Exclamatives and heightened emotion: Extracting pragmatic generalizations from large corpora

Christopher Potts and Florian Schwarz
UMass Amherst

September 7, 2008

Abstract

Exclamatives like *What a dump!*, *Wow!*, and *Boy, you've grown!* are, when uttered in context, rich in information about the speaker's attitudes. Drawing on evidence from about 100,000 online product reviews with associated meta-data, we develop a frequency-based characterization of this pragmatic contribution. This allows us to make precise predictions about the *exclamativity* that inheres in these constructions. In addition, we build logistic regression models and use the resulting statistics to state general, corpus- and language-independent hypotheses about what it means to be an exclamative pragmatically. These hypotheses allow us to identify previously unnoticed exclamatives, and they highlight the importance of purely expressive meanings.

Keywords corpus pragmatics, exclamatives, expressives, logistic regression

1 Corpus pragmatics

The phenomena of linguistic pragmatics involve complex interactions among a variety of factors, most notably (i) sentence meanings, (ii) the discourse participants' understanding of the utterance context, and (iii) general principles of rational communication. While considerable progress has been made in annotating corpora with highly nuanced semantic representations (Palmer et al. 2005; Cahill et al. 2007; Pradhan et al. 2007), it is, at present, not even clear how to adequately represent (ii) or (iii). Thus, the very nature of these problems might seem to place them outside the bounds of present-day corpus methods.

Nonetheless, the past decade has seen a flowering of work in corpus pragmatics. Models and data sets have been developed for a variety of pragmatic phenomena, including discourse

structuring (Webber 2006; Prasad et al. 2008), sentiment (Nigam et al. 1998; Pang et al. 2002; Pang and Lee 2004, 2005; Beineke et al. 2004), speaker attitudes (Shanahan et al. 2005), coreference (Webber et al. 2003; Culotta et al. 2007), and emotionality (Liberman 2002; Liscombe et al. 2003). The lesson is a familiar one: where exact methods are infeasible, approximate methods can still yield deep results.

Unfortunately, these approaches have not really made their way into our own field of theoretical linguistic pragmatics (though see Jurafsky 2004 and references therein). There is no intellectual reason for this divide. As theoretical pragmaticists, we too deal exclusively in approximations of (i)–(iii) when designing test scenarios, and, in recent years, conducting experiments with human subjects (Noveck and Sperber 2004; Sedivy 2007; Grodner and Sedivy 2008; Schwarz et al. 2008; Clifton et al. 2008). Thus, corpora should prove useful in addressing the problems of linguistic pragmatics, even if they only approximate the complexity of the discourse situations involved.

The present paper reports on one effort to bring corpus evidence to bear on questions in lexical and constructional pragmatics. *Exclamatives* like those in (1) are our empirical focus.

(1) a. What a hotel!

b. A must read!!

c. Wow, . . .

We study the distribution of a variety of exclamatives in online collections of product reviews. Three central results emerge from this study. First, we obtain a quantitative perspective on the heightened emotion that attaches to examples like (1). Second, we are able to estimate how reliable exclamatives are at conveying emotionality. Third, the corpus evidence allows us to build logistic regression models that support a mathematical characterization of exclamative content, one that holds across diverse data sets and, we suspect, across languages.

In the next section, we introduce exclamatives more fully and highlight some of their important linguistic properties, concentrating on *use conditions*. Section 3 describes the data collections we use in this paper. With this background in place, we turn, in section 4, to the distribution of exclamatives in naturally occurring texts, using this information to flesh out the qualitative theoretical picture with quantitative results. Section 5 puts those quantitative results to work in motivating a pragmatics of exclamativity that is based on the interactions of speaker and hearer expectations. In section 6, we develop a method for identifying exclamatives in context using logistic regression. We close, in section 7, by describing our ongoing efforts to extend the results into new theoretical and experimental areas.

2 Exclamatives and exclamativity

The usual basic perspective of linguistic meaning studies is interpretive: we ask what the morphemes, phrases, and constructions of language mean. However, *use conditions* can be equally important to understanding linguistic communication. Kaplan (1999) writes:

> When I think about my own understanding of the words and phrases of my native language, I find that in some cases I am inclined to say that I know what they *mean*, and in other cases it seems more natural to say that I know how to *use* them.

Here, Kaplan takes the speaker's perspective on use conditions: much of knowing how to use an expression E is being able to identify the conditions under which E is appropriately used ('expressively correct', in Kaplan's terms). One can adopt the hearer's perspective as well: utterance understanding depends on one's ability to extract, from the speaker's language, certain information about the context of utterance.

Languages do not draw a divide between items that are characterized by their meanings and those that are characterized by their use conditions. Invariably, we call upon both perspectives. Exclamatives, illustrated in (2), provide especially rich examples of how meanings and use conditions can play off each other.

(2) a. What big eyes you have!

b. How you've grown!

c. Is he ever fast!

In each case, we discern a regular propositional component of the overall message.[1] Example (2a) conveys that the addressee has big eyes. Example (2b) says roughly that the addressee has grown a lot. And so forth. Though it can be challenging to say just what the meanings are and how they are derived systematically (Zanuttini and Portner 2003; Castroviejo Miró 2006; Rett 2008), it is clear that, at some level, exclamatives can be evaluated for truth.

Exclamative meanings do not end with this truth-conditional component, though. While (2a) shares something with *You have big eyes*, the two are not to be equated. The exclamative encodes excitement, or surprise, or enthusiasm. Let's call this extra meaning *exclamativity*. A major research question in the study of exclamatives (and other clause types) is just how

[1] Ginzburg and Sag (2001:§3.2.3) argue that exclamatives denote *facts*, rather than propositions. In the context of their ontology, this divorces the semantics of exclamatives from that of declaratives and yields a different perspective on the sense in which exclamatives have a truth-conditional component.

to characterize exclamativity. In sections 4 and 6, we use our corpus evidence to address this issue.

The examples above are canonical English exclamatives. They have roughly the same morphosyntax as embedded interrogatives, an affinity that is well attested cross-linguistically (Ginzburg and Sag 2001; Zanuttini and Portner 2003; Castroviejo Miró 2006). They can, though, be disambiguated from interrogatives with intonational cues (partially reflected in the exclamation point), with a variety of sentence-initial particles — *Boy*, *My*, *Oh* — that are incompatible with interrogatives, and by the fact that they appear routinely as matrix clauses. In addition, not all interrogatives correspond to exclamatives. Systematically excluded from exclamative readings are interrogative phrases that cannot have a degree-based semantics (Rett 2008). For example, none of the examples in (3) are grammatical exclamatives, though minor changes can fix them (e.g., changing *who* to *what people* in (3b)).

(3) a. *Which size eyes you have!

b. *Who you've met in your time!

c. *When we visited Rome!

Canonical interrogative-style exclamatives are not the only vehicles for exclamativity. A variety of clause-types can receive exclamative readings in context. In writing, we can tack strings of exclamation points onto the end of declaratives to imbue them with exclamativity, and certain adverbials — e.g., *absolutely*, *totally* — can also layer an exclamative semantics atop a declarative foundation:

(4) a. You have big eyes!!

b. It was an absolutely wonderful stay!

c. I am totally fed up with my computer!

Simple nominal phrases can be used in an exclamative fashion, as in (5), and we have a variety of particles whose sole function is to convey amazement, excitement, and so forth — little packages of pure exclamativity, as in (6).

(5) a. The bus! (as it finally appears in the distance)

b. Your bag! (as you leave the room without it)

(6) a. Wow!

b. Whew!

Even this small sample makes it evident that there is no single meaning associated with exclamatives. Each of these constructions is uniquely exclamative. One important dimension along which exclamatives differ is what we might call the 'polarity' of the emotion: whether it is positive or negative. It is typically hard to determine which direction a given exclamative tends to lean. For example, the polarity of canonical exclamatives like *What a PRED* seems to be determined largely by the nature of the lexical items involved. Thus, the examples in (7) are positive, whereas those in (8) are negative.

(7) a. What a wonderful view!

b. What a pleasure!

(8) a. What a dump!

b. What a disappointment!

This suggests that these exclamatives do not encode polarity as part of their lexical meaning; their exclamativity seems to be more like generalized heightened emotion. However, a bias can emerge where the lexical content is not inherently evaluative:

(9) a. What a view!

b. What a hotel!

c. What a read!

Thus, in some sense, these exclamatives default to positive meanings. In contrast, declaratives used exclamatively seem, in and of themselves, unbiased; if one hears,"This movie is by Peckinpah!", one might feel unsure as to the intended polarity of the exclamativity. Our corpus evidence is ideal for teasing out subtle general trends like this, so we will not attempt to be precise about these biases (or lack thereof) just yet. For now, suffice it to say that this is a component of exclamativity that we would like to better understand.

There are excellent reasons to think that canonical exclamatives like (2) represent a linguistically coherent clause-type, and thus many studies rightly set aside the more heterogeneous group suggested by (4)–(6), perhaps even separating out the elliptical versions in (7)–(9). However, our goal is to better understand exclamativity itself. We seek to determine just what kind of signal it is, how reliably speakers signal it, and how reliably hearers apprehend it. Thus, we consider the full range of constructions represented by (2) and (4)–(9), and we use the naturally occurring data to expand the class even further, uncovering previously overlooked pockets of exclamativity and defining a method for automatically extracting exclamatives from a body of labeled data. The next section introduces our data sets, and then sections 4, 5, and 6 apply them to the linguistic expression of exclamativity.

3 Data sets

Our investigations are based on data gathered from online book reviews from Amazon.com and online hotel reviews from Tripadvisor.com. Each review comes with a great deal of meta-data, including information about the author, other reviewers' reactions to the review, a summary statement, and a rating of the product in question from one to five stars. In this paper, we deal exclusively with the review text, the short summary text, and the associated product rating. Figure 1 provides two representative examples.

Amazon.com

Summary:	Phenomenal!
Rating:	5 of 5 stars
Review:	This book is absolutely earth shatteringly outstanding. It is so funny, clever and the details make your heart sing. It is easy to understand, entertaining and totally relevant. You have to read it!

Tripadvisor.com

Summary:	Bring Ear Plugs
Rating:	1 of 5 stars
Review:	From 12:30 AM to 3:30 AM, there was a disturbance outside from patrons of a nearby nightclub so loud it woke everyone in my room up. I'm usually a heavy sleeper – it takes a lot to wake me up. When I called the front desk around 2:00 AM to ask them to do something about the noise, they said this was a usual occurrence on weekend nights and they would have to call the police to get the crowd to go home. [. . .]

Figure 1: Example reviews with ratings

For our experiments, we extracted meta-data and distributional information from each of these websites. From Amazon.com, the information came from about $53,000$ reviews of 347 books, for a total of about 8.1 million words in the reviews and about $277,000$ words in the summaries. From Tripadvisor.com, we looked at about $55,000$ reviews, of 500 hotels, for a total of about 8.7 million words in the reviews and $256,000$ words in the summary field. The two different sites and the two 'review' and 'summary' fields from each site give us four data collections:

(10) a. Amazon summary: Summaries of book reviews

b. Amazon review: Reviews of books

c. Tripadvisor summary: Summaries of hotel reviews

d. Tripadvisor review: Reviews of hotels

	R_1	R_2
word$_1$	20	29
word$_2$	5	54
word$_3$	10	10
⋮		

(a) Table of word counts for two rating categories.

Tripadvisor summary		**Amazon summary**	
Category pair	χ^2 score	Category pair	χ^2 score
3, 5	22823.933	*3, 5*	15976.4233
2, 5	19000.268	2, 5	10273.0141
4, 5	13745.964	1, 5	9442.1387
1, 5	12853.315	**4, 5**	8372.2496
1, 4	11658.816	1, 4	4149.5776
2, 4	10270.770	*1, 3*	2924.8805
3, 4	8315.794	*2, 4*	2673.7162
1, 3	6403.132	**3, 4**	2573.4935
1, 2	2855.584	**1, 2**	1028.4324
2, 3	2070.025	**2, 3**	983.8994

(b) Pairs with a rating difference of 1 are in bold, and those with a rating difference of 2 are italicized. The lower the χ^2 score, the more similar the categories being compared.

Figure 2: Estimating the similarity of pairs of rating categories using the χ^2 statistical test.

To each of the documents in each of these collections, a star-rating is attached. Appendix A is a more extensive overview of the quantitative aspects of the data.

If one reads through reviews at these sites, one quickly develops intuitions about what the reviews are like. Authors adopt a wide variety of registers, from very formal to very colloquial. The texts are generally fairly short: the average summary, across the various corpora, is about 6 words, and the average review, again across the corpora, is about 200 words.

How, and to what extent, do reviews from the various rating categories resemble each other? Or, to take a different perspective, if we asked someone to guess, based on the words occurring in the text alone, what the author's assigned rating was, how successful would we expect the guesses to be? To address these questions, we employ the χ^2-based measure of corpus similarity developed by Kilgarriff and Rose (1998) (and reported on by Manning and Schütze (1999:171)). Given a pair of rating categories R_1 and R_2, we construct a table of the form in figure 2(a), in which the rows are labeled with words and the cells are filled with token counts. We then calculate the χ^2 statistic to obtain a measure of how similar the ratios of counts are between the two columns. The lower the χ^2 score, the more similar the categories being compared are. The results of running this test on the Tripadvisor summary collection and the Amazon summary collection are given in figure 2(b). We've ordered the rows according to the size of the χ^2 statistic to bring out the pattern. Pairs with a rating

difference of 1 are in bold, and those with a rating difference of 2 are italicized.[2]

Adjacent category pairs — e.g., (2, 3), (1, 2), (3, 4) — are fairly consistently the most similar, with pairs like (1, 5) and (2, 5) much less similar. The pairs (3, 5) and (4, 5) are exceptions to this pattern, but the general trend is evident. All our corpora deliver approximately this ordering, and we obtain similar results by measuring the KL-divergence between pairs of rating categories (conceived of as probability distributions; Cover and Thomas 1991:§2.3; Manning and Schütze 1999:§2.2.5). These results suggest that a person guessing ratings by looking at the words occurring in the texts alone might be expected to be in the correct rating area — low, middle, or high — but that it would be easier to confuse a two-star review with a three-star review than it would be to confuse a one-star review with a five-star review.

However, there are linguistic features that run counter to these measures. One important thing that these similarity measures fail to capture is a dimension of similarity between extreme reviews, namely, the emotionality inherent in one-star and five-star reviews. People who have had a bad hotel stay are fairly unconstrained in expressing their dissatisfaction, and people newly back from wonderful vacations are effusive. It turns out that these two extreme emotions are reflected in the use of strikingly similar linguistic means. Exclamatives abound in such reviews. Generalized emotionality is the driving force.

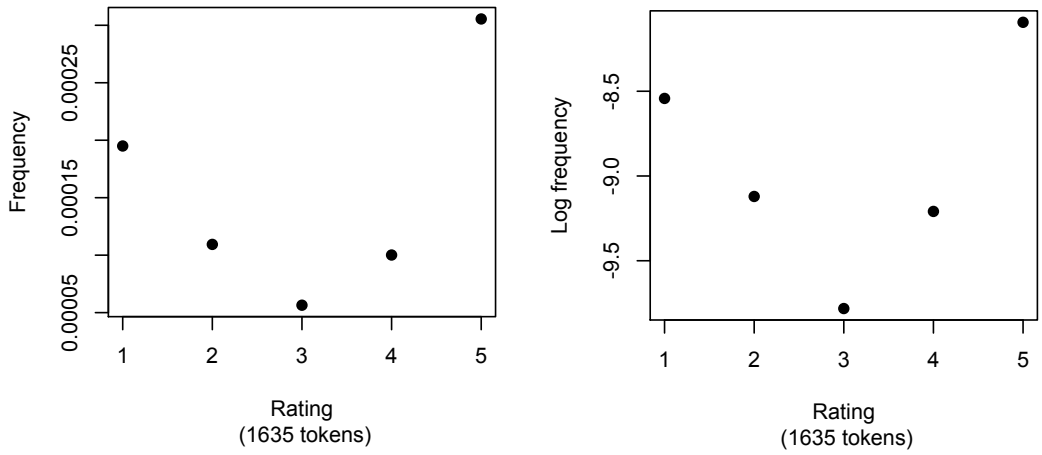
4 The corpus evidence

As noted above, the reviews in our collections (10) contain a lot of emotional language, including large numbers of the exclamatives discussed in section 2 (as well as many others that fit the mould). This abundance of data means that we can get reliable distributional information across the different star-rating categories and across corpora. The goal of this section is to define a method for gathering such distributional information. We put this evidence to use in a pragmatic theory in section 5.

By way of introducing our investigative strategy, we open with a detailed discussion of the bigram (two-word sequence) *what a*. This bigram is an approximate identifier of canonical exclamatives like those in (7), (8), and (9) above. Though it picks up some spurious examples (e.g., *they asked us what a persimmon looks like*), the vast majority of such tokens are exclamatives.[3]

[2] These results are for the subset of the vocabulary with at least 200 tokens in each of the corpora.

[3] For example, in the Tripadvisor summary corpus, there are 455 tokens of *what a(n)*, only two of which are not exclamatives: *What an hotel should be* and *If this is what a 4 star hotel is about, I'll save my money!*. We note also that our corpora are large enough to tolerate a little noise: the four collection contain between 193 and 1,635 tokens of this phrase. What's more, any noise of this sort can only dampen the effects we are after, by introducing constructions that lack exclamativity. Our results, though, are robust.



(a) Frequencies, calculated as in (13)

(b) Log odds, calculated as in (15)

Figure 3: *what a* in the Tripadvisor review corpus

Figure 3 depicts two different perspectives on the frequency of *what a* in the Tripadvisor review data. In both, the rating categories are arrayed along the x-axis. The y-axis depicts the frequency of *what a* in those rating categories. Figure 3(a) gives the basic frequencies, derived directly from the empirical counts: for each category R, we divide the number of occurrences of *what a* by the total number of bigram tokens in texts in category R. In (11)–(13), we describe the general calculation for any word sequence x_n of length $n > 0$.

(11) $\text{count}(x_n, R) \stackrel{def}{=}$ the number of tokens of x_n in documents with rating R

(12) $\text{count}_n(R) \stackrel{def}{=}$ the number of tokens of word sequences of length n in documents in rating category R (i.e., $\sum_{x_n} \text{count}(x_n, R)$).

(13) $$\text{frequency}(x_n, R) \stackrel{def}{=} \frac{\text{count}(x_n, R)}{\text{count}_n(R)}$$

It is important to relativize to the number of tokens in each rating category. As the numbers in appendix A show, these categories are highly uneven in size, with the preponderance of reviews falling in the five-star category. This is a linguistically uninteresting fact about our corpora that we abstract away from with our approach to frequencies.

Figure 3(b) provides a log-odds perspective on the frequency data in 3(a), which facilitates a proper statistical analysis of the distributions of phrases. To move to this

perspective, we first shift from frequencies (probabilities) to odds, as in (14).[4] We then take the natural logarithm (ln) of the odds, shown in (15).

$$\text{odds}(x_n, R) \stackrel{def}{=} \frac{\text{count}(x_n, R)}{\text{count}_n(R) - \text{count}(x_n, R)} \tag{14}$$

$$\text{log-odds}(x_n, R) \stackrel{def}{=} \ln\big(\,\text{odds}(x_n, R)\big) \tag{15}$$

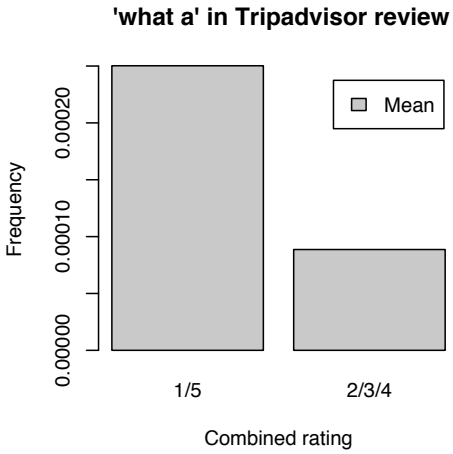
The log-odds perspective enables us to compare differences in frequency (or, more precisely, odds) across different orders of magnitude of frequency. It also corresponds to the approach underlying the appropriate statistical tool for our purposes, logistic regression (discussed in more detail in section 6). One way to illustrate the effect of taking the log-odds perspective is to compare the effect of a small difference in frequency (or odds) in the middle of the frequency scale and in the lower end. The overall effect of looking at log-odds is that differences at the extreme ends of the scale are enhanced relative to those at the more central part of the scale. Since all our odds-values are very small, this means that the differences between lower frequencies are enhanced in the log-odds perspective, relative to the regular frequency (or odds) perspective. For example, the difference in odds in the last two rows in (16) is identical, but in log-odds terms, the first is a lot smaller than the second, as shown in the last column. In log-odds terms, the difference between 0.0001 and 0.0002 corresponds to that between 0.3 and 0.6. This reflects the fact that in both cases, the second frequency is twice as large as the first.

(16) Distances between frequencies on the basic and log scales (abstract)

Odds A	Odds B	Log-odds A	Log-odds B	Log-odds A - Log-odds B
0.3	0.6	−1.203973	−0.5108256	−0.6931472
0.3001	0.3002	−1.203640	−1.203306	−0.0003332
0.0001	0.0002	9.21034	−8.517193	−0.6931472

Figures 3(a) and 3(b) provide a concrete example that brings out the same point. In the basic frequencies, the distance between the two-star frequency and the three-star frequency is clearly smaller than the difference between the one-star frequency and the two-star frequency. In the log-odds, however, the two distances are closer, and one can just barely make out that the 2-to-3 distance is larger here than the 1-to-2 distance. Here are the precise figures:

[4] Given the size of our rating categories and the overall quite small frequencies of the phrases we are looking at, the differences between frequency and odds is negligible, and we sometimes use the terms somewhat loosely as being more or less interchangeable.



(a) Occurrences of *what a* are about three times more frequent in the combined extreme rating categories than in the combined middle.

Category pair	Distance
1, 4	2.265626
1, 3	1.860744
4, 5	1.634529
1, 2	1.430694
3, 5	1.229647
2, 4	0.834932
2, 5	0.799597
1, 5	0.631097
2, 3	0.43005
3, 4	0.404882

(b) Log-odds differences between rating categories in Amazon summary, sorted from largest to smallest.

Figure 4: *what a* is a reliable indicator of an extreme rating, but it is not a reliable indicator of which extreme.

(17) Distances between frequencies on the basic and log scales (for figure 3)

	Frequency	Log-odds
1-star − 2-star	0.000085617	0.578332
2-star − 3-star	0.00005290513	0.661277

In (17), the important comparisons are column-wise: for basic frequencies, the 1-to-2 distance is larger. For log-odds, the 2-to-3 distance is larger. We see again that the differences in frequency are enhanced as we go lower on the frequency scale. The distributions have the same basic shape, but the log-odds 'stretch out' the lower part of the graph.

Stepping back from these technical details and returning to a more general perspective on the type of distribution of phrases like *what a*, it is the U-shaped nature of this distribution that we wish to highlight. An expression with such a distribution is more likely to fall into one of the extreme ends of the rating scale than it is to fall into the middle. Figure 4(a) highlights this bias by comparing the mean frequencies for two supercategories — extreme (1/5) and middle-of-the-road (2/3/4). Occurrences of *what a* are about three times more frequent in the combined extreme rating categories than in the combined middle categories. Thus, an exclamative is a solid indicator of which of these two supercategories a given exclamative falls into.

In contrast, the chances of a given token of *what a* falling into a one-star review tend to be about the same as the chances of it falling into a five-star review, so *what a* is not much help in signaling this distinction. The largest difference we measure between one-star and five-star is in the Amazon summary collection, and it is only about 0.63 in log-odds terms. This number doesn't mean much in isolation, but it is among the smallest spreads in the between-category distances in this corpus. Figure 4(b) provides the full set of such distances for the Amazon summary corpus. The pair $(1, 5)$ is third from the bottom. Thus, at the level of individual categories, *what a* is solid evidence for deciding between a one-star and a four-star guess. By comparison, it does not reliably indicate the difference between 1 and 5 (or 2 and 3, etc.).

In sum, there is a close correlation between the likelihood of *what a* appearing in a review and the rating category of the review, with an increase in the likelihood towards the extreme ends of the rating spectrum. As we will discuss in more detail in section 5, we can interpret this correlation in two directions, corresponding to the hearer's and the speaker's perspective. For the hearer, hearing *what a* is a good indicator that the speaker is in a heightened emotional state (associated with extreme reviews, either positive or negative). For the speaker, being in a heightened emotional state (again, either positive or negative) is a precondition (at least as a tendency) for using *what a*. Any item with a genuinely U-shaped distribution will have these properties.

We have so far mostly used the Tripadvisor review collection to study *what a*. The distributional facts are not peculiarities of that collection, though. Rather, the same patterns hold consistently across all four of the collections listed in (10). Figure 5 illustrates with log-odds distributions for *what a* in all four collections. What is striking about these figures, first and foremost, is their uniformity. In order to substantiate this visual impression, we have included quadratic logistic regression lines (in gray) for each distribution, along with their associated p-values. The p-values are uniformly significant, providing an initial statistical basis for the visual impression that the distributions are U-shaped. (Section 6 describes the reasoning behind this connection more fully.)

Linguistically, this distributional shape jibes with the discussion in section 2 above: canonical exclamatives express heightened emotion, without themselves indicating the direction of that emotion. A look at some specific examples from the corpora strengthens this overall impression:

(18) a. What a Find! (from a five-star Tripadvisor review)

b. What an awesome place!! (from a five-star Tripadvisor review)

c. What a mess! (from a one-star Tripadvisor review)

d. What an Overpriced Dump (from a one-star Tripadvisor review)

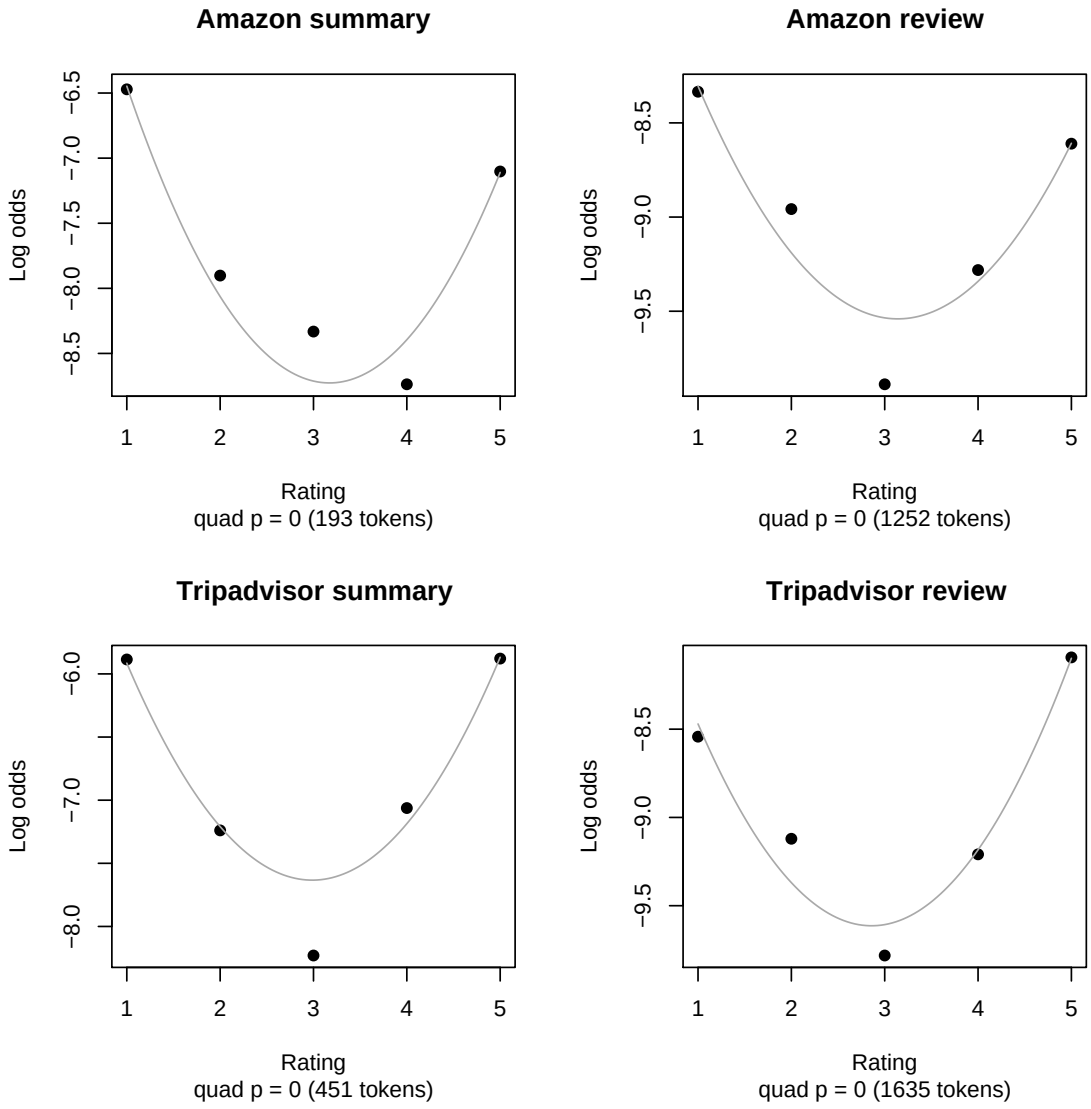


Figure 5: *what a*. The empirical log-odds are marked with black dots. The quadratic regression lines are gray.

As noted above, whether the sentiment is positive or negative is determined by the lexical content — which, in these examples, is the argument to the indefinite determiner. However, where the predicate is non-evaluative, the overall meaning is positive. For example, if we look at the specific string *what a hotel* in the Tripadvisor summary collection, we find a total of 9 occurrences: 2 in four-star reviews and 7 in five-star reviews. In the Tripadvisor review collection, there are 19 occurrences, 17 of them in the five-star review and 2 in the four-star reviews.

Many other words and phrases have this basic U-shaped profile, including *absolute*, *anyone*, and the bigram *I wish*; section 6 gives a complete accounting for our data sets. This is not to say, though, that all exclamatives are free from a bias for one direction or another on the scale. Figure 6 provides distributional information for *wow* using the same format as in figure 5. This particle also has a U-shaped distribution, but it seems more accurately described as J-shaped, suggesting a bias for positivity over negativity. Other items with this overall profile include *fabulous*, *absolutely*, *truly*, and sequences of exclamation points. We can also find Reverse-J distributions, suggesting a bias for negativity (without, though, excluding positive uses, if the context and the lexical content are right). These include a variety of negative forms along with negative polarity items and negative superlatives like *worst* and, perhaps surprisingly, *please*.

5 The pragmatics of exclamativity

The shapes identified in the previous section are, we claim, important sources of pragmatic information. In particular, they are a window into the nature of exclamativity. As is evident from the discussion in section 2, it is hard, perhaps impossible, to completely characterize, in familiar semantic terms, just what exclamativity is or what it adds to an utterance in context. This need not leave us silent, though. It just suggests that we need to approach the issue from a different empirical perspective. Our corpus data facilitate one such alternative.

In the opening of this paper, we noted that the context is a factor in determining pragmatic phenomena. Together with the semantic content of the uttered phrase and some general principles of rational communication, it determines the pragmatically enriched meanings that we perceive. In this paper, we treat the rating categories as approximations of one dimension of the context. One-star ratings identify contexts in which the speaker is feeling negatively about the things she is discussing. Three-star ratings indicate that the speaker has a relatively mixed reaction. And five-star ratings take us to the other extreme of speaker emotionality. We summarize these general assumptions in the hypothesis in (19).[5]

[5]We recognize, of course, that a speaker's emotional state might change in complex ways during the writing of a full review, even one of just 200 words. Nonetheless, our results for the review corpora largely

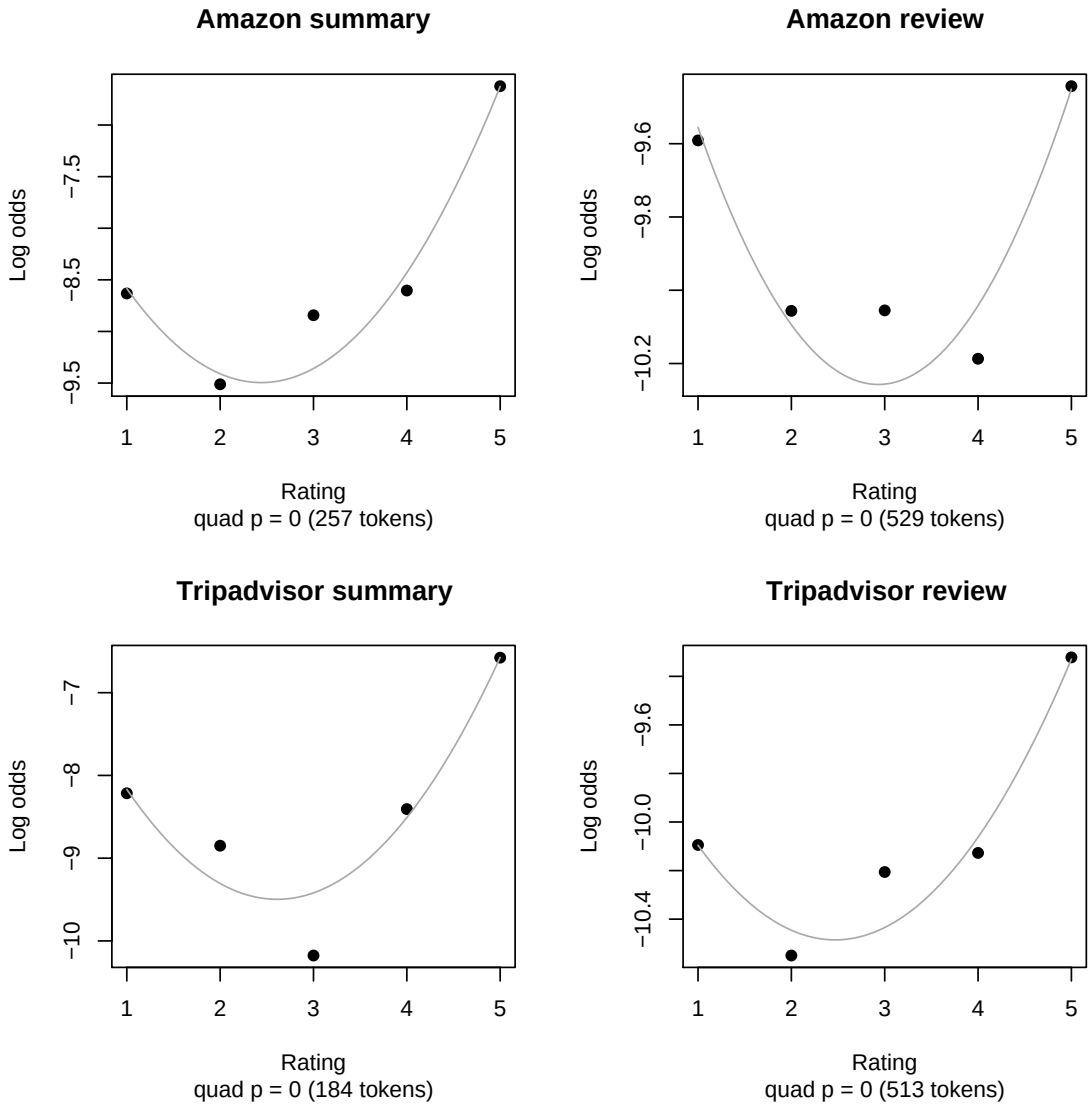


Figure 6: *wow*. The empirical log-odds are marked with black dots. The quadratic regression lines are gray.

(19) Speakers writing one-star or five-star reviews are (or seek to create the impression that they are) in more heightened emotional states than speakers who are writing two-, three-, or four-star reviews.

Generalization (19) states a connection between contexts and ratings. Exclamatives connect contexts with linguistic forms, by indicating that the speaker is in a heightened emotional state (section 2). We formulate this as the hypothesis in (20).

(20) A speaker who uses an exclamative is in a heightened emotional state (or at least seeks to create such an impression).

Taken together, hypotheses (19) and (20) make a straightforward prediction, namely, the one in (21), which we have begun to test with the corpus experiments reported here.

(21) Exclamatives are more frequent in reviews with extreme ratings (both positive and negative).

The data for the expressions discussed above indicate that this prediction is correct, and thus support our hypotheses. The results are statistically robust, and are based on 100,000 reviews from two different kinds of review (books, hotels), written by more than 70,000 distinct authors.[6] This is a large enough data set to withstand anomalous cases in which the ratings are not indicators of the sort that (19) defines or the exclamatives are used with nonstandard (particularized) effects.

We can now use (19), (20), and our frequency data confirming (21) to address the question of just what exclamativity is. In doing so, we make use of both the speaker's perspective and the hearer's perspective. The frequency data tell us that speakers are significantly more likely to use exclamatives in extreme reviews. Thus, if one hears an exclamative, one should infer, from experience and (19), that the person who used it is (probably) in a heightened emotional state, or at least wishes to create that impression. We get a sense for the strength of that inference if we imagine a speaker who uses an exclamative when he *isn't* in an emotional state (nor intending to create that impression). This speaker's usage runs counter to our linguistic experiences, and he is likely to pay for it with a lowered rate of successful communication. His strategy just isn't a stable one for making oneself understood. In this way, exclamatives emerge as reliable windows into certain aspects of the speaker's mental state.

This direction of analysis is reminiscent of that taken by Lewis (1969, 1975), who introduces signaling games as a means of understanding how conventions, linguistic and

match those for the summary corpora, in which the individual texts are short enough (6 words, on average) that significant emotional variation internal to a single text is unlikely.

[6]The author-count is based on authors' self-identification. If many authors write under multiple names, then the actual number of distinct authors could be much lower.

otherwise, arise in a population (see also van Rooy 2004). After informally describing a signaling game, Lewis writes:

> I have now described the character of a case of signaling without mentioning the meaning of the signals [...] But nothing important seems to have been left unsaid, so what has been said must somehow imply that the signals have their meanings (Lewis 1969:124–125).

We have done something similar with our frequency data: we gave a description of the pragmatics of exclamativity entirely in terms of the expectations established, and reinforced, by frequency data. We can trace these effects to something about the lexical content of the words and phrases involved, but our purely pragmatic explanation says a lot about exclamativity without the need to resort to loose paraphrase or highly underspecified context-dependent meanings. What's more, our predictions are very clear: other corpora should yield similar distributional effects, and competent speakers should be able to make good predictions about this aspect of their interlocutors' emotional states based on the exclamatives they do (or don't) hear.

Up to now, we have discussed a small number of more or less well known exclamatives. What other expressions should we look at? Can we go beyond the already known cases of exclamatives? In the next section, we turn the reasoning above around in order to identify a general statistical profile for exclamatives. This allows us to locate all the exclamatives in a corpus with this structure, without appeal to native speakers' intuitions, a close reading of the text, or deep understanding of the context.

6 A statistical profile for exclamatives

We have so far restricted our analyses to items that we know independently to harbor exclamativity. Our corpus-based approach permits a different starting point, though: rather than looking at known exclamatives and investigating their distribution across rating categories, we can take a statistical approach to identifying potential exclamatives by systematically searching for all phrases that have the same type of distribution.

In order to classify types of distributions, we employ a logistic regression model to characterize the distribution of individual phrases. Simply put, a typical regression model tries to determine the relationship between two continuous variables by finding the function that characterizes a line that, in sum, is closest to all the data points for the two variables. To take a hypothetical example, if we were trying to establish the relationship between age and height based on data from an elementary school, a regression analysis would provide us with a function that specifies the estimated height per year of age. Since these

relationships can be not only linear, but also quadratic (or higher-order polynomial), the equation specifying this function can include polynomial terms as well. The general form for a quadratic relationship is given in (22)

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 \tag{22}$$

The constants β_i, the *coefficients*, each provide important information about the shape of the curve described by the function. We make use of this below to classify different types of distributions. However, the nature of our data doesn't allow us to use a regular regression analysis, because the variable that we are treating as the dependent one is binomial: the way that frequency information for a given phrase is encoded for the purposes of statistical analysis is to determine for each token in the corpus whether or not it is of the type of the phrase being looked at (e.g., each bigram token would be coded for whether it is of the type *what a*). The scale of rating categories for the reviews (from one to five stars), on the other hand, can be treated as a continuous one. We therefore use a logistic regression, which is used for continuous predictors and binomial dependent variables, regressing frequency of a phrase on rating category.[7] This corresponds to taking the perspective of the speaker and seeing the presence of a heightened emotional state as a precondition for using an exclamative.

Logistic regression has many statistical advantages when looking at frequencies. The basic technical move is to look not at frequencies directly, but rather at the log-odds values corresponding to them, as we did in section 4. This ensures, among other things, that the frequencies predicted by the regression are all between 0 and 1. Apart from shifting to log-odds values, however, logistic regression works just like a linear regression. As Jaeger 2008 puts it, "we can think of ordinary logit models as linear regression in log-odds space."[8]

Given that logistic regression can be seen as a linear regression in log-odds space, the basic equation still is the one in (22). We can now turn to the question of what the individual coefficents tell us about the shape of the curve described by the function. The *intercept* β_0 is not of particular interest for us, as it simply determines the value of y when $x = 0$. In a purely linear regression (without polynomial coefficients), β_1 determines the slope of the

[7] Given our discussion above, it is reasonable to see the direction of influence of one factor on another in both directions, depending on whether we take the hearer or speaker's perspective. So, in principle, we could have analyzed the data in the reverse direction as well. However, in that case, we would loose the ability to look at quadratic relationships, since squaring a binomial predictor (encoded as 0 and 1) does not make sense.

[8] This brief and informal sketch of logistic regression is only intended to make the basic statistics of the corpus experiment reported below more accessible to a broader audience. For recent discussion of the general properties and advantages of logistic regression, see Jaeger 2008 and Baayen In press:§6.3.2, where these models are defined and explored in contrast to other statistical models. (See also Bresnan and Nikitina 2008 for an application to the theory of how speakers choose among competing syntactic structures.)

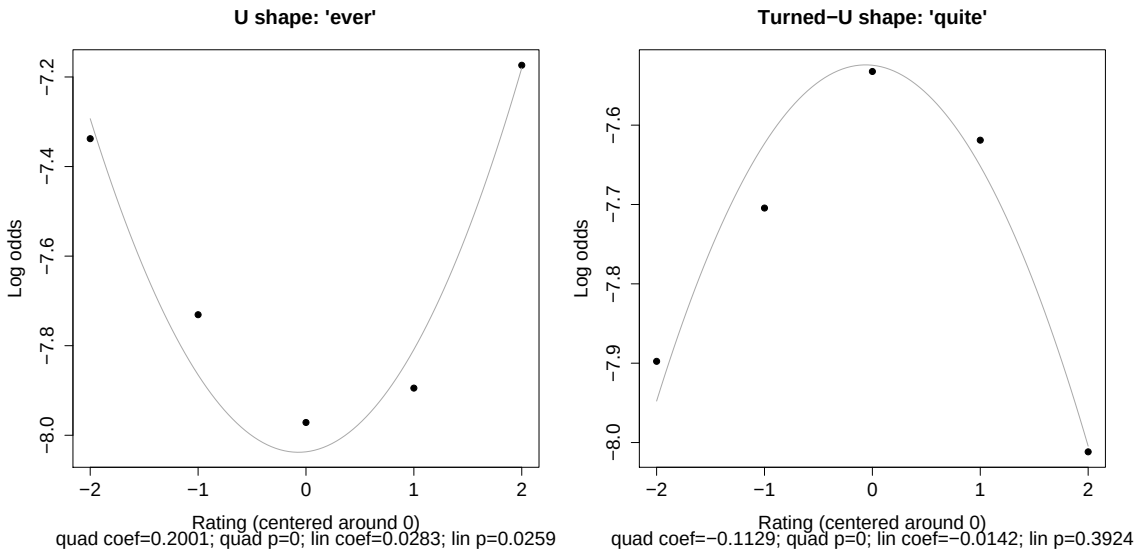
line. In a quadratic regression, β_2 determines how narrow or wide the curve is and whether it is facing up or down.

To illustrate how we make use of the information conveyed by the coefficients, we provide, in figure 7, analyses of representative phrases of each of the shapes that is directly relevant to exclamativity: the U shape, the J shape (bias for positivity), and the Reverse-J shape (bias for negativity). We also include the Turned-U distribution, a counterpoint to these exclamatives.

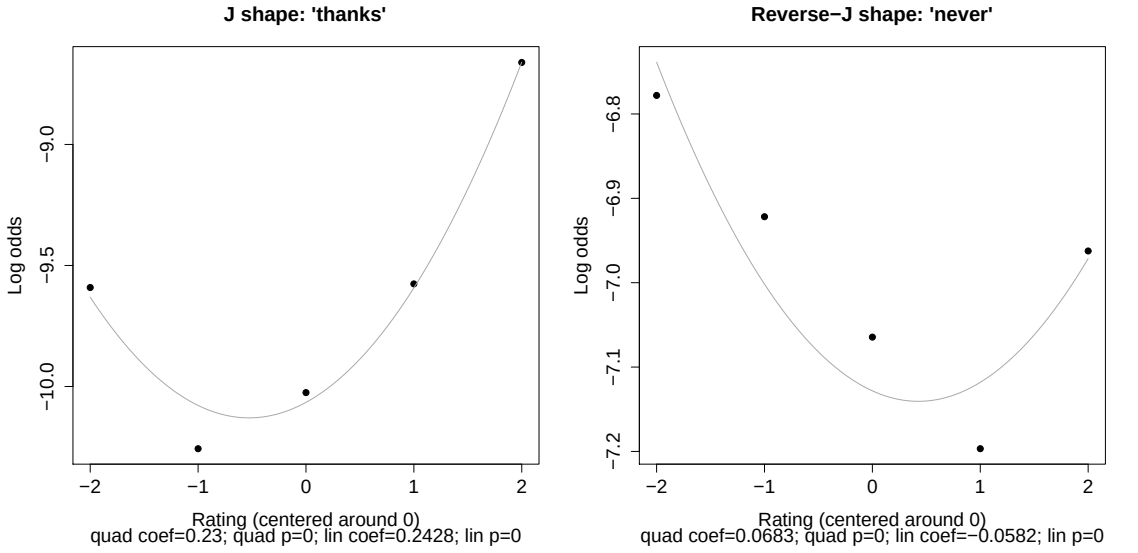
Figure 7(a) is a classic U-shape, with a positive quadratic coefficient. The U is not as deep as it is for *what a* (figure 5), but it is still evident. The only categorical difference between this shape and the Turned-U, exemplified in 7(b), is the orientation of the quadratic coefficient: for the Turned-U, it is negative.

Distinguishing U-shapes, J-shapes, and Reverse-J shapes requires us to look at the linear coefficient. Its effect in a quadratic regression is not as straightforward. Let's begin by considering what happens when it is 0. In that case, the quadratic curve will have its turning point exactly at $x = 0$, since $x^2 = 0$, and for a positive β_2, all other values of y will be bigger than the value of y at $x = 0$ (which, as we mentioned above, is identical to β_0). Therefore, if we have a scale centered around 0 and are looking at a symmetric window of the scale, a true U-shape should not have a β_1 that is significantly different from 0. In order to make use of this property with our rating scale, we shift it so that it is centered around 0 (i.e., 1 becomes -2, 2 becomes -1, 3 becomes 0, etc.). True U-shapes then are ones that have a significant quadratic coefficient, but no significant linear coefficient.

How about the J and Reverse-J shapes? Examples are given in 7(c) and 7(d). They too have significant quadratic coefficients, but the linear coefficient (β_1) now has an important role to play. If it is positive, it shifts the turning point of the U-shaped curve to the right (and also affects the height of the turning point, if the intercept remains the same). Looking at a symmetric window around 0 on the x-axis, as we are doing by looking at values from -2 to 2, this results in a curve shaped like a J, with the y-value for $x = 2$ being bigger than the y-value for $x = -2$. If the linear coefficient is negative, on the other hand, we find the reverse effect, which results in a Reverse-J-shaped curve.



(a) (b)



(c) (d)

Figure 7: Items with the shapes identified in (23). These distributions are from the Amazon review collection.

Our characterization of what it means for a set of items to have one of these distributions is summarized in (23).[9]

(23) Identifying distribution-types with a logistic regression model

Shape	Quadratic coef	Quadratic p	Linear coef	Linear p
U	positive	significant	–	nonsignificant
Turned-U	negative	significant	–	nonsignificant
J	positive	significant	positive	significant
Reverse-J	positive	significant	negative	significant

To test (23) experimentally, we ran a logistic regression for every phrase-type occurring in all four of our corpora.[10] We filtered out those phrase-types that had non-significant quadratic p-values (see the 'Quadratic p' column of (23)), and then sorted the remaining items using the quadratic coefficients and linear statistics. One can then gauge the success of (23), by appeal to intuitions or by engaging in further experiments to see whether speakers genuinely regard these phrases as signals that the speaker is in a heightened emotional state. We have only just begun the work of seriously evaluating these ideas, but the initial results are promising. In figure 8, we list the items that have U, J, or Reverse-J distributions in all four corpora (figure 8(a)) and at least three of the four corpora (figure 8(b)), along with the items that have Turned-U shapes in at least three of the four corpora (figure 8(c)).

We find obviously exclamative items like *what a*, *wow*, and sequences of exclamation points throughout figures 8(a) and 8(b), suggesting that (23) and similar hypotheses are promising sources of new data and new predictions. However, the lists also point up some shortcomings of the current approach. One clearly problematic class of items consists of function words, such as *my*, *I*, and perhaps *this*. These items often end up with significant shapes simply in virtue of their very high frequency: in logistic regression, even a very shallow U shape can be significant if the phrase is frequent enough. Figure 9 illustrates by plotting *my* (40,371 tokens) and *!!* (6,802 tokens) along the same y-axis. This direct comparison reveals that the U-shape of *my* is extremely shallow relative to the exclamative marker. Thus, we might revise (23) to be more sensitive to the size of the quadratic coefficient, so that our official U shapes are really U shapes, rather than merely 'sideways parentheses'. One could also simply exclude functions words using an independently available list of stopwords.

[9] One can also identify Turned-J and Turned-Reverse-J with this statistical model. We focus, though, on the U-shapes in this paper. In addition, one can also look at linear relationships. It turns out that a very large number of expressions have a significant (increasing or decreasing) linear relationship with the rating categories, which suggests that this class is not quite as interesting as the ones discussed above.

[10] These experiments were conducted in R (R Development Core Team 2005) using its `glm()` function. The code and datasets are available upon request.

! !	!	absolutely	all
best	ever	every	i've ever
i've	i	it !	my
the best	this	what a	wow

(a) The items whose shapes are limited to U, J, and Reverse-J for all four corpora.

! !	!	absolute	absolutely
again !	all	am	any
anyone	best	book	couldn't
even	ever !	ever	ever had
every	have ever	i	i am
i could	i have	i've	i've ever
is the	it !	life	must
my	never	new	one of
simply	the best	this	this is
time	what	what a	will
will never	wow !	wow	

(b) The items whose shapes are limited to U, J, and Reverse-J in at least three of the four corpora.

- but	a few	a good	average
basic	but nothing	but some	but still
cons	decent	few	fine
little too	mostly	near the	not bad
not quite	otherwise	part	points
pretty	pros	quite	short
some	somewhat	though	two
with some			

(c) The items whose shapes are limited to Turned-U in at least three of the four corpora.

Figure 8: Some results of applying hypotheses (23) to our corpora with the level of statistical significance set at 0.001.

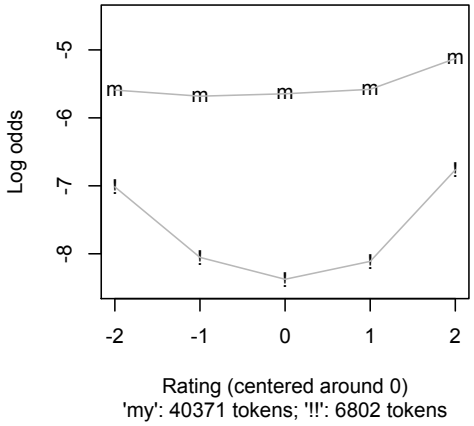


Figure 9: Log-odds plots for *my* (marked with 'm') and *!!* (marked with '!') from the Amazon review corpus, in which they both have U shapes. Very high frequency items often have significant U shapes, but shallow ones.

The lists seem also to contain a subclass of elements that harbor exclamativity (or at least have features closely correlated with exclamativity) in the context of the domain of the reviews, but not more generally. For example, *baby* is reliably U-shaped in the Tripadvisor corpora, presumably because people feel strongly about good and bad places for babies to stay. One might be able to turn this into valuable information about *particularized* exclamativity, but it would be hard to shake the close dependence on the quirks of the given data set. Thus, it seems more promising to filter off these items by bringing in more diverse data sets, and thereby getting closer to truly domain-independent exclamativity.

Finally, working with parsed corpora would also allow us to weed out spurious hits and identify more abstract patterns. The major benefit of working with raw text, as we have done, is that we can make use of massive amounts of data with a minimal investment of resources. The major drawback is that we can only approximate most linguistic information. *what a* is by no means the only constructional exclamative, as we saw in section 2. It is just one that we could easily approximate with a string.

Despite these drawbacks, we feel that figure 8 illuminates the nature of exclamativity and also phenomena that are conceptually and linguistically related to it. For example, we see items that assist in conveying exclamativity without fully determining it. The polarity-sensitive element *ever* is a clear illustration. It has a U-like distribution in all of our data sets and participates in some of the identified bigrams as well. In retrospect, this is not surprising: *ever* is often a primary signal that a sentence with the overt syntax of a polar interrogative should be read as an exclamative (cf. *Is she tall?* and *Is she ever tall!*.)

Scalar-endpoint items are another, related group represented in figures 8(a) and 8(b): *absolutely*, *even*, and a handful of superlatives. The appearance of *any* and *anyone* might trace to the same sources: these items have been argued to be explicitly endpoint-oriented (Horn 2000), and the domain-widening analysis of them is also concerned with extremes (Kadmon and Landman 1993; van Rooy 2003; Chierchia 2006). With these polarity-sensitive items come negative words as well, generally with a Reverse-J distribution. Such items inform our understanding of the distribution of polarity items (Hoeksema 1997; Lichte and Soehn 2008), and they also help identify the factors that make the exclamative signal itself pragmatically negative.

In sum, even with the statistically course-grained hypotheses (23) and our reduction of the data sets to just words and bigrams, we are still able to identify items that determine exclamativity as well as a host of items that relate intimately to it.

7 Future work

The U-shaped distributions (U, J, and Reverse-J) are not the only linguistically interesting ones that we can identify. For example, items with a Turned-U distribution are 'un-exclamatives' — hallmarks of balanced reasoning. This information too can be put to good use in understanding pragmatic inferences, especially those that concern the speaker's emotional state.

The J and Reverse-J distributions also demand closer inspection than we have given them here. These resemble exclamatives in conveying information about the speaker's emotional state, but they also carry information about whether those emotions are positive or negative. This information is not very specific ('positive' and 'negative' are very rough approximations), but it is nonetheless useful. We think that the *expressives* studied by Kaplan (1999) and Potts (2007) tend to fall into these categories. Our corpus-based approach can, therefore, provide a broader empirical basis for theoretical claims about the emotionality that these items encode and the effect that this information has on utterance understanding.

We see potential applications outside of theoretical linguistics as well. In particular, the shapes we have identified are robust across different corpora. The *generalized expectation* approach of Druck et al. (2007, 2008) permits us to take advantage of these features when constructing statistical models for document classification. We have conducted preliminary experiments using the MALLET suite of machine learning tools (McCallum 2002). The results suggest that the words and bigrams with statistically significant shapes, as described in this paper, can enhance the performance of maximum entropy classifiers (Berger et al. 1996; Nigam et al. 1999). Whereas *pool* is a good indicator of a positive review in the

Tripadvisor corpus, it is fairly useless on the Amazon data. The features of interest to us have less specialized lexical content. They are basically domain-independent, so they transfer well to new settings.

We think that pursuing sentiment classification from this perspective can have engineering payoffs, and also that it can serve as a useful empirical test of how much we know, at this point, about exclamatives and related constructions. These studies also continue to support the general claim that this expressive language is vital to communication.

References

Baayen, R. Harald. In press. *Analyzing Linguistic Data: A Practical Introduction to Statistics*. Cambridge University Press.

Beineke, Philip; Trevor Hastie; Christopher Manning; and Shivakumar Vaithyanathan. 2004. Exploring sentiment summarization. In *AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, 4–7.

Berger, Adam L.; Stephen A. Della Pietra; and Vincent J. Della Pietra. 1996. A maximum entropy approach to natural language processing. *Computational Linguistics* 22(1):39–71.

Bresnan, Joan and Tatiana Nikitina. 2008. The gradience of the dative alternation. In Linda Uyechi and Lian Hee Wee, eds., *Reality Exploration and Discovery: Pattern Interaction in Language and Life*. CSLI Publications.

Cahill, Aoife; Mairead McCarthy; Michael Burke; Josef van Genabith; and Andy Way. 2007. Deriving quasi-logical forms from F-structures for the Penn Treebank. In Harry Bunt and Reinhard Muskens, eds., *Computing Meaning*, volume 3, 33–53. Dordrecht: Springer.

Castroviejo Miró, Elena. 2006. *Wh-Exclamatives in Catalan*. Ph.D. thesis, Universitat de Barcelona.

Chierchia, Gennaro. 2006. Broaden your views: Implicatures of domain widening and the "logicality" of language. *Linguistic Inquiry* 37(4):535–590.

Clifton, Chuck; Lyn Frazier; and Britta Stolterfoht. 2008. Scale structure: Processing minimum standard and maximum standard adjectives. *Cognition* 106(1):299–324.

Cover, Thomas M. and Joy A. Thomas. 1991. *Elements of Information Theory*. New York: Wiley.

Culotta, Aron; Michael Wick; Robert Hall; and Andrew McCallum. 2007. First-order probabilistic models for coreference resolution. In *Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*.

Druck, Gregory; Gideon Mann; and Andrew McCallum. 2007. Generalized expectation criteria. Technical Report 2007-60, UMass Amherst, Amherst, MA.

Druck, Gregory; Gideon Mann; and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of ACM Special Interest Group on Information Retreival*.

Ginzburg, Jonathan and Ivan A. Sag. 2001. *Interrogative Investigations: The Form, Meaning, and Use of English Interrogatives*. Stanford, CA: CSLI.

Grodner, Daniel and Julie Sedivy. 2008. The effects of speaker-specific information on pragmatic inferences. In Neal Pearlmutter and Edward Gibson, eds., *The Processing and Acquisition of Reference*.

Hoeksema, Jack. 1997. Corpus study of negative polarity items. University of Groningen, URL http://www.let.rug.nl/hoeksema/docs/barcelona.html.

Horn, Laurence R. 2000. Pick a theory (not just *any* theory): Indiscriminatives and the free-choice indefinite. In Laurence R. Horn and Yasuhiko Kato, eds., *Negation and Polarity*, 147–192. Oxford: Oxford University Press.

Jaeger, T. Florian. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* .

Jurafsky, Daniel. 2004. Pragmatics and computational linguistics. In Laurence R. Horn and Gregory Ward, eds., *Handbook of Pragmatics*, 578–604. Oxford: Blackwell.

Kadmon, Nirit and Fred Landman. 1993. Any. *Linguistics and Philosophy* 16(4):353–422.

Kaplan, David. 1999. What is meaning? Explorations in the theory of *Meaning as Use*. Brief version — draft 1. Ms., UCLA.

Kilgarriff, Adam and Tony Rose. 1998. Measures for corpus similarity and homogeneity. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, 46–52. ACL-SIGDAT.

Lewis, David. 1969. *Convention*. Cambridge, MA: Harvard University Press. Reprinted 2002 by Blackwell.

Lewis, David. 1975. Languages and language. In Keith Gunderson, ed., *Minnesota Studies in the Philosophy of Science*, volume VII, 3–35. Minneapolis: University of Minnesota Press. Reprinted in Lewis 1983, 163–188.

Lewis, David. 1983. *Philosophical Papers*, volume 1. New York: Oxford University Press.

Liberman, Mark. 2002. Emotional prosody speech and transcripts. Linguistic Data Consortium, Philadelphia.

Lichte, Timm and Jan-Philipp Soehn. 2008. The retrieval and classification of negative polarity items using statistical profiles. In Sam Featherston and Wolfgang Sternefeld, eds., *Roots: Linguistics in Search of its Evidential Base*. Mouton de Gruyter.

Liscombe, Jackson; Jennifer Venditti; and Julia Hirschberg. 2003. Classifying subject ratings of emotional speech using acoustic features. In *Proceedings of Eurospeech*, 725–728.

Manning, Christopher D. and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: MIT Press.

McCallum, Andrew. 2002. MALLET: A machine learning for language toolkit. Software package, UMass Amherst, URL http://mallet.cs.umass.edu/.

Nigam, Kamal; John Lafferty; and Andrew McCallum. 1999. Using maximum entropy for text classification. In *IJCAI-99 Workshop on Machine Learning for Information Filtering*, 61–67.

Nigam, Kamal; Andrew McCallum; Sebastian Thrun; and Tom M. Mitchell. 1998. Learning to classify text from labeled and unlabeled documents. In *Proceedings of the 15th National Conference on Artificial Intelligence*, 792–799. Menlo Park, CA: American Association for Artificial Intelligence.

Noveck, Ira A. and Dan Sperber, eds. 2004. *Experimental Pragmatics*. Houndsmills, Basingstoke, Hampshire: Palgrave Macmillan.

Palmer, Martha; Dan Gildea; and Paul Kingsbury. 2005. The Proposition Bank: A corpus annotated with semantic roles. *Computational Linguistics* 31(1):71–106.

Pang, Bo and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Pang, Bo and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Pang, Bo; Lillian Lee; and Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Potts, Christopher. 2007. The expressive dimension. *Theoretical Linguistics* 33(2):165–197.

Pradhan, Sameer S.; Eduard Hovy; Mitch Marcus; Martha Palmer; Lance Ramshaw; and Ralph Weischedel. 2007. OntoNotes: A unified relational semantic representation. In *Proceedings of the First IEEE International Conference on Semantic Computing*. Irvine, CA.

Prasad, Rashmi; Nikhil Dinesh; Alan Lee; Eleni Miltsakaki; Livio Robaldo; Aravind Joshi; and Bonnie Webber. 2008. The Penn Discourse Treebank 2.0. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*.

R Development Core Team. 2005. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, URL http://www.R-project.org/.

Rett, Jessica. 2008. A degree account of exclamatives. In Masayuki Gibson and Tova

Friedman, eds., *Proceedings of Semantics and Linguistics Theory 18*. Ithaca, NY: CLC Publications.

van Rooy, Robert. 2003. Negative polarity items in questions: Strength as relevance. *Journal of Semantics* 20:239–273.

van Rooy, Robert. 2004. Signalling games select Horn strategies. *Linguistics and Philosophy* 27(4):493–527.

Schwarz, Florian; Chuck Clifton; and Lyn Frazier. 2008. Strengthening 'or': Effects of focus and downward entailing contexts on scalar implicatures. Ms., UMass Amherst.

Sedivy, Julie. 2007. Implicatures in real-time conversation: A view from language processing research. *Philosophy Compass* 2/3:475–496.

Shanahan, James G.; Yan Qu; and Janyce Wiebe, eds. 2005. *Computing Attitude and Affect in Text: Theory and Applications*. Dordrecht: Springer.

Webber, Bonnie. 2006. Accounting for discourse relations: Constituency and dependency. In Mirian Butt; Mary Dalrymple; and Tracy King, eds., *Intelligent Linguistic Architectures*, 339–360. CSLI.

Webber, Bonnie; Matthew Stone; Aravind Joshi; and Alistair Knott. 2003. Anaphora and discourse structure. *Computational Linguistics* 29(4):545–587.

Zanuttini, Raffaella and Paul Portner. 2003. Exclamative clauses at the syntax–semantics interface. *Language* 79(1):39–81.

A Information about the data sets

This appendix provides basic information about the data sets used in this paper. The word and vocabulary counts include all stop-words, and tokenization was done by removing all case distinctions, stripping off all HTML mark-up, and treating most word-final punctuation marks as separate words. When counting reviews, we ignored pages that lacked any content for the field of interest, which is why the review counts differ slightly between the 'summary' and 'review' sections. The rightmost column gives totals, which are the sum of the columns in that row, except for the vocabulary row, where the total is a count of the union of the vocabularies for each rating category.

The Tripadvisor.com reviews were written by a total of 35,713 authors, and the Amazon.com reviews were written by a total of 40,625 authors. These numbers are based on the authors' self-identification. The actual number of distinct authors might be much lower.

A.1 Tripadvisor.com summaries

	1 star	2 star	3 star	4 star	5 star	total
reviews	2,989	4,300	5,410	17,950	25,200	55,849
words	14,794	20,908	26,266	80,597	114,305	256,870
vocab	2,417	3,052	3,134	5,272	5,651	10,819

A.2 Tripadvisor.com reviews

	1 star	2 star	3 star	4 star	5 star	total
reviews	2,896	4,130	4,948	15,801	22,450	50,225
words	605,207	877,854	974,271	2,726,796	3,577,764	8,761,892
vocab	19,128	23,534	24,761	42,776	49,492	85,425

A.3 Amazon.com summaries

	1 star	2 star	3 star	4 star	5 star	total
reviews	3,322	2,684	3,993	8,598	34,946	53,543
words	16,830	13,518	20,779	43,607	182,377	277,111
vocab	3,434	3,019	3,785	6,025	11,482	15,930

A.4 Amazon.com reviews

	1 star	2 star	3 star	4 star	5 star	total
reviews	3,323	2,687	3,994	8601	34,952	53,557
words	570,687	512,643	767,958	1,513,776	4,769,921	8,134,985
vocab	27,352	26,239	32,818	46,306	80,569	112,323