT WAS ______________.
JIM
BOB
Continue

Jim uses one of the phrases below to describe the coffee maker. For each of the phrases below that Jim could have said, indicate what you think the price of the coffee maker was.

truly expensive $

expensive $

very expensive $

Figure 6: Screenshot from Experiment 3 target question.

a fixed effect and random intercepts for word and participant, word type was a significant predictor of difference score (estimate=-31.39, p=0.0226).

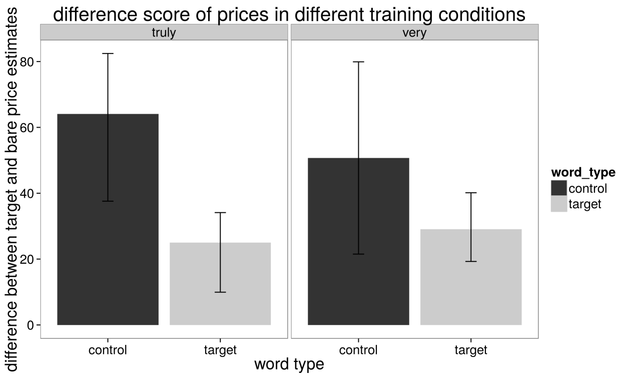


Figure 7: Results of Experiment 3. Price estimate for intensifier is lower after the intensifier is repeated (target condition), showing that overuse within a dialect results in a less strong meaning.

In a linear regression with word type (target or control) as a fixed effect and random intercepts for word and participant, word type was a significant predictor of frequency (estimate=34.06, p=0.0405).

this is cool because we manipulated the frequency and the price estimate consequently dropped.

Discussion

revisit our point and summarize briefly our results.

discuss other probable contributions to intensifier meaning: polarity, affect, others?

if room could say something about next steps. also relation to other's work on meaning.

close by restating our big picture point, that effective meaning is a combination of both arbitrary convention and non-arbitrary factors mediated by pragmatic inference.

conclusion

Acknowledgments

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

Table 2: Intensifier Lists from Experiment 2: Rankings.

List A	List B	List C	List D
surpassingly	colossally	terrifically	frightfully
astoundingly	phenomenally	uncommonly	outrageously
fantastically	mightily	supremely	insanely
strikingly	acutely	awfully	decidedly
excessively	extraordinarily	exceedingly	intensely
markedly	amazingly	radically	unusually
remarkably	terribly	exceptionally	desperately
utterly	notably	incredibly	seriously
truly	significantly	totally	extremely
particularly	quite	especially	very