

On the negativity of negation*

Christopher Potts
Stanford University

Abstract Natural language negation is persistently negative in the pragmatic sense, and emphatic and attenuating negative polarity items modulate this effect in systematic ways. I use large corpora of informal texts with meta-data approximating features of the context to characterize this pragmatic negativity, and I attempt to explain it in terms of the ways in which negative sentences engage the questions under discussion. The discussion highlights some of the ways in which quantitative corpus methods can be used to achieve novel results in linguistic pragmatics.

Keywords: scales, negation, negative polarity, expressives, corpus pragmatics

1 Negative tendencies

The negation of linguistic semantics and pragmatics tends to be a benign reverser of truth-values, true to the Fregean injunction that “A negation may occur anywhere in a sentence without making the thought indubitably negative” (Frege 1919). There is a sense in which this analysis is obviously correct; many negative sentences are unburdened by pragmatic negativity, so it would seem mistaken to build such expressivity into the basic meaning of negative morphemes.

What, then, are we to make of the observation that negation is “Learned early on with the association of ‘unpleasant feelings’ ” (Russell 1948, cited by Horn 1989: 164), it is associated with “falsity, absence, deprivation, and evil” (Israel 2004: 706), and its name is synonymous with repudiation, nullification, and rejection? This is the portrait of a caustic, taboo expressive. Our theories should explain why being a literal nay-sayer will make you seem, well, negative.

The primary goal of this paper is to help characterize, and quantify, the sense in which negation tends to be pragmatically negative. To do this, I draw on large corpora of informal texts collected from the Internet, especially from the Internet

* My thanks to David Beaver, Larry Horn, Sven Lauer, Marie de Marneffe, Tyler Schnoebelen, and the audience at SALT 20. A special thanks to Anastasia Giannakidou, Michael Israel, and Bill Ladusaw for inspiring discussions of negation and negative polarity, and to Chris Davis, Alex Djalali, Scott Grimm, and Florian Schwarz for detailed comments on an earlier draft. This research was supported in part by ONR award N00014-10-1-0109. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the view of ONR.

Movie Database and the Experience Project. The most important property of the texts in these corpora is that each is associated with meta-data approximating features of the context in which it was produced, yielding information about speaker/author goals and attitudes as well as hearer/reader reactions. In these corpora, negation patterns with scalar modifiers like *bad*, *depressing*, and *terrible*. Furthermore, as in effect predicted by Israel (1996, 2001, 2004), emphatic negative polarity items (e.g., *any*, *give a damn*) amplify this negativity whereas attenuating negative polarity items (e.g., *all that*, *overmuch*) soften it. The pressing theoretical question is then what causes these distributional profiles. I seek to explain them in terms of the ways in which negative sentences engage the abstract questions under discussion (Ginzburg 1995; Roberts 1996; Schoubye 2009; Beaver & Clark 2008; Simons, Tonhauser, Beaver & Roberts 2010), and I offer some additional evidence for that proposal from the Switchboard Dialog Act Corpus (Jurafsky, Shriberg & Biasca 1997).

In sec. 2, I introduce the data and methods. Sec. 3 is a first case study: positive and negative scalar modifiers, which illustrate the important correlations between language and the contextual meta-data and which set the stage for identifying the negativity of negation. Sec. 4 studies negation without associated polarity items, and sec. 5 shows how polarity phenomena affect the basic negative patterns. Finally, sec. 6 draws connections with other areas of lexical and constructional pragmatics, seeking in particular to highlight the ways in which quantitative corpus methods can be used to achieve novel non-categorical results about language use.

2 Data and methods

I rely primarily on two corpora. The first is a collection of nearly all the user-supplied movie and TV reviews from the Internet Movie Database (IMDB) as of March, 2010. The second is the full set of ‘confession’ texts at the Experience Project website as of June, 2010. The word-level data from these corpora, as well as two others used in secs. 4–5, are available from my website or upon request. This data distribution also includes an easy-to-use R function (R Development Core Team 2010) that performs the core calculations defined below and plots the results. Space considerations force me to pass over opportunities for statistical modeling (Constant, Davis, Potts & Schwarz 2009; Davis & Potts 2010; Potts & Schwarz 2010), but I think the visualizations and raw numbers paint a clear picture.

2.1 IMDB user-supplied reviews

The IMDB corpus consists of about 1.36 million user-supplied reviews of 45,772 movies and television shows. Each review has associated with it a star-rating, 1-star (most negative) to 10-star (most positive), chosen by the author to summarize her

Rating: 1 out of 10 stars
Review: For fans of the North and South series, this should never have been produced. Never, never, never never!! (If you have seen the first two Books and enjoyed them as most do, don't even consider viewing the third [...])

Rating: 5 out of 10 stars
Review: Two women compete with each other, seeing who can stay the youngest looking. Both go to a beautiful witch who has a youth potion, but they get more than they bargained for. Not all that funny to me.

Rating: 10 out of 10 stars
Review: This is the greatest TV series ever! I hope it hits the shelves! A movie would be da bomb! The special f/x are so cool! Too bad the series died. Hope for a renewal!!

Table 1 Short sample reviews from IMDB.

evaluation. Some sample reviews are given in tab. 1. These samples are about 40 words long, shorter than the corpus average of 233 words, but they are typical of the prose one finds in the corpus: direct and emotive, with little plot summarizing relative to evaluation, especially as compared with professional reviews. The emotive language is especially prominent at the extremes of the rating scale (1-star and 10-star), where the reviews tend to be impassioned pleas for or against the product under discussion (Potts & Schwarz 2010).

Tab. 2 gives summary numbers for the corpus as a whole, broken down by rating category.¹ The most noteworthy fact about the distribution is that the majority of reviews are highly positive, with 6-star to 10-star reviews accounting for 73% of the texts. This kind of imbalance is very common with review corpora collected from the Web (Pang & Lee 2008: §5.2.3.2) — an important observation for advertisers but something we need to abstract away from in linguistic work.

In this paper, I primarily study correlations between the rating categories and the words, phrases, and constructions in the review texts. The basic mode of counting is def. 1, which functions conceptually as though we took each token (at the linguistic level we care about) to be annotated with the star-rating of the text containing it.

Definition 1 (IMDB counts). Let $C = \{1 \dots 10\}$ be the set of rating categories for the IMDB corpus, and let π be a linguistic type (e.g., a morpheme, word):

$$\text{Count}_{\text{IMDB}}(\pi, c) \stackrel{\text{def}}{=} \text{the number of tokens of } \pi \text{ in IMDB reviews in } c \in C$$

¹ The word-level statistics differ slightly from those of de Marneffe, Manning & Potts (2010) because the tokenization algorithm used here preserves emoticons and more thoroughly strips off punctuation.

Rating	Reviews	Words	Vocabulary	Mean words/review
1	124,587 (9%)	25,395,214	172,346	203.84
2	51,390 (4%)	11,755,132	119,245	228.74
3	58,051 (4%)	13,995,838	132,002	241.10
4	59,781 (4%)	14,963,866	138,355	250.31
5	80,487 (6%)	20,390,515	164,476	253.34
6	106,145 (8%)	27,420,036	194,195	258.33
7	157,005 (12%)	40,192,077	240,876	255.99
8	195,378 (14%)	48,723,444	267,901	249.38
9	170,531 (13%)	40,277,743	236,249	236.19
10	358,441 (26%)	73,948,447	330,784	206.31
Total	1,361,796	317,062,312	800,743	232.83

Table 2 Basic statistics by rating category for the IMDB corpus.

By summing over these counts for linguistic types, we obtain a measure of the overall size of the categories, as in def. 2. This definition generalizes over corpora T and categories C so that it can also be used with both the IMDB counting method, def. 1, and the Experience Project and the other corpora introduced later.

Definition 2 (Category counts). Let T be a corpus partitioned by categories C , π a linguistic type, and Π the set of all linguistic types of the same class as π :

$$\text{Count}_{T,\Pi}(c) \stackrel{\text{def}}{=} \sum_{\pi \in \Pi} \text{Count}_T(\pi, c)$$

The notion of ‘class’ in this definition is hard to define generally, but it is typically clear for specific cases. At the level of words, Π is the full vocabulary and $\text{Count}_{\text{IMDB}}$ values correspond to column 3 in tab. 2. If we were looking at passive VPs, then Π would group all the different types of VPs. The crucial thing going forward is that the types that make up these classes do not share any tokens, or even any parts of tokens, so that nothing is counted more than once in def. 2.

The following uses the above counts to define two conditional distributions relating words and categories:

Definition 3. Let T be a corpus partitioned by categories C , π a linguistic type, and Π the set of all linguistic types of the same class as π :

- i. The probability of π given $c \in C$: $\Pr_{T,\Pi}(\pi|c) \stackrel{\text{def}}{=} \frac{\text{Count}_T(\pi, c)}{\text{Count}_{T,\Pi}(c)}$

- ii. The probability of $c \in C$ given π : $\Pr_{T,\Pi}(c|\pi) \stackrel{\text{def}}{=} \frac{\Pr_{T,\Pi}(\pi|c)}{\sum_{c' \in C} \Pr_{T,\Pi}(\pi|c')}$

Where the class is clear from context, I leave off the Π subscript from these values.

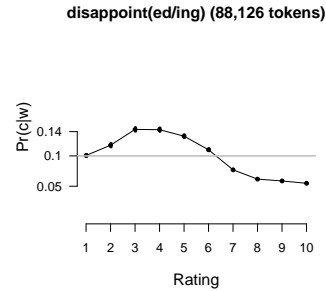
Clause (i) of def. 3 defines the conditional probability of π given category c . This is a speaker/author perspective: the author begins by selecting a rating that abstractly summarizes her attitude towards the product under review. She then composes the review with that rating category in mind. Clause (ii) turns this around: given a unit π , it determines a distribution over the categories. This is intuitively a hearer/author perspective: in light of the language the speaker is using, what is her internal state likely to be, as measured by the category? The two conditional distributions are related by Bayes rule under the assumption that each rating category is equally likely. (We have seen that this is not true — positive ratings are more common than negative ones. However, as noted above, this is a fact about online reviewing, not about language, so we abstract away from it.)

Throughout this paper, I work with the values provided by clause (ii). More specifically, for a given linguistic unit π and corpus T , I study the distribution of values $\Pr_T(c|\pi)$ across the different categories c . My primary reason for favoring $\Pr_T(c|\pi)$ values is that they abstract away from the overall frequency of π in the corpus, making it easy to place items with very different frequencies on the same scale for direct comparison. I'll have more to say about the comparative aspects of the analysis later on, when we begin studying specific linguistic phenomena.

Tab. 3(a) gives the distribution of $w = \text{disappoint}(ed|ing)$. The Count column gives the values for $\text{Count}_{\text{IMDB}}(w, c)$ for each rating c , and the Total column gives the $\text{Count}_{\text{IMDB}}(c)$ values. Dividing Count values by corresponding Total values gives the distribution $\Pr(w|c)$. Dividing those values by the sum of all the $\Pr(w|c)$ values yields $\Pr(c|w)$. These values highlight why ratings-relative values are so important. For example, there are about twice as many tokens of w in 10-star reviews as in 3-star reviews, but this is only because the 10-star category is about 5.25 times larger than the 3-star category. When we correct for this by using the relative values $\Pr(c|w)$ (or $\Pr(w|c)$), we see that w is nearly three times more frequent in 3-star reviews than in 10-star reviews.

Fig. 3(b) gives a richer, more intuitive picture of $w = \text{disappoint}(ed|ing)$. The black dots are the $\Pr_{\text{IMDB}}(c|w)$ values from tab. 3(a). Each is given with a 95% confidence interval. (In the IMDB data, these are often hard to see; the counts are so large that the intervals tend to be tiny. Confidence intervals play a more central role for the smaller Experience Project corpus.) In addition, the horizontal line marks the frequency we would expect if $\text{disappoint}(ed|ing)$ were equally probable in all rating categories. That is, this line depicts the hypothesis that the frequencies are independent of the rating categories, and the measured values and confidence intervals help support inferences about the effects of the rating categories on probabilities.

Cat.	Count	Total	$\text{Pr}_{\text{IMDB}}(w c)$	$\text{Pr}_{\text{IMDB}}(c w)$
1	8,557	25,395,214	0.0003	0.10
2	4,627	11,755,132	0.0004	0.12
3	6,726	13,995,838	0.0005	0.14
4	7,171	14,963,866	0.0008	0.14
5	9,039	20,390,515	0.0004	0.13
6	10,101	27,420,036	0.0004	0.11
7	10,362	40,192,077	0.0003	0.08
8	10,064	48,723,444	0.0002	0.06
9	7,909	40,277,743	0.0002	0.06
10	13,570	73,948,447	0.0002	0.05



- (a) Count gives the $\text{Count}_{\text{IMDB}}(w,c)$ values, Total the $\text{Count}_{\text{IMDB}}(c)$ values. For $\text{Pr}_{\text{IMDB}}(w|c)$, divide Count values by corresponding Total values. For $\text{Pr}_{\text{IMDB}}(c|w)$, divide $\text{Pr}_{\text{IMDB}}(w|c)$ values by the sum of all the $\text{Pr}_{\text{IMDB}}(w|c)$ values, as in def. 3.
- (b) The black dots represent $\text{Pr}_{\text{IMDB}}(c|w)$ values. The error bars around each point mark 95% confidence intervals. The horizontal line is the probability we would expect (always 0.1 for IMDB) if the word were equally probable in all categories.

Table 3 *disappoint(ed|ing)* in IMDB.

Here and throughout, the plots for the IMDB data are given with a y-axis that stretches from 0 to 0.3, the largest $\text{Pr}_{\text{IMDB}}(c|\pi)$ value reported in this paper, and the y-axis labels include the minimum and maximum values for the word in question. This makes it easy to compare distributions (Tufte 2001: §6). For example, as we see here, *disappoint(ed|ing)* has a maximum value of 0.14 (3-star and 4-star) and a minimum value of 0.05 (10-star). Skipping head to the leftmost panel in fig. 3(c), one can see that *not good* has a similar shape, but more pronounced: the peak is at 0.16 (4-star) and the minimum is at 0.03 (at 9-star).

For the IMDB, it is most natural to take the speaker’s perspective, in the sense that we can assume that the reviewer began the review with a specific emotive range in mind that is reflected in the assigned rating, and then wrote the review from that particular attitudinal vantage point. I can offer indirect evidence that the hearer’s (reader’s) perspective is also legitimate. Fig. 1 reports on an experiment conducted with Amazon’s Mechanical Turk (Snow, O’Connor, Jurafsky & Ng 2008; Munro, Bethard, Kuperman, Melnick, Potts, Schnoebelen & Tily 2010) in which subjects were presented with 130-character reviews from OpenTable.com and asked to guess which rating the author of the text assigned, where the ratings here are 1-star to 5-star. The figure depicts the actual rating assigned by the author on the x-axis, with

participants' guesses on the y-axis. (The responses have been jittered around so that they don't lie atop each other. The guesses were integers 1...5. The jittering helps to reveal where the responses clustered.) The plot also includes median responses (the black horizontal lines) and boxes surrounding 50% of the responses. The figure reveals that participants were able to guess with high accuracy which rating the author assigned; the median value is always the actual value, with nearly all subjects guessing within one star rating. A linear model using the actual rating to predict participants' guesses produces an excellent fit, with the average prediction just 0.81 stars from the empirical value (residual standard deviation = 0.81; $R^2 = 0.65$).

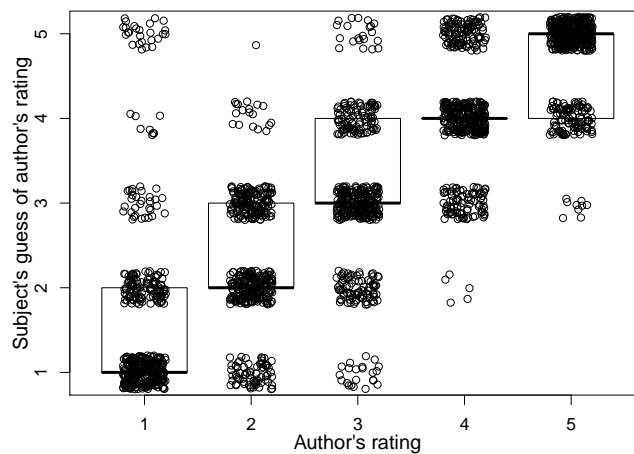


Figure 1 Results of an experiment in which participants read a short review from OpenTable.com and guessed which star rating, 1-5 stars, was assigned by the review's author. The individual data points have been jittered so that they don't all sit atop each other. The dark horizontal lines indicate the median guesses, with boxes surrounding 50% of the guesses.

It's worth mentioning that the IMDB website offers additional information beyond the reviews and the star ratings. There are also usernames, dates, reader assessments of helpfulness, and short summary texts that aim to capture the overall thrust of the review. In this paper, I set this additional meta-data aside for reasons of space, but it is there awaiting future experiments that bring this contextual information into the analysis for a sharper, more nuanced picture, perhaps by building multi-level regression models that can synthesize, e.g., utterance-level data with speaker- and topic-level data (Gelman & Hill 2007; Baayen, Davidson & Bates 2008; Jaeger 2008).

I note also that the experiments reported on here can be reproduced using data from OpenTable.com, Amazon.com, Tripadvisor.com, and Goodreads.com. All those sites use a basic five-star rating system, which makes them coarser than the IMDB scale, but they have the advantage of moving away from domain-specific effects deriving from TV and movies. In secs. 4–5, I draw on these corpora to further bolster the findings for negation.

2.2 Experience Project confessions

The IMDB meta-data are fairly straightforward: they roughly measure the author’s overall attitude towards the product under discussion, and they correlate well with the evaluative language people use. The next corpus I turn to, from the Experience Project (EP) website, is more subtle in the sense that its meta-data are more emotively complex. I focus on the ‘confessions’ at the site, which are short texts in which people tell brief revealing stories about themselves. The site offers readers a chance to react by clicking buttons for the categories ‘sorry, hugs’, ‘you rock’, ‘teehee’, ‘I understand’, and ‘wow, just wow’. Thus, each text is associated with a distribution over these reaction categories. Some sample confessions are given in tab. 4, along with their reactions. It should be born in mind that the reactions are *reader* responses. Whereas IMDB ratings captured something about the authors’ perspectives, EP reactions capture something about the readers’ perspectives.²

Confession: I really hate being shy ... I just want to be able to talk to someone about anything and everything and be myself... That’s all I’ve ever wanted.
Reactions: <i>hugs</i> : 1; <i>rocks</i> : 1; <i>teehee</i> : 2; <i>understand</i> : 10; <i>just wow</i> : 0;

Confession: subconsciously, I constantly narrate my own life in my head. in third person. in a british accent. Insane? Probably
Reactions: <i>hugs</i> : 0; <i>rocks</i> : 7; <i>teehee</i> : 8; <i>understand</i> : 0; <i>just wow</i> : 1

Confession: I have a crush on my boss! *blush* eeek *back to work*
Reactions: <i>hugs</i> : 1; <i>rocks</i> : 0; <i>teehee</i> : 4; <i>understand</i> : 1; <i>just wow</i> : 0

Table 4 Sample Experience Project confessions with associated reaction data.

² Readers can also comment on confessions. If we study the relationships between the comment texts and the reaction data, then we take a speaker perspective. I set the comments aside in what follows, for reasons of space. The patterns seem to be broadly the same for them, though I suspect there are pragmatically interesting differences to be uncovered. For example, a sad confession might elicit mostly ‘hugs’ and ‘I understand’ responses, but with comment texts dominated by the cheering language of pep-talks and motivational speeches.

The EP corpus is much smaller than the IMDB, but it is still substantial enough to study quantitatively. There are 27,187 texts with reaction data (out of 31,675 in total). The total word count for this part of the corpus is 3,132,620 with a vocabulary of 45,719 and an average text length of 99 words. Tab. 5 summarizes the reaction data. The counts listed there are the number of times people chose each reaction for some text. As with IMDB, the categories are not balanced: ‘I understand’ accounts for 48% of the responses, and ‘wow, just wow’ for only 4% of them. Thus, again, the methods need to correct for this imbalance, so that we get a picture of the linguistic usage conditions as opposed to the general tendencies of the site’s users.

Category	Count
‘sorry, hugs’	3,733 (16%)
‘you rock’	3,781 (16%)
‘teehee’	3,545 (15%)
‘I understand’	11,277 (48%)
‘wow, just wow’	916 (4%)

Table 5 Reactions in the Experience Project corpus.

The goal is again to study words and constructions relative to the meta-data. The situation is more complicated here than it was for IMDB, though, because we have not a single annotation (star rating), but rather a distribution over categories. For this paper, following suggestions by Tyler Schnoebelen (p.c.), I embrace the richness of these response distributions, as follows. Each text in the corpus is directly associated with a mapping from the set of EP reaction categories $C = \{hugs, rocks, teehee, understand, \text{and } just\ wow\}$ to the number of reactions each category elicited for that text. Just as we took each token in the IMDB corpus to be annotated with the rating of the text containing it, so too here we take each token to be annotated with the reaction distribution of the confession containing it. The following two definitions flesh out this basic method of counting:

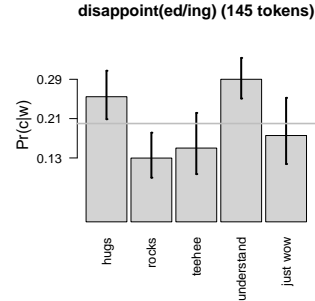
Definition 4 (EP token annotations). Let t be a token in the EP corpus and c an EP reaction category. Then $R(t, c)$ is the number of c choices for the text containing t .

Definition 5 (EP counts). Let π be a linguistic type (e.g., morpheme, word), π_t the set of tokens of π in the EP corpus, and c an EP reaction category:

$$\text{Count}_{\text{EP}}(\pi, c) \stackrel{\text{def}}{=} \sum_{t \in \pi_t} R(t, c)$$

With def. 5, we can use def. 2 to define $\text{Count}_{\text{EP}}(c)$ values for each of the EP categories, and we can use def. 3 to build conditional probability distributions. I illustrate in tab. 6(a), again using $\text{disappoint}(ed|ing)$ (cf. fig. 3).

Cat.	Count	Total	$\text{Pr}_{\text{EP}}(w c)$	$\text{Pr}_{\text{EP}}(c w)$
<i>hugs</i>	108	2,153,134	0.00005	0.25
<i>rocks</i>	34	1,330,084	0.00002	0.13
<i>teehee</i>	25	845,397	0.00003	0.15
<i>understand</i>	197	3,447,377	0.00006	0.29
<i>just wow</i>	29	838,059	0.00004	0.18



- (a) Count gives the word-level reactions for the corpus (def. 5), and Total sums over all the words' reaction counts, using def. 2 with Count_{EP} values. The $\text{Pr}_{\text{EP}}(c|w)$ values are again obtained by dividing each $\text{Pr}_{\text{EP}}(w|c)$ value by the sum of the $\text{Pr}_{\text{EP}}(w|c)$ values.
- (b) The $\text{Pr}_{\text{EP}}(c|w)$ values at left, with 95% confidence intervals, and a horizontal line marking the expected frequency assuming an even distribution across the categories (0.20). The token count is the number of times w occurs, which is different from the Count_{EP} values defined in def. 5 but which conveys a sense for the corpus's coverage of the item.

Table 6 $\text{disappoint}(ed|ing)$ in Experience Project.

The EP visualizations follow the same basic logic as employed for the IMDB, except now the categories are discrete and unordered rather than (arguably) continuous. Fig. 6(b) is the plot for $w = \text{disappoint}(ed|ing)$. The plot title includes the corpus-wide token count, to help provide a sense for how well represented the words are in the corpus. The bars represent the $\text{Pr}_{\text{EP}}(c|w)$ values, and the error bars delimit 95% confidence intervals.

The horizontal line is the probability we would expect if the word were equally distributed across the rating categories, adjusting for the categories' size differences. If a word's error bar is entirely above (below) this line for a category c , then we can be reasonably sure that it is genuinely over- (under-) represented in c , and if the error bars for two categories do not overlap in the y-axis, then we can be reasonably sure that their values are genuinely different. In fig. 6(b), both *hugs* and *understand* are highly probable, with *rocks* and *teehee* somewhat under-represented. It is very common for these categories to pair off like this in the data. Similarly, we don't have a good estimate for *just wow*, a common problem that seems to derive mainly from

this category being so small (tab. 5).

All the categories are quite emotively subtle. The *just wow* category is especially tricky. I am confident that it is not a straightforwardly exclamative response, though. It seems more accurate to say that it identifies negative heightened emotions like shock and disbelief. The words that are over-represented in this category tend to be things like *knife*, *cocaine*, *charge*, and *rage*. The more clearly positive exclamative category is *rocks*; confessions containing *wow*, *amazing*, *absolutely*, *what a(n)* and other markers of exclamation and intensification tend to elicit this response from readers.

3 Scalars

This section begins to lay the groundwork for the study of negation and polarity sensitivity by looking at scalar modifiers in the IMDB and EP corpora. The plots, given in fig. 3, employ the conventions described for the two corpora in the previous sections. The IMDB y-axis always stretches from 0 to 0.3, and the EP axis from 0 to 0.4. The y-axis labels include the minimum and maximum values for each item. These conventions should facilitate comparisons between shapes.

The positive scalar modifiers are in the first two subfigures, fig. 3(a) and fig. 3(b). Consider first the relationship between *great* and *amazing*. Both are heavily biased towards the positive end of the scale, with a climb in probability from 1-star to 10-star. This is in keeping with the fact that both are lexically positive scalar modifiers. The difference between the two is that *great* is milder than *awesome*; whereas *great* climbs in a roughly linear fashion, *awesome* is dramatically curved, nearly flat in the negative part of the scale and ramping up in the positive part to create a J-shaped distribution (Potts & Schwarz 2010; Davis & Potts 2010). We see a comparable pattern in the EP data. For both *great* and *awesome*, *rocks* and *teehee* are over-represented, but the pattern is much more pronounced for *awesome*.

The word *amazing* presents a more subtle profile. In the IMDB data, it looks much like another highly positive scalar modifier, with a distribution that is hard to distinguish from *awesome*. In the EP data, though, the well-known exclamativity of *amazing* shines through: *rocks* is by far the best represented category. Thus, the more nuanced categories of the EP corpus bring out an emotive dimension that is obscured by the simpler IMDB scale.

The word *good* also fits into the overall scalar picture, but the facts are slightly more complex, owing to the scale-reversing effects of negation. Whereas it is easy to negate *good*, such predications are highly infrequent and hard to motivate for *awesome* and *amazing*, so negation is less of a factor. To control for this, the *POS good* plots involve *good* outside the scope of negation, and the *NEG good* plots involve *good* in the scope of a negative morpheme (*not*, *n't*, *never*, and forms of *no*).

For both corpora, these data were collected at the clause level. I used punctuation to approximate this for the IMDB corpus. For the smaller EP corpus, greater precision was called for, so I parsed the data using the freely available Stanford parser (Klein & Manning 2003a,b) and then extracted the requisite patterns using the Stanford Tregex program (Levy & Andrew 2006), which provides intuitive functionality for extracting patterns from parsetrees.

In the leftmost panel of fig. 3(a), we see that *POS good* has a peak value just to the right of the middle of the rating scale. It is a mild positive modifier, fairly frequent throughout the corpus but skewed positive. The picture is similar in the EP data, leftmost in fig. 3(b): mild over-representation in the *rocks* category. The leftmost panels for the negative scalar modifiers, fig. 3(c) and fig. 3(d), are sharper in the sense that the negative bias is more pronounced in both cases: the peak value for *NEG good* is 4-star, and the *hugs* and *understand* categories, both of which express solidarity, sympathy, and compassion, are over-represented. As noted earlier, this is like a stronger version of *disappoint(ed|ing)* seen in tab. 3(a) and tab. 6(a).

The words *bad* and *terrible* are negative counterparts of *great* and *awesome*. Here, the bias is towards the negative end of the scale in the case of IMDB and towards *hugs* and *understand* in the case of EP. Once again, the EP data seem to be more subtle; whereas the pattern for *depress(ed|ing)* is mixed for the IMDB, we see a sharp rise in *understand* in the EP data, though *hugs* remains prominent.

Readers who try out the data and associated code can check that forms of *depress* are more clearly negative in the 5-star corpora. The mixed pattern for IMDB might derive from the fact that some good movies are intended to be depressing. An amusing example of such a genre effect is the distribution of *gross* in subcorpora consisting just of reviews of horror and romance movies. Grossness is apparently independent of quality for horror movies but a real drawback in romances:

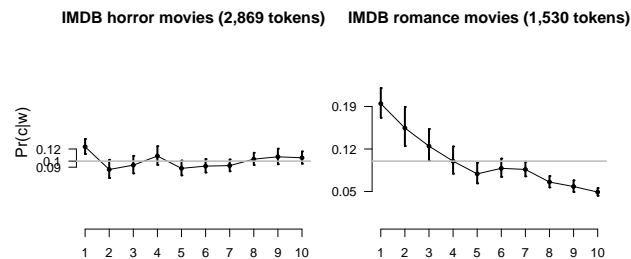


Figure 2 *gross* in subcorpora consisting of horror and romance movies.

On the negativity of negation

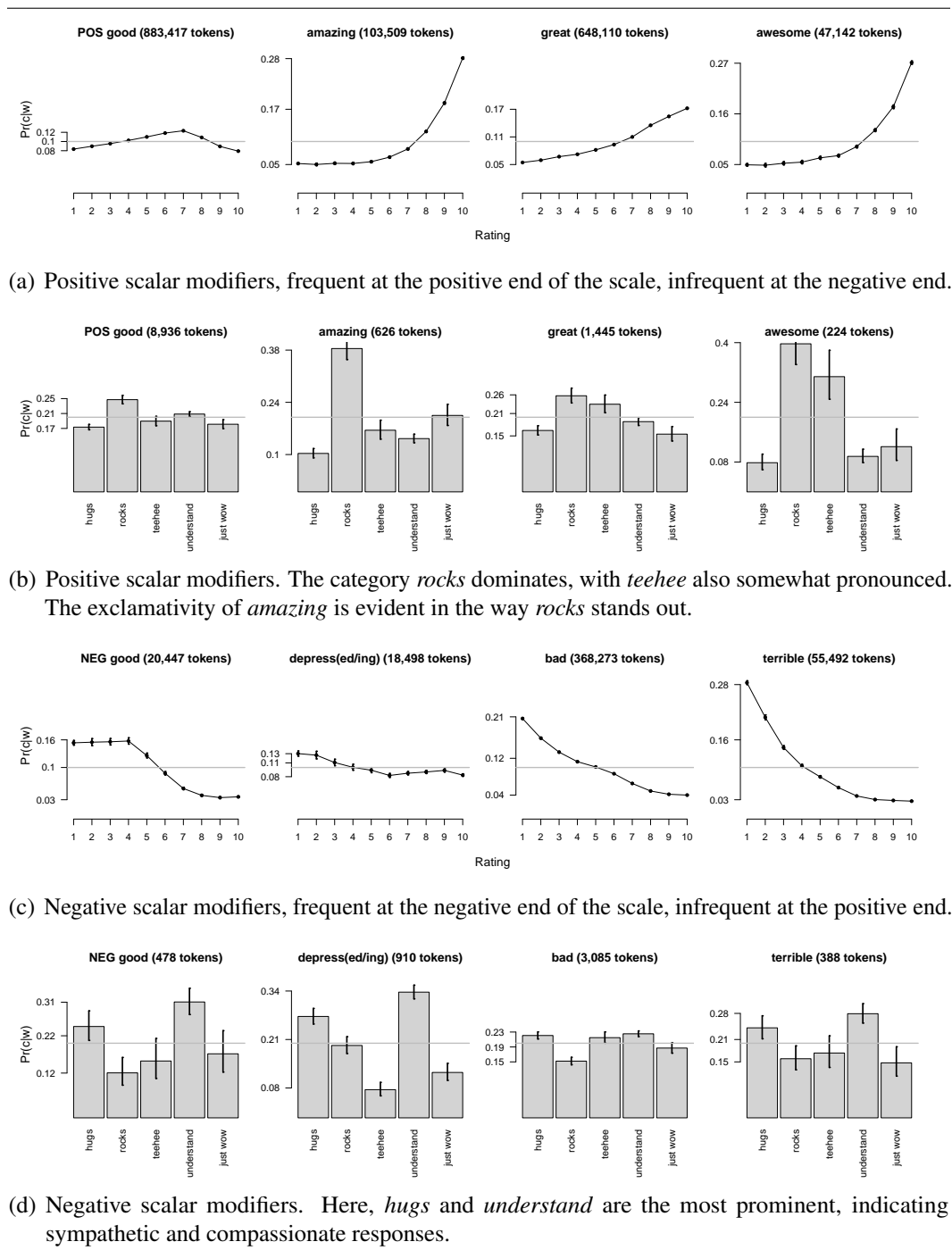


Figure 3 Scalar modifiers.

4 Negation

Students learning to reason with logical negation must work to set aside their linguistic intuitions about words like *false* and *negative*. One truth-value is as desirable as another, and negation moves us between them with equal ease in both directions. Iterating negation might take you back and forth between values (classically), or to something else (non-classically), but it certainly doesn't amplify any negative vibes. There are undesirable possible worlds, but negation could as easily steer us away from them as towards them. And so forth; the negation of logic is non-judgmental.

What about natural language negation? Linguistic semantic theories tend to define negative morphemes using those same non-judgmental logical operators. This might lead us to expect negation to have no special pragmatic import; p and $\neg p$ each simply pick out propositions, after all, so there is, on the face of it, no particular reason to favor \neg when delivering unwelcome news. In terms of the corpus methods defined here, the prediction is clear. Complaints and commendations can be delivered with negation or without it, and negation itself should be equally likely to elicit a *rocks* reaction as an *understand* reaction (for example). The distribution of negation should hover around the null hypothesis line, independent of the categories and thus (roughly) equally likely in all of them.

This is not at all what we find, though. The panels in fig. 4 pool *not*, *n't*, *no*, *never*, and compounds formed with *no*, studying their distribution without associated polarity items (for reasons addressed in the next section). The left panel of fig. 4 shows the IMDB distribution. Negation patterns almost exactly like the mild negative scalar adjective *bad* in fig. 3(c): its probability peaks in the most negative reviews and steadily drops off on the way to the most positive ones. The similarity with *bad* holds also in the EP corpus, as one can see by comparing fig. 3(d) with the middle right panel of fig. 4. In both, *understand* is over-represented; both negative scalars and negation are more common in stories that elicit this reaction of solidarity than they are, for example, in stories that elicit *teehee*. The EP picture is not as clear as one might like, since *hugs*, the other sympathetic category, is not also elevated, but the general trend seems in line with the negative scalars of fig. 3(d), and the polarity data in the next section help confirm this impression.

To help further solidify this point, and push back against the hypothesis that this has something to do with the IMDB and EP, I've included two supplementary panels using new data. First, the middle left panel in fig. 4 depicts negation in a corpus of 1,094,219 reviews (63,278,065 words) drawn from Amazon (a wide variety of products), Goodreads (books), OpenTable (restaurants), and Tripadvisor (hotels). Each review in this corpus is associated with a rating on a scale of 1 to 5 stars, which makes it coarser than IMDB, but the patterns are generally the same. Second, the

On the negativity of negation

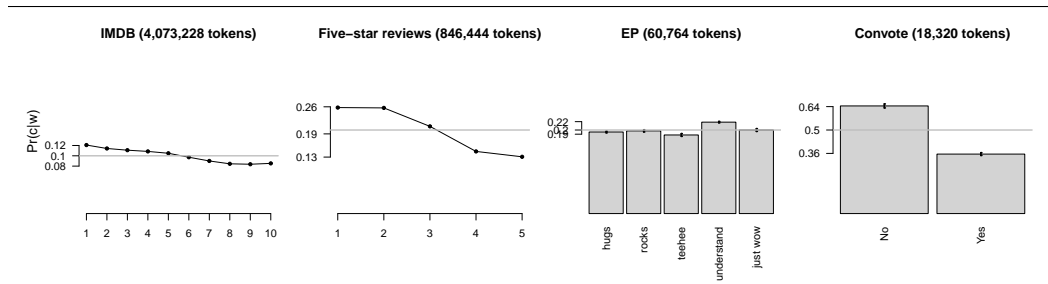


Figure 4 Negation without associated polarity items.

rightmost panel in fig. 4 uses data from the Convote corpus of congressional speeches (Thomas, Pang & Lee 2006). The corpus consists of selections from 8,121 speeches made by members of the U. S. Congress during the debate period prior to legislative votes. The categories here are ‘Yes’ if the speaker voted in favor of the bill, and ‘No’ if he voted against the bill. The panel depicts the distribution of negative morphemes across these two categories, with *no* left out of the calculations to avoid undue influence from speakers simply saying “Vote no” and the like. The correlation between negation and negativity is striking here, helping to further support the idea that negation is persistently negative no matter what the general context is like.

What are the underlying causal mechanisms of this correlation between contextual features and negation? If our goal is to reconcile the non-judgmental logical negation with the pragmatic negativity of negative morphemes, then we should look outside the grammar for an answer. Horn (1989: 164) reports on Russell’s (1948) attempt at an extra-linguistic explanation. Russell assumes that the pragmatically negative conception is prior: children learn to associate *no* with being rejected, chastised, or hurt. He then writes, “We may say that ‘not’ means something like ‘You do right to reject the belief that . . .’ . And ‘rejection’ means, primarily, a moment of aversion. A belief is an impulse towards some action, and the word ‘not’ inhibits this impulse”. While Horn grants that this approach has “some plausibility”, he is blunt in his general assessment: “Russell must execute a number of prodigious leaps of faith over the apparent holes in his argument”.

I think it is more fruitful to assume that the logical conception is prior and attempt to derive the pragmatic negativity from it. The logical conception is blind to the difference between negative and positive at the level of individual denotations, but asymmetries arise when we consider the two kinds of sentence relative to the context. Following Roberts (1996), I view contexts as structured by a set of abstract questions under discussion (QUDs), which jointly characterize the participants’ goals for the discourse and shape their behavior (see also Beaver & Clark 2008; Schoubye 2009; Simons et al. 2010). For simplicity, suppose each QUD is a partition on the common

ground (Groenendijk & Stokhof 1984). Negative and positive sentences are, by logical non-judgmentalism, in principle equally capable of resolving such QUDs, by identifying cells of these partitions. In practice, though, positive sentences tend to identify small numbers of cells (near resolution), whereas negative sentences tend to exclude small numbers of cells (far from resolution, since so many alternatives remain). This is not invariably the case — for example, interrogatives containing negation can reverse the informativity ordering — but it holds often enough to create a bias. This is the (defeasible) sense in which negative sentences are less informative than positive ones (Ducrot 1972; Anscombe & Ducrot 1983; Horn 1989: 60): they tend to be less resolving.

How does ‘less resolving relative to the QUDs’ become ‘pragmatically negative’? I think the seeds of this lie in a suggestion of Groenendijk & Stokhof (1984: 571), citing Wittgenstein (1953), that assertions in effect raise and resolve their corresponding polar questions, which are typically addressed as subquestions of more overarching ones. Thus, consider a discourse in which the immediate QUD is a polar question $?p = \{q_A, q_B\}$, where q_A and q_B partition (the salient region of) logical space. Both positive and negative sentences can resolve $?p$, but the positive ones will tend to be more resolving with regard to more general QUDs, so they are favored. Suppose, then, that a positive form S is uttered. If another speaker feels compelled to react to S (say, because it denotes an incorrect resolution of a QUD), then she is likely to use a negated form $\neg S$. Such disagreements are often pragmatically negative. Thus, over time, a hearer expectation develops: negation is used in case of a clash, which leads speakers to avoid it when they wish to avoid appearing to clash, which further strengthens the hearer expectations, and so forth, in a feedback loop that enhances the negativity of negation.

The reactive disposition of negation is arguably behind the corpus patterns seen above. In the case of the product reviews and the Convote data, the presumption is that the objects under discussion should possess certain properties, and the negative reviewers are compelled to counter that assumption, using negation to enhance the feeling that they are being reactive. In the case of the EP data, it is much less clear what is being evaluated, but my impression based on reading samples is that the confessions eliciting predominantly *understand* and *hugs* reactions are about missed opportunities, failed attempts, and other situations in which negation’s reactivity can enhance the author’s intended message.

QUD-based theories of the context focus on multi-speaker interactions, so ideally we would test these hypotheses about negation against data involving dialogues in which the QUDs are explicitly characterized and annotated. As far as I know, there is no such data set in existence; major research issues have to be resolved before we even know what it would mean to annotate a corpus with QUDs. However, the Switchboard Dialog Act Corpus (Jurafsky et al. 1997) provides annotations that get

fairly close. The Dialog Act Corpus consists of over 200,000 utterances from the Switchboard Corpus (Godfrey & Holliman 1993) annotated with information about the nature of the discourse move. The tag set is extensive. Here, I focus on just the annotations ‘ar’ (reject), ‘arp’ (partial reject), and ‘aa’ (accept). The coders manual (Jurafsky et al. 1997) says that these tags, “mark the degree to which speaker accepts some previous proposal, plan, opinion, or statement”. Thus, they are fundamentally reactive. The current hypotheses about negation predict that negation will be more frequent than normal in utterances marked with the reject tags (‘ar’, ‘arp’), and less frequent than normal in utterances marked with the accept tag (‘aa’).

To test these predictions, I first removed from consideration all utterances containing *no* as a free-standing morpheme, since many of these have conventionalized reactive uses, as in responses to polar questions. This left 219,556 utterances. Of the resulting set of utterances, 28,757 (13.1%) contain negation. This provides a baseline frequency for negation. In the restricted set, 133 utterances are tagged with ‘ar’ or ‘arp’, and 79 (59.4%) of them contain negation — well above the baseline frequency. Conversely, 9,581 utterances are tagged with ‘aa’ (accept), and just 169 (1.8%) of those contain negation — well below the baseline frequency. Tab. 7 summarizes these findings, which are statistically significant ($\chi^2 = 1728.41$; $df = 1$; $p < 0.001$).³

One might worry that this result traces not to linguistic intuitions but rather to the coders manual. Perhaps the instructions biased the annotators to chose ‘ar’ or ‘arp’ whenever they saw a negation in a reactive statement. (I thank Florian Schwarz for raising this issue, p.c.) However, the coders manual explicitly advises against moving too quickly from the presence of negation to one of the reject tags. It says, “A negative response to a question, statement or proposal is not necessarily a ‘reject’. If the previous statement is phrased in the negative, a ‘no’ could be an agreement [...]” (§6.1). Thus, the bias for negation likely reflects genuine linguistic intuitions about its use conditions.

	Contain negation	Lack negation	Total
Tagged (partial) reject	79 (59.4%)	54 (40.6%)	133
Tagged agree/accept	169 (1.8%)	9,412 (98.2%)	9,581

Table 7 Reject and accept utterances in the Dialog Act Corpus, excluding free-standing forms of *no*, which have conventionalized rejection uses.

³ If *no* is treated as a regular negation, then the result just becomes stronger, with 83.3% of ‘ar(p)’ cases containing negation and 6.9% of ‘aa’ cases containing negation.

5 Negative polarity items

Above, I was careful to look at negation without associated NPIs, to try to get a look at the pragmatic signals of negation alone. This section studies the effects that NPIs have on those signals. The approach is inspired by Israel's (1996; 2001; 2004) distinction between emphatic and attenuating polarity items. (The categories further subdivide into positive and negative; I set the positive ones aside.) Some examples:

Emphatic: *any, ever, at all, whatsoever, give a damn*
 Attenuating: *much, overmuch, long, all that, infinitival need*

These examples are meant to be representative of much larger classes. I've chosen to focus on them because they are all relatively frequent in the IMDB and EP corpora, so it is precisely these that I use in the corpus investigations below.

Israel (2004: 717) sums up the pragmatic effects that these expressions can have:

The pragmatic functions which polarity items encode, emphasis and attenuation, reflect two antithetical ways in which scalar semantics may be deployed for rhetorical effect: emphatic expressions serve to mark commitment or emotional involvement in a communicative exchange, while attenuation both protects a speaker's credibility and shows deference to a hearer by minimizing any demands on his credulity.

The best-studied emphatic NPIs are *any* and *ever*, and Israel's characterization hews closely to the proposal of Kadmon & Landman (1993) that these items are domain wideners. Under negation, domain widening corresponds to strengthening, so including such items results in a stronger statement. I can say "I won't eat spinach" with the intention of conveying an absolute restriction on my dining, but I can still strengthen this claim pragmatically with NPIs: "I won't eat any spinach" is stronger, and "I won't eat any spinach at all" is stronger still. The effect is additive in a manner similar to that of *big, big boat*.

This makes a prediction for the corpus data: emphatic NPIs should enhance the negativity of their licensing negations, by emphasizing an already generally negative discourse move. And this is precisely what we find. Fig. 5(a) depicts the effects of emphatic NPIs. The EP data were once again collected using the parsed forms and Tregex expressions, whereas the IMDB and other review data were obtained by chunking the texts heuristically into clauses and then finding co-occurring negation and NPIs in those clauses. For both IMDB and EP, the picture is the same as that of fig. 4, but sharper: the disparity between the negative and positive categories is greater for IMDB and the five-star reviews, and the *hugs* and *understand* categories

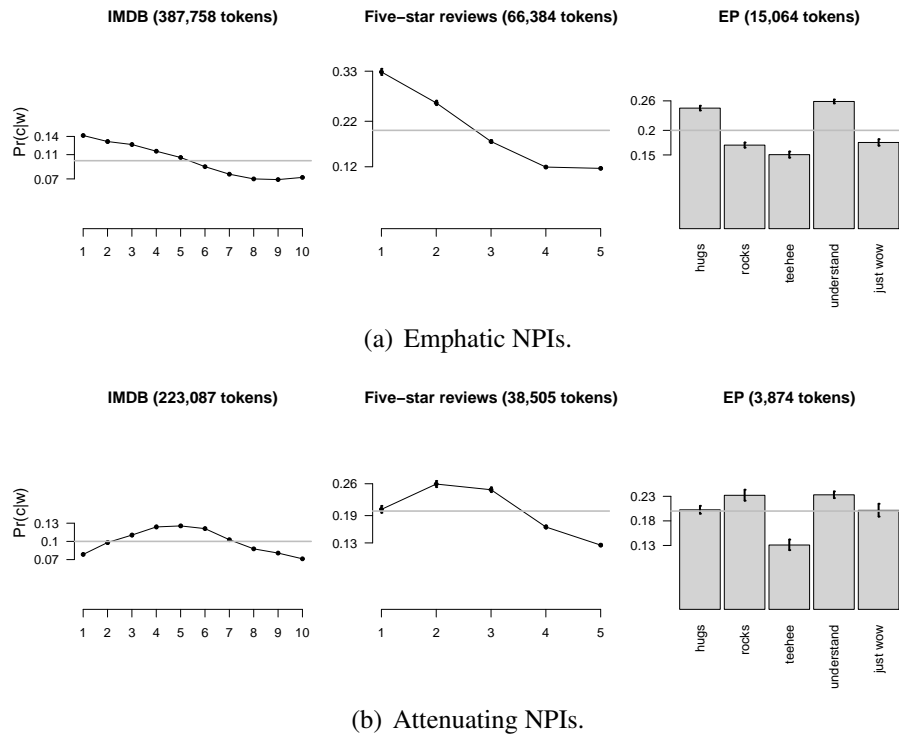


Figure 5 Negation with emphatic and attenuating polarity items.

are even more prominent in EP. This effect makes sense in terms of the QUD-based approach to explaining the pragmatic negativity: emphatic NPIs are likely to create a situation in which the speaker reacts to S with a sentence whose content strictly entails that of $\neg S$.

Israel’s predictions for attenuating NPIs are also born out. Attenuating NPIs soften, which makes them less confrontational. In a general sense, they resemble the sort of politeness markers that weaken in order to allow the speaker to save face (Brown & Levinson 1987; Sawada 2010). This is reflected in the corpus data. In the presence of attenuating NPIs, negation looks more like the slightly positive scalar modifier *good*; compare fig. 5(b) with the leftmost panels in fig. 3(a) and fig. 3(b). Other words that typically have this shape⁴ are *somewhat* (\wedge), *but* (\wedge), and *quite* (\wedge) — words that are used to make balanced, unimpassioned claims. For discussion of this class of items in corpora like this, see Potts & Schwarz 2010: §3.2.

In reading through the corpus texts, I became aware of a new (to me) use of negation. At least in some dialects of English, negation can be repeated in order to

⁴ The following small graphics are sparklines (Tuft 2006) representing the IMDB distributions.

strengthen the overall claim:

- (1) i do not not not like it when my connection goes all 1995ish on me
- (2) I am NOT NOT letting someone take out part of my liver!
- (3) I am NOT getting sick. Not not NOT.
- (4) Because never never never i saayyy!
- (5) Oh, no. No, no, no, no. [...]

I conjecture that this usage is related to the repetition of scalar modifiers, as in *big*, *big problem* and *sick*, *sick*, *sick*, where repetition does work similar to that of *very* and other scalar adverbials. I connect this with Israel's fundamentally scalar view of polarity items by supposing further that repeated negations like *never never never* involve a single negation (perhaps abstract; Ladusaw 1992) and then a series of emphatic polarity items. I note also that the rhetorical effect of the repeated negations is to enhance the speaker's overall commitment to the proposition expressed; the speaker of (1) means to say that he does not like it when his (Internet) connection "goes all 1995ish" (becomes slow), and the repeated negations emphasize this assertive act, imbuing it with a new level of emotionality. This is a kind of speech-act-oriented use of these negative morphemes, perhaps akin to the intensive uses of *totally* and stressed *SO*, as in *I am totally/SO seeing that movie*.

6 Emergent expressivity

This paper is about uncovering the ways in which negation is pragmatically negative, but it's also about introducing new data and methods. In sec. 2, I presented two large corpora. Their power for pragmatic research lies in their combinations of language and contextual meta-data. In the analyses, secs. 3–5, I focused on correlations between language and a few pieces of evaluative meta-data: ratings and reader reactions. This just hints at the possibilities; the texts on those sites also come with usernames, demographics, dates, summaries, comments from other users, reader estimates of helpfulness, and subcategorizations. There is a surprising amount of contextual information simply attached to these texts.

Linguistic pragmatics has, to date, focused on what is possible and impossible — the range of potential implicatures, the (non-)optionality of certain presuppositions, the options for anaphora resolution, and so forth. Questions about what *actually happens* in language use have been much less central, though such questions are fundamental to understanding human talk exchanges. Corpora of the sort explored here can help us understand what is possible, but their great strength is that they can help us with these questions of what *actually happens*. Negation and negative polarity illustrate this. Negation is not invariably negative, but the corpus findings

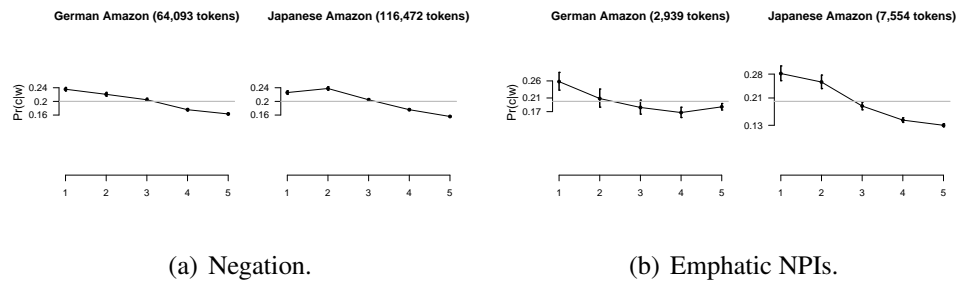


Figure 6 Negation and emphatic negative polarity in German and Japanese.

indicate that it tends to be, across a variety of different contexts. Similarly, emphatic and attenuating NPIs do not invariably interact with this pragmatic content, but they do so reliably enough to shape hearer expectations about speaker intentions.

The result is an emergent expressivity for negation and negative polarity (Ladusaw 2009; Giannakidou & Yoon 2010). The expressivity is arguably not encoded anywhere in the lexical entries involved but nonetheless reliably perceived by hearers and, in turn, intended by speakers. Indeed, encoding it in the lexical entries would seem to predict that this is potentially a point of cross-linguistic variation. Preliminary investigations suggest that this is not so. Fig. 6 summarize negation and emphatic polarity data for German and Japanese. The patterns are broadly the same throughout, suggesting that the negative bias is not specific to English. For similar cross-linguistic correspondences, see Constant et al. 2009; Potts & Schwarz 2010; Davis & Potts 2010. (Unfortunately, the attenuating NPIs known to me are too sparse in the German and Japanese data to yield a clear picture of how they work there. This is likely not a deficiency in the corpora, but rather in my understanding of polarity phenomena in those languages.)

Potts & Schwarz (2010) argue for similar cross-linguistic effects for demonstratives. Lakoff (1974) first identified a class of affective demonstratives (‘emotional deixis’), conveying solidarity and shared sentiment, and Potts & Schwarz (2010) find the effects of this affectivity in large corpora. There again, the expressivity arguably does not derive solely from the lexical meanings of demonstratives (though they contribute, to be sure), but rather also from a complex set of pragmatic factors including markedness (relative to the rest of the determiner paradigm), metaphorical extension, and optionality. The present argument for negation charts a similar path: the logical underpinnings of negation contribute to its pragmatic negativity, but they do not tell the whole story. At a certain point, hearer expectations and the effects they have on speaker choices ensure their own stability.

References

- Anscombe, Jean-Claude & Oswald Ducrot. 1983. *L'argumentation dans la langue*. Bruxelles: Mardaga.
- Baayen, R. Harald., Douglas J. Davidson & Douglas M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59. 390–412.
- Beaver, David & Brady Zack Clark. 2008. *Sense and sensitivity: How focus determines meaning*. Oxford: Wiley-Blackwell.
- Brown, Penelope & Stephen C. Levinson. 1987. *Politeness: Some universals in language use*. Cambridge University Press.
- Constant, Noah, Christopher Davis, Christopher Potts & Florian Schwarz. 2009. The pragmatics of expressive content: Evidence from large corpora. *Sprache und Datenverarbeitung* 33(1–2). 5–21.
- Davis, Christopher & Christopher Potts. 2010. Affective demonstratives and the division of pragmatic labor. In Maria Aloni, Harald Bastiaanse, Tikitou de Jager & Katrin Schulz (eds.), *Proceedings of the 17th Amsterdam Colloquium*, Springer.
- Ducrot, Oswald. 1972. *Dire and ne pas dire*. Paris: Hermann.
- Frege, Gottlob. 1919. Die Verneinung. Eine logische Untersuchung. *Beiträge zur Philosophie des deutschen Idealismus* 1. 143–157. Reprinted in translation in Peter T. Geach and Max Black (eds.), *Translations from the Philosophical Writings of Gottlob Frege*, 1952, 117–135.
- Gelman, Andrew & Jennifer Hill. 2007. *Data analysis using regression and multi-level/hierarchical models*. Cambridge University Press.
- Giannakidou, Anastasia & Suwon Yoon. 2010. The subjective mode of comparison: Metalinguistic comparatives in Greek and Korean. To appear in *Natural Language and Linguistic Theory*.
- Ginzburg, Jonathan. 1995. Resolving questions, part I. *Linguistics and Philosophy* 18(5). 549–527.
- Godfrey, John J. & Ed Holliman. 1993. Switchboard-1 transcripts. Linguistic Data Consortium, Philadelphia.
- Groenendijk, Jeroen & Martin Stokhof. 1984. *Studies in the semantics of questions and the pragmatics of answers*: University of Amsterdam dissertation.
- Horn, Laurence R. 1989. *A natural history of negation*. University of Chicago Press. Reissued 2001 by CSLI.
- Israel, Michael. 1996. Polarity sensitivity as lexical semantics. *Linguistics and Philosophy* 19(6). 619–666.
- Israel, Michael. 2001. Minimizers, maximizers, and the rhetoric of scalar reasoning. *Journal of Semantics* 18(4). 297–331.
- Israel, Michael. 2004. The pragmatics of polarity. In Laurence Horn & Gregory

- Ward (eds.), *The handbook of pragmatics*, 701–723. Oxford: Blackwell.
- Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59(4). 434–446.
- Jurafsky, Daniel, Elizabeth Shriberg & Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual, draft 13. Tech. Rep. 97-02 University of Colorado, Boulder Institute of Cognitive Science Boulder, CO.
- Kadmon, Nirit & Fred Landman. 1993. Any. *Linguistics and Philosophy* 16(4). 353–422.
- Klein, Dan & Christopher D. Manning. 2003a. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st annual meeting of the Association for Computational Linguistics*, vol. 1, 423–430. ACL.
- Klein, Dan & Christopher D. Manning. 2003b. Fast exact inference with a factored model for natural language parsing. In Suzanna Becker, Sebastian Thrun & Klaus Obermayer (eds.), *Advances in neural information processing systems 15*, 3–10. Cambridge, MA: MIT Press.
- Ladusaw, William A. 1992. Expressing negation. In Chris Barker & David Dowty (eds.), *Proceedings of Semantics and Linguistic Theory 2*, 237–259. Columbus, OH: OSU Working Papers in Linguistics.
- Ladusaw, William A. 2009. Still puzzled why there are polarity items. Talk delivered at the 35th Annual Meeting of the Berkeley Linguistics Society.
- Lakoff, Robin. 1974. Remarks on ‘this’ and ‘that’. In *Proceedings of the Chicago Linguistics Society 10*, 345–356.
- Levy, Roger & Galen Andrew. 2006. Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the 5th edition of the International Conference on Language Resources and Evaluation*, 2231–2234.
- de Marneffe, Marie, Christopher D. Manning & Christopher Potts. 2010. “Was it good? It was provocative.” Learning adjective scales from review corpora and the Web. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics*, 167–176. ACL.
- Munro, Robert, Steven Bethard, Victor Kuperman, Robin Melnick, Christopher Potts, Tyler Schnoebelen & Harry Tily. 2010. Crowdsourcing and language studies: The new generation of linguistic data. In *Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Mechanical Turk*, 122–130. ACL.
- Pang, Bo & Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval* 2(1). 1–135.
- Potts, Christopher & Florian Schwarz. 2010. Affective ‘this’. *Linguistic Issues in Language Technology* 3(5). 1–30.

- R Development Core Team. 2010. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing Vienna, Austria. <http://www.R-project.org>.
- Roberts, Craige. 1996. Information structure: Towards an integrated formal theory of pragmatics. In Jae Hak Yoon & Andreas Kathol (eds.), *OSU working papers in linguistics*, vol. 49: Papers in Semantics, 91–136. Columbus, OH: The Ohio State University Department of Linguistics. Revised 1998.
- Russell, Bertrand. 1948. *Human knowledge, its scope and limits*. New York: Simon and Schuster.
- Sawada, Osamu. 2010. *Pragmatic aspects of scalar modifiers*: University of Chicago dissertation.
- Schoubye, Anders. 2009. Descriptions, truth value intuitions, and questions. *Linguistics and Philosophy* 32(6). 583–617.
- Simons, Mandy, Judith Tonhauser, David Beaver & Craige Roberts. 2010. What projects and why. In David Lutz & Nan Li (eds.), *Proceedings of Semantics and Linguistic Theory 20*, Ithaca, NY: CLC Publications.
- Snow, Rion, Brendan O’Connor, Daniel Jurafsky & Andrew Y. Ng. 2008. Cheap and fast — but is it good? Evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, 254–263. ACL.
- Thomas, Matt, Bo Pang & Lillian Lee. 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. In *Proceedings of EMNLP*, 327–335.
- Tufte, Edward R. 2001. *The visual display of quantitative information*. Cheshire, CT: Graphics Press 2nd edn.
- Tufte, Edward R. 2006. *Beautiful evidence*. Cheshire, CT: Graphics Press.
- Wittgenstein, Ludwig. 1953. *Philosophical investigations*. New York: The MacMillan Company. Translated by G. E. M. Anscombe.

Christopher Potts
Department of Linguistics, Building 460
Stanford University
Stanford CA 94305
cgpotts@stanford.edu