Scalar implicature in adverbial vs. nominal quantifiers: Two experiments

Johanna Alstott
Massachusetts Institute of Technology

1 Overview

This work experimentally investigates scalar implicature (SI) in an understudied domain, adverbial quantifiers (*sometimes*, *usually*), and compares SI in this domain to SI in nominal quantifiers (*some*, *most*). This investigation probes the parallels between adverbials and nominals posited in the literature and touches on core questions of which alternatives factor into SI computation. The two experiments use a Degen & Tanenhaus (2015)-inspired "calendar paradigm," in which participants see calendars showing a character's shirts over a two-week period and rate the naturalness of quantified sentences about the calendars. Experiment 1, in which participants rated only adverbials or only nominals, finds that *some* and *sometimes* have similar SI profiles, though *most* carries a stronger *not all* inference than *usually* does. Experiment 2 shows that participants think about nominals and adverbials more similarly when they see both. I unify the results of the two experiments by appealing to the different domain specification in the adverbial and nominal stimuli.

2 Introduction and background

Consider the following sentences:

(1) a. In the last two weeks, Morgan wore a blue shirt on some days.
b. In the last two weeks, Morgan wore a blue shirt on most days.

Both (1a) and (1b) imply that Morgan did not wear a blue shirt every day in the last two weeks, while (1a) additionally implies that Morgan did not wear a blue shirt on most days. Following Horn (1972), these inferences are known as scalar implicatures (SI). A large literature has devoted itself to investigating SI. While there is broad agreement that SIs arise via some form of competition between the sentence spoken and similar alternative utterances the speaker could have said (e.g. competition between (1a), (1b), and a similar sentence using *every day*), semanticists argue about whether this competition is a product of Gricean pragmatics (Horn 1972), the lexical entries for scalar items (Levinson 2000), or a silent operator in the syntax (Chierchia *et al.* 2012). In tandem with this theoretical debate, an ever-growing body of experimental research has investigated the acquisition (e.g. Papafragou & Musolino 2003), processing (e.g. Degen & Tanenhaus 2015) and interpretation of SI (e.g. Chemla & Spector 2011).

While the SI literature is immense in both breadth and depth, it is narrow in the scalar items it studies. The fact that SI occurs with a wide variety of expressions has been acknowledged for half a century (Horn 1972), but the literature has

mainly focused on three: nominal quantifiers as in (1a-b), logical connectives (*or* implies *not and*) and natural numbers (*three* implies *not four*, *not five*, etc.). Because of the overwhelming focus on these cases, other SI domains poised to enrich our understanding of the phenomenon have gone understudied. In this paper, I initiate experimental research into SI in one such domain, adverbial quantifiers (*sometimes*, *usually*), and compare SI in this domain to SI in nominal quantifiers like in (1).

(2) a. In the last two weeks, Morgan sometimes wore a blue shirt.
$\rightsquigarrow$ Morgan did not usually/always wear a blue shirt.

b. In the last two weeks, Morgan usually wore a blue shirt.
$\rightsquigarrow$ Morgan did not always wear a blue shirt.

Although some have noted in passing the existence of adverbial quantifiers as an SI domain (Papafragou & Musolino 2003), no SI literature to my knowledge has investigated this domain or compared SI in adverbial and nominal quantifiers.

Despite this lack of attention, SI in adverbial quantifiers merits study for three reasons, the first of which comes from cross-linguistic work. While languages like Straits Salish, Mohawk, and Warlpiri have been shown to lack nominal quantification altogether (Bach *et al.* 1995), quantification via adverbs, adjectives, or auxiliaries (A-quantification) appears across attested languages (Partee 2008). Within A-quantification, Partee (2008) singles out adverbial quantification as a candidate for linguistic universality, so researching SI in adverbial quantifiers can give insight into how this process works in a potentially universal domain.

Researching SI in adverbial quantifiers experimentally also allows us to test the predictions of formal semantic theories of adverbial quantifiers. These theories all assume close parallels between adverbial quantifiers and nominal ones. For example, Lewis (1975)'s approach treats *always* as equivalent to *every* except when it comes to the number of variables they can bind. Similarly, the competing framework of de Swart (1991) and von Fintel (1994) assumes that *always* universally quantifies over situations in an analogous way to how *every* universally quantifies over individuals. Lewis (1975), de Swart (1991), and von Fintel (1994) posit these parallels not just for *every* and *always* but also for *some* and *sometimes*, *most* and *usually*, *no* and *never*, etc. With the exception of Alstott & Jasbi (2020), these parallels have never been probed experimentally. Alstott & Jasbi (2020) find evidence for the parallels proposed in the literature, but their task involved judgments of quantificational force rather than SI. An experimental comparison of SI in nominal vs. adverbial quantifiers thus allows us to investigate whether the parallels from the literature hold in the kinds of scalar contexts critical to the use of quantifiers.

Finally, investigating SI in adverbial quantifiers alongside SI in nominal quantifiers touches on a core question in the SI literature: if SIs arise as a result of competition between alternatives (as is commonly accepted), what counts as an alternative? From the perspective of formal theories of alternative sets (e.g. Matsumoto 1995; Katzir 2007), adverbial quantifiers are alternatives for one another and nominal quantifiers are alternatives for one another; however, adverbials would not count as alternatives for nominals or vice versa in these theories. For example, in Katzir (2007)'s system, alternatives must be derived from one another via substitution or deletion. As words of different syntactic categories, one cannot substitute

a nominal quantifier like *some* for an adverbial quantifier like *sometimes* and get a grammatical result, so they are not alternatives in Katzir (2007)'s system.

Although nominal and adverbial quantifiers are not formal alternatives, it does not follow that nominal quantifiers have no effect on SI computation for adverbials or vice versa: SI computation can be affected by unsaid utterances other than the formal alternatives ("contextual alternatives" in Degen & Tanenhaus (2015)'s terms). For example, Degen & Tanenhaus (2015) find that contexts that make numbers salient alternatives to *some* affect SI processing for *some*, even though numbers and *some* are generally not considered formal alternatives for one another.

Since adverbially and nominally quantified sentences are not formal alternatives for one another but coexist as 'things the speaker could have said' in certain contexts (e.g. in (1) and (2)), the question arises of whether they count as contextual alternatives. Regardless of whether the answer to this question is positive or negative, answering it can help demarcate the space of possible contextual alternatives. If nominals and adverbials are contextual alternatives, we expect that SI computation in contexts that make both nominal and adverbial quantifiers salient will differ from SI computation in contexts in which only one domain is salient.

Having identified some motivations for embarking on an investigation of SI in adverbial quantifiers, we can identify two guiding questions. First is a question of *comparison*, intended to probe the parallels assumed by theoretical work: (I) do we see identical alternative-based scalar reasoning among adverbial quantifiers (*sometimes*, *usually*, *always*) as we do among intuitively parallel nominal quantifiers (*some*, *most*, *every*)? Second is a question of *interaction*, inspired by the above considerations about alternative sets: (II) If the answer to (I) is "yes," can nominals and adverbials function as (contextual) alternatives in the kinds of settings where nominals can for each other and adverbials can for each other?

To distill these questions into a formulation answerable by a set of experiments, I draw inspiration from the methodology of Degen & Tanenhaus (2015), who studied SI in *some (of)*. In their trials, participants saw contexts in which they "obtained" varying numbers of gumballs from a pool of thirteen and judged the naturalness of *you got some (of) the gumballs* given the context. Via this design, they ascertained how natural it is to use *some (of) the gumballs* to describe zero gumballs, one gumball, two gumballs, etc., all the way up to the maximal set of thirteen.

This sort of experimental design is well-suited to our two guiding questions. For one thing, this approach offers a way to compare SI computation between two quantifiers, making it suitable for investigating question (I). Degen & Tanenhaus (2015) use their approach to compare the SI profiles of *some* and *some of*. They found that naturalness for *some* and *some of* decrease in tandem on trials where the participant "got" more than half of the gumballs, indicating that *some* and *some of* have a *not most* inference of similar strength. However, *some* was more natural than *some of* to describe the maximal set of thirteen gumballs, indicating that *some* generates a weaker *not all* inference than *some of* does. Just like Degen & Tanenhaus (2015) compare the SI profiles of *some* and *some of* by looking at their naturalness across set sizes, we can compare the SI profiles of *some* and *sometimes* or the profiles of *most* and *usually* by ascertaining their naturalness across set sizes.

Degen & Tanenhaus (2015)'s approach is also suitable for investigating question (II), as their approach was built to test the influence of contextual alternatives

on SI computation. Degen & Tanenhaus (2015) investigated whether naturalness ratings for *some* and *some of* were affected by the inclusion vs. exclusion of natural numbers from the stimuli. Similarly, to probe (II), we could look at how naturalness ratings for adverbial quantifiers are affected by the inclusion vs. exclusion of nominal quantifiers in the stimuli and vice versa.

Armed with Degen & Tanenhaus (2015)'s design philosophy, we can reformulate our guiding questions in concrete form. (I) **Comparison**: when participants see only nominally quantified sentences or only adverbially quantified sentences, do naturalness ratings for adverbials mirror those for intuitively parallel nominals across set sizes? (II) **Interaction**: If adverbial quantifiers are introduced alongside nominal quantifiers via a within-subjects design where participants rate both, how are naturalness ratings affected?

I address these questions via two experiments, with Experiment 1 focusing on question (I) and Experiment 2 focusing on question (II).

3 Experiment 1

3.1 Methodology

3.1.1 Participants and materials

200 adult participants reporting their first language as English were recruited via Prolific. Participants received $1.10 for their participation.

The experiment used an original "calendar paradigm" similar to Degen & Tanenhaus (2015)'s "gumball paradigm" but more suitable for comparison of nominals and adverbials. *You got some gumballs* has no adverbially quantified counterpart, making the gumball paradigm not amenable to comparison of nominals and adverbials. For example, *You sometimes got a gumball* quantifies over instances instead of individuals, so we cannot make a comparison with *You got some gumballs*.

On each trial of the calendar paradigm (see Figure 1 for example), participants saw a two-week mini-calendar populated by stick figures wearing different colored

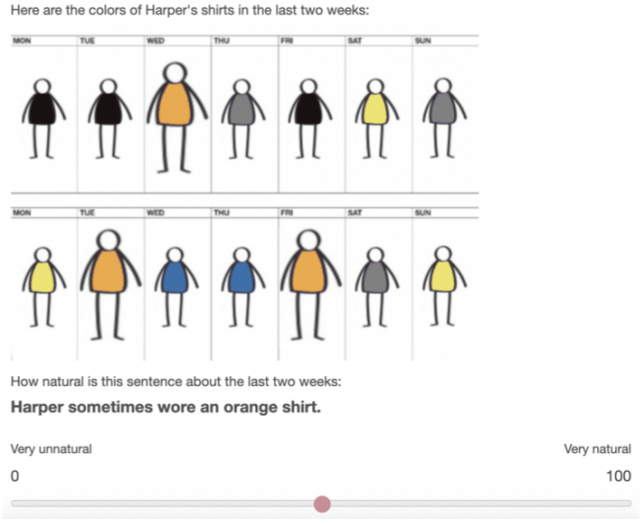

Figure 1: A sample trial within the calendar paradigm.

shirts; each calendar was prefaced by a specification that the calendar depicts "the colors of [NAME]'s shirts in the last two weeks". Figures wearing the "target color" (the color used in the rated sentence) were larger than the others; as participants were told in the task instructions, the size difference had no significance beyond flagging the target color. Below each mini-calendar was a quantified sentence that made a claim about the frequency with which the character wore shirts of the target color over the last two weeks. The sentences, inspired by the method of comparing adverbial and nominal quantifiers in Alstott & Jasbi (2020), involved adverbial quantification as in Figure 1 or nominal quantification over days (e.g. *Harper wore an orange shirt on some days*). Participants then rated the naturalness of the sentence using a slider that ranged from 0 ("very unnatural") to 100 ("very natural"). Trials differed with respect to the quantifier, the number of days out of 14 on which the target color was worn (which I will call "set size" going forward), the name of the character, and the target color (orange, blue, yellow, black, or gray).

The quantifiers tested using this calendar paradigm were *sometimes*, *usually*, *always*, and *never* on the adverbial side and *some*, *most*, *every*, and *no* on the nominal side. One could have used *mostly* instead of *usually*, but the *most/usually* comparison is particularly interesting as a case of two quantifiers not compositionally built from each other that are nonetheless treated in the literature as having parallel semantics. I used *no/never* and *every/always* as "baseline" quantifiers of sorts. *No/never* should be rated very highly when the calendar shows no figures wearing the target color, while *every/always* should be rated very highly when every figure on the calendar is wearing the target color. As such, we can use *no/never* with set size 0 and *every/always* with set size 14 as benchmarks for high naturalness.

3.1.2 Procedure

After a consent sheet, participants saw practice trials where they rated the naturalness of sentences like *Evan wore a black shirt on Thursdays* against the backdrop of a calendar that made the sentence true as well as a calendar that made it false. After the practice trials, participants were randomly sorted into one of 40 lists, 20 of which consisted only of adverbial trials and 20 of which consisted only of nominal trials. The nominal and adverbial lists were identical to one another except for the quantifiers. For example, nominal list 1 and adverbial list 1 were identical except that the *every*, *most*, *some*, and *no* sentences in nominal list 1 showed up as *always*, *usually*, *sometimes*, and *never* sentences (respectively) in adverbial list 1.

Like Degen & Tanenhaus (2015)'s Experiment 1, each list contained 16 trials. Nominal quantifier lists contained six *some* trials, six *most* trials, two *no* trials, and two *every* trials, while adverbial quantifier lists contained six *sometimes* trials, six *usually* trials, two *never* trials, and two *always* trials. Within a given list, the six trials with *some* or the six trials with *sometimes* consisted of one trial with set size 0 (i.e. a trial with a calendar that showed zero shirts of the target color), two trials in the subitizing range (between 1-4 shirts of the target color), one trial in the mid-range (5-9), one trial in the high range (10-13), and one trial with set size 14. The six trials with *most* or the six trials with *usually* had a similar breakdown, but these quantifiers had two trials with high-range set sizes (10-13) and only one in the subitizing range (1-4). The two trials with *every* or the two trials with *always*

consisted of one trial with set size 14 and one trial in the mid range. The two trials with *no* or the two trials with *never* consisted of one trial with set size 0 and one trial in the mid range. The 16 trials in each list were presented in random order.

Although the main scalar quantifiers of interest (*some*, *sometimes*, *most*, and *usually*) were not rated with every set size in a given list, I implemented experiment-wide controls for how often each quantifier occurred with each set size. I also implemented within-list and experiment-wide controls for character name and target color, which are detailed at `https://tinyurl.com/2p8fakn4`.

3.2 Results and analysis

Figure 2 shows mean ratings for *some/sometimes* and *most/usually* with set sizes 0-14, *no/never* with set size 0, and *every/always* with set size 14. From the original pool of 200 participants, I chose to exclude those who failed to rate the true practice trial more natural than the false practice trial. This resulted in 29 participants being excluded from analysis. Of the 171 participants whose data were used, 86 had been assigned to an adverbial list and 85 had been assigned to a nominal list.

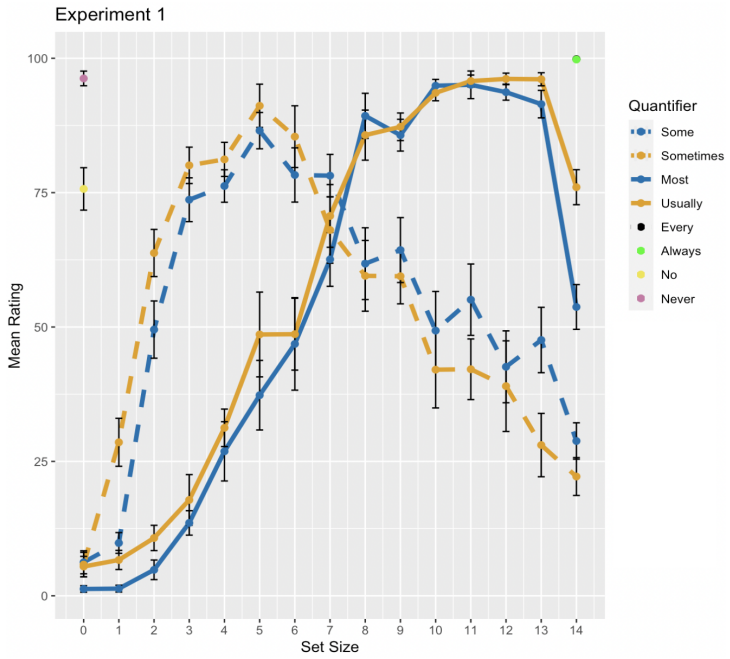

Figure 2: Experiment 1 results. Error bars are SEM. *Every* and *always* were rated so similarly at set size 14 that *every* is barely visible, and the SEMs were infinitesimal.

My analyses aimed at answering the main question behind Experiment 1: did naturalness ratings for the scalar quantifiers *sometimes* and *usually* mirror those of *some* and *most* across set sizes? In other words, for which (if any) set sizes did *some* and *sometimes* differ in naturalness, and for which did *most* and *usually* differ?

To tackle this question, I adopted the style of analysis Degen & Tanenhaus (2015) used when comparing naturalness across set sizes for *some* and *some of*. For *most/usually* and *some/sometimes* used with each set size (0-14), I fit one linear regression model to each subset of the data corresponding to that set size and pair of

quantifiers. For example, I fit one model to the subset of the data involving *most* or *usually* with set size 12, a different model to the subset involving *some* or *sometimes* with set size 3, etc. Each model predicted naturalness rating from quantifier.

Mean ratings for *most* and *usually* only significantly differed for the maximal set of 14 (β = 22.282, SE = 5.279, p $<$ 0.0001). Ratings for *some* and *sometimes* only differed significantly for set size 1 (β = 18.722, SE = 4.86, p $<$ 0.001), where *sometimes* was more natural, and set size 13 (β = -19.542, SE = 8.466, p = 0.02), where *some* was more natural. *Sometimes* was marginally more natural than *some* for set size 2 (β = 14.246, SE = 6.833, p = 0.04).

I made similar comparisons for *every*/*always* at set size 14 and *no*/*never* at set size 0, finding no significant difference for the former but finding that *never* was rated significantly higher than *no* at set size 0 (β = 20.550, SE = 4.157, p $<$ 0.0001).

3.3 Experiment 1 discussion

Experiment 1 set out to answer an overarching comparative question: do we see similar alternative-based scalar reasoning among adverbial quantifiers as we do among intuitively parallel nominal quantifiers? As a means to the end of answering this question, Experiment 1 ascertained whether naturalness ratings for *sometimes* and *usually* mirror those for the intuitively parallel *some* and *most* across the set sizes of a Degen & Tanenhaus (2015)-style task. Analyses indicate that naturalness ratings across set sizes were overwhelmingly similar for the pairs *some*/*sometimes* and *most*/*usually*. There were only a small handful of set sizes where ratings for *most* significantly differed from those for *usually* or where ratings for *some* significantly differed from those for *sometimes*: *most*/*usually* with 14 and *some*/*sometimes* with 1, 2, and 13. The overall very similar ratings for *some* and *sometimes* and for *most* and *usually* indicate that these words have very similar meanings and provide further evidence on top of Alstott & Jasbi (2020) that the parallels between nominals and adverbials proposed in the literature are on the right track.

What do these results tell us about SI in nominal vs. adverbial quantifiers? The results suggest that SI computation in *some* and *sometimes* is very similar. *Some* and *sometimes* received similarly low naturalness ratings on trials where the target color was worn every day, indicating that *some* and *sometimes* generate a *not all* inference of similar strength. Naturalness for both *some* and *sometimes* peaked with set size 5 and decreased in tandem on trials where the target color was worn on more than half of the days, indicating a *not most*/*not usually* inference of similar strength. While ratings for *some* and *sometimes* did differ on a few set sizes, the reason for these disparities likely has little to do with SI (see Section 5).

While SI computation for *some* and *sometimes* was very similar, SI computation for *most* differed from SI computation for *usually*: *usually* was rated much higher than *most* on trials where the target color was worn every day, indicating that *usually* generates a weaker *not all* implicature than *most* does. If *usually* and *most* express perfectly parallel meanings (as in the theories of adverbial quantifiers discussed in Section 2), this disparity is surprising. As such, these results suggest the need for a revision to theories of *most* and *usually* that captures their differences when it comes to SI as well as the many ways in which they are similar.

The strangely low ratings for *no* with set size 0 are likely a result of some par-

ticipants judging *on no days* ungrammatical (?*Quinn wore a blue shirt on no days*).

4 Experiment 2

While Experiment 1 gave a sense of how SI compares between adverbial and nominal quantifiers, the fact that participants saw only adverbials or only nominals means that questions of how the two SI domains interact remain open. For example, when both nominals and adverbials are made salient, do the divergences in naturalness between the domains from Experiment 1 become more or less pronounced? Experiment 2 set out to answer this question, which has consequences for the issue of whether nominal and adverbial quantifiers can be contextual alternatives. If they can, we expect divergences in naturalness between the domains to be more pronounced when both nominals and adverbials are salient. Consider Degen & Tanenhaus (2015)'s Experiment 2: they had evidence that numbers can be contextual alternatives to *some* because including numbers in the stimuli decreased naturalness for *some* at set sizes where numbers were more natural alternatives. Similarly, if an adverbial like *usually* is a contextual alternative to a nominal quantifier like *most*, we expect that interspersing *usually* trials with *most* trials will decrease naturalness for *most* at the set size where *usually* is a more natural alternative (i.e. set size 14).

4.1 Methodology

200 adult participants reporting their first language as English were recruited via Prolific. Participants received $1.80 for their participation.

The "calendar paradigm," quantifiers tested, and practice trials were the same in Experiments 1 and 2. After the practice trials, participants were randomly sorted into one of 40 lists, each of which contained both adverbially quantified and nominally quantified sentences.[1] Like Degen & Tanenhaus (2015)'s Experiment 2, each list consisted of 32 trials. The lists in Degen & Tanenhaus (2015)'s Experiment 2 had 16 nominal quantifier trials and 16 number trials. Similarly, my Experiment 2's lists had 16 nominal trials and 16 adverbial trials: six *some* trials, six *most* trials, two *every* trials, two *no* trials, six *sometimes* trials, six *usually* trials, two *always* trials, and two *never* trials. The make-up of the six *some* trials, the six *most* trials, etc. followed the same pattern as Experiment 1. For example, *some* and *sometimes* each occurred once with set size 0, twice with subitizing range set sizes (1-4), once with a mid-range set size (5-9), once with a high-range set size (10-13), and once with set size 14. The trials in each list were presented in random order.

4.2 Results and analysis

Figure 3 shows mean ratings for *some*, *sometimes*, *most*, and *usually* with set sizes 0-14, *no*/*never* with set size 0, and and *every*/*always* with set size 14. Out of the 200 participants, I excluded those who failed to rate the true practice trial higher than the false one. This resulted in 20 participants being excluded from analysis.

[1]The 40 lists in experiment 2 consisted of 20 "list types," with each list type having a Version A and Version B. Each of the list types retained the 16 trials from their Experiment 1 counterparts; the 16 Experiment 1 trials were assigned to adverbials in Version A and nominals in Version B. Details on each list's 16 new trials can be found at `https://tinyurl.com/4tphytbe`.

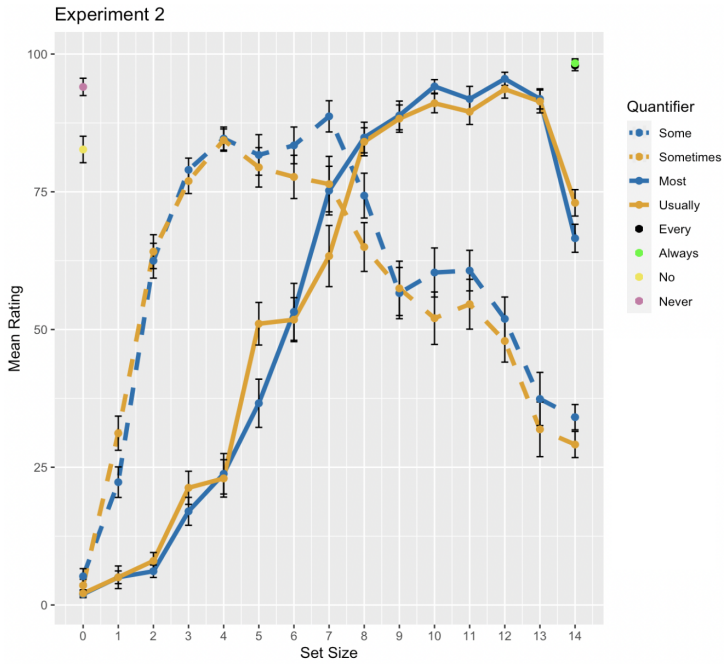

Figure 3: Experiment 2 results. Error bars represent SEM.

I ran three sets of analyses. The first set of analyses, which aimed to ascertain the extent to which Experiment 2 replicated Experiment 1, involved running the analyses I previously performed on the Experiment 1 data (Section 3.2) on the Experiment 2 data. For *most/usually* and *some/sometimes* with each set size (0-14), I fit one linear regression model to subsets of the data corresponding to that set size and pair of quantifiers. Each model predicted naturalness rating from quantifier.

Ratings for *most* and *usually* in Experiment 2 differed for set size 5 (β = 14.436, SE = 5.839, p = 0.01) and marginally at set size 14 (β = 6.433, SE = 3.496, p = 0.06), where *usually* was more natural. Ratings for *some* and *sometimes* differed for set size 1 (β = 8.892, SE = 4.158, p = 0.03), where *sometimes* was more natural, and set size 7 (β = -12.303, SE = 5.776, p = 0.03), where *some* was more natural. Ratings for *some* and *sometimes* did not differ for set sizes 2 and 13, unlike Experiment 1.

The second set of analyses aimed at determining if the divergences in naturalness observed in both experiments (*most/usually* at 14; *some/sometimes* at 1) became more or less pronounced in Experiment 2. For subsets of the data involving *most* at 14, *usually* at 14, *some* at 1, and *sometimes* at 1, I fit one regression model apiece that predicted mean rating from experiment (1 vs. 2).

At set size 14, *most* and *usually* were rated more similarly in Experiment 2: ratings for *most* at set size 14 were significantly higher in Experiment 2 (β = 12.832, SE = 4.681, p $<$ 0.01), while ratings for *usually* did not differ between experiments. At set size 1, *some* and *sometimes* were also rated more similarly in Experiment 2: ratings for *some* at set size 1 were higher in Experiment 2 (β = 12.456, SE = 4.406, p $<$ 0.01), while ratings for *sometimes* did not differ between experiments.

The final set of analyses aimed at determining the source of the two divergences in naturalness observed in Experiment 2 but not in Experiment 1 (*most/usually* at 5;

some/*sometimes* at 7). For example, did *some* and *sometimes* at set size 7 differ in Experiment 2 but not in Experiment 1 due to higher ratings for *some* in Experiment 2? To answer these sorts of questions, I fit one linear regression model apiece for subsets of the data involving *some* at 7, *sometimes* at 7, *most* at 5, and *usually* at 5 that predicted mean rating from experiment. Ratings for *some* at 7 were significantly higher in Experiment 2 (β = 10.539, SE = 4.792, p = 0.03), while ratings for *sometimes* at 7, *most* at 5, and *usually* at 5 did not differ between experiments.

4.3 Experiment 2 discussion

Experiment 2 set out to answer an overarching question of interaction: can nominal and adverbial quantifiers function as contextual alternatives in the kinds of settings where nominals can for each other and adverbials can for each other? To address this question, Experiment 2 showed participants both nominal and adverbial trials using the calendar paradigm and ascertained whether and how naturalness ratings for *some*, *sometimes*, *most* and *usually* changed vis-à-vis Experiment 1, in which participants saw only adverbials or only nominals. Analyses indicate that Experiment 2 mostly replicated Experiment 1: naturalness ratings across set sizes were very similar for the pairs *some*/*sometimes* and *most*/*usually*. However, one consistent difference between the results of the two experiments bears mention: the divergences in naturalness between adverbials and nominals observed in Experiment 1 either narrowed (*some*/*sometimes* at 1 and *most*/*usually* at 14) or collapsed entirely in Experiment 2 (*some*/*sometimes* at 2 and 13).

These results provide evidence against the hypothesis that nominal and adverbial quantifiers can be contextual alternatives to one another. Recall that if adverbial and nominal quantifiers are contextual alternatives, we expect divergences in naturalness between nominals and adverbials to become more pronounced when both domains are salient. For example, if *sometimes* is a contextual alternative to *some* at set size 1, we expect ratings for *some* at set size 1 to be lower among those exposed to the more natural *sometimes* (Experiment 2) than among those who were not (Experiment 1). Similarly, if *usually* is a contextual alternative to *most* at set size 14, we expect ratings for *most* at set size 14 to be lower among those exposed to the more natural *usually* than among those who were not. But the opposite occurred: *some* at set size 1 and *most* at set size 14 were rated as more natural in Experiment 2 than in Experiment 1, near the naturalness level of their adverbial counterparts.

Some/*sometimes* and *most*/*usually* did differ for some new set sizes in Experiment 2 (*some*/*sometimes* at 7, *most*/*usually* at 5). However, neither of these new disparities provide evidence that adverbial and nominal quantifiers can function as contextual alternatives, as neither of them stem from a quantifier's naturalness decreasing in Experiment 2. Since these new disparities have no bearing on the main question that guided this experiment, I set them aside going forward.

5 General discussion

Experiments 1 and 2 set out to address (I) whether the parallels between nominal and adverbial quantifiers proposed in the literature hold in scalar contexts; and (II) whether nominal and adverbial quantifiers can be contextual alternatives to one

another. Experiment 1, which focused on question (I), compared SI in nominal and adverbial quantifiers by ascertaining whether naturalness ratings for *sometimes* and *usually* mirror those for *some* and *most* across the set sizes of a Degen & Tanenhaus (2015)-style task. Experiment 1 found that naturalness ratings across set sizes were strikingly similar for *some* and *sometimes* and for *most* and *usually*, which indicates that the parallels between these quantifiers proposed in the literature are on the right track. However, Experiment 1 also found that *usually* generates a weaker *not all* implicature than *most* does, which suggests that formal semantic theories should draw a more substantive distinction between *most* and *usually* when it comes to SI.

Experiment 2 investigated question (II) by showing participants both nominal and adverbial trials and ascertaining how naturalness ratings changed from Experiment 1, in which participants only saw adverbials or only nominals. Experiment 2 found no evidence indicating adverbial alternatives for nominals or vice versa. If adverbials and nominals were contextual alternatives, we would expect divergences in naturalness between the two domains to become more pronounced in Experiment 2, where both domains are salient. However, participants actually rated adverbial and nominal quantifiers more similarly in Experiment 2 than in Experiment 1.

Two questions about these results remain: why did ratings for *some* and *sometimes* significantly differ for set sizes 1, 2, and 13 in Experiment 1, and why did the divergences in naturalness between adverbials and nominals observed in Experiment 1 narrow or collapse entirely in Experiment 2? I suggest that the answer to both questions lies in the differential levels of domain specification in the nominal and adverbial stimuli. The overt reference to days in the nominal quantifier trials (*on some days*, *on most days*, etc.) likely made it clear to anyone seeing these trials that the 14 days in the calendar were the domain of quantification. By contrast, the domain of quantification was underspecified in the adverbial sentences, which contained no overt reference to days (*Harper sometimes wore an orange shirt*, e.g.).

Because Experiment 1 participants were exposed to only adverbials or only nominals, some adverbial condition participants might have interpreted the domain differently than those in the nominal condition, translating to small differences like those observed for *some* and *sometimes* at set sizes 1, 2, and 13. For example, it is quite unnatural to use an existential quantifier to describe one or two entities out of 14, so the higher ratings for *sometimes* vis-à-vis *some* at set sizes 1 and 2 would make sense if some participants judging *sometimes*, unlike the participants judging *some*, interpreted the domain as something other than 14 days. For example, perhaps some adverbial condition participants interpreted *sometimes* as quantifying over moments in time rather than days. For these participants, *sometimes* would be fairly natural for set sizes 1 and 2 because there are far more than a couple times (moments in time) in which the target color was worn. The higher ratings for *sometimes* at set sizes 1 and 2 would also make sense if some in the adverbial condition interpreted the calendars as representing a general shirt-wearing pattern rather than a specific 14-entity domain; *sometimes* is intuitively natural in a situation where someone wears, say, a blue shirt for one or two days every two weeks in perpetuity.

Some might have been more natural than *sometimes* for set size 13 due to adverbial condition participants interpreting the calendars as showing a general shirt-wearing pattern. An existential quantifier is perhaps particularly unnatural in a situation where someone wears a yellow shirt for 13 out of 14 days in perpetuity

vis-à-vis a situation where someone happens to do so for one specific set of 14 days.

I suggest that the narrowing effects in Experiment 2 are due to the fact that all participants saw both nominals and adverbials in Experiment 2, leading them to converge on the domain and therefore rate nominals and adverbials more similarly. The existence of a narrowing effect with *most*/*usually* at 14 alongside narrowing effects for *some*/*sometimes* at 1, 2, and 13 suggests that *most* and *usually* may not differ in their *not all* SIs as strongly as Experiment 1 indicated.

6 Further directions

The results of Experiments 1 and 2 open up several avenues for future research in formal and experimental semantics. On the formal side, one further direction concerns how to model the semantic relationship between *most* and *usually* given their similarity overall but differences with regards to SI. On the experimental side, one could test whether domain restriction played a role in the results by tweaking my paradigm so that the adverbial and nominal stimuli have similar levels of domain specification. Finally, given that numbers can serve as contextual alternatives to *some* (Degen & Tanenhaus 2015) but adverbials and nominals cannot serve as contextual alternatives to one another, future work should seek to establish a precise characterization of what can and cannot count as a contextual alternative.

References

Alstott, J., & M. Jasbi. 2020. Lexicalization of quantificational forces in adverbial and determiner domains. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, 2001–2006.

Bach, E., E. Jelinek, A. Kratzer, & B. Partee. 1995. *Quantification in natural languages*.

Chemla, E., & B. Spector. 2011. Experimental evidence for embedded scalar implicatures. *Journal of semantics* 28. 359–400.

Chierchia, G., D. Fox, & B. Spector. 2012. The grammatical view of scalar implicatures and the relationship between semantics and pragmatics. *Semantics: An international handbook of natural language meaning* 3. 2297–2332.

de Swart, H. 1991. *Adverbs of quantification: A generalized quantifier approach*. University of Groningen dissertation.

Degen, J., & M. K. Tanenhaus. 2015. Processing scalar implicature: A constraint-based approach. *Cognitive science* 39. 667–710.

von Fintel, K. 1994. *Restrictions on quantifier domains*. University of Massachusetts, Amherst dissertation.

Horn, L. 1972. *On the semantic properties of logical operators in English*. University of California, Los Angeles dissertation.

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

Lewis, D. 1975. Adverbs of quantification. In *Papers in philosophical logic*, 5–20. Cambridge University Press.

Papafragou, A., & J. Musolino. 2003. Scalar implicatures: experiments at the semantics-pragmatics interface. *Cognition* 86. 253–282.

Partee, B., 2008. A-quantification and d-quantification: Background. Ms., University of Massachusetts.