A corpus-based study of (non-)exhaustivity in *wh*-questions

A key issue in *wh*-question interpretation regards the distribution of exhaustive (Mention-All, MA) vs. non-exhaustive (Mention-Some, MS) question readings (see (1) and (2)):

(1) Who came to the party?
   a. Who is every person that...? MA
   b. Who is a person that...? #MS

(2) Where can I find coffee?
   a. What is every place that...? MA
   b. What is a place that...? MS

Theories of question interpretation have typically assumed that a MA reading is always appropriate [1,2]. Linguistic factors that have been argued to generate variation in readings include the specific *wh*-word—e.g., *who*-questions are biased for MA, while *where*/*how*-questions are biased for MS [3-4]—and existential (priority) modality—e.g., *can* purportedly licenses MS, as in (2) [5-9]. Recent work [10] tested these judgements in lab-controlled experiments with artificial stimuli and found evidence for some biases, but these biases can be overridden by features of the context like speaker/discourse goals [cf. 3-4,11]. However, there is to date no systematic investigation of *naturally occurring questions* that tests the intuitions reported in the literature. We ask: (Q1) How much does question interpretation vary in natural discourse contexts? Is there indeed a bias for MA? (Q2) Is the distribution of interpretations modulated by linguistic form?

Methods. **Step 1: Naturalistic Stimuli from a Corpus Database.** Using TGrep2 and the Tgrep2 Database Tools [12-14], we extracted all occurrences of *wh*-questions (10,009) from the Switchboard corpus [15] and coded the questions for syntactic structure (e.g., embedded, root), *wh*-word, and presence of modality. To curate stimuli for step 2, we constrained the database to the most frequently discussed cases: root *who*-, *where*-, and *how*-questions. We also removed degree (*How much sugar do you need?*) and identity (*Who is that?*) questions because MS and MA meanings converged, with 335 questions remaining. The distribution of *wh*-word and modality in this database is reported in Table 2. **Step 2: Paraphrase Rating Task.** The remaining cases were divided into 11 lists with occurrence of critical factors roughly proportional to the overall database. Participants (n=385) on Prolific were presented with each question and the 10 preceding lines of dialogue, and asked to rate the likely intended meanings (paraphrases), using a slider task (Fig. 1). Question paraphrases were selected to reflect MS/MA readings: *a* indicates MS, *every* MA, while in *the*-paraphrase the two readings collapse. There was a fourth option (*something else*) in case no other was appropriate. Performance on 6 catch trials functioned as exclusion criterion (n=19).

Results. Questions with highest ratings for *something else* (17%) were excluded because they were rhetorical (see Tab. 1). *The*-paraphrases, where MS=MA, had the highest mean rating (.59), suggesting that only one reading was possible for most cases. Data were analysed using linear mixed effects regression. To investigate the posited MA bias, we compared *every* vs. *a* ratings, as these represent MA and MS (Fig. 2): contrary to the literature, there was no bias for *every* (Q1). However, significant two-way interactions between paraphrase and linguistic form factors partially support reports from the literature (Q2): first, the presence of a modal resulted in higher ratings of *a* ($p<.0001$, Fig. 3) [5-9,10] but not *every*. Second, ratings for *how*-questions resulted in higher *a* than *every* ratings ($p<.04$), confirming [3-4, 10], but not for *where* or *who*-questions.

Conclusion. In contrast to theoretical predictions, we find no bias for MA question readings in naturalistic dialogue (Q1). With respect to (Q2), we find support for some, but not all, observations about the effect of linguistic form on question interpretation reported in the literature. We suggest that MS/MA readings result from reasoning about the speaker's goal in the context, consistent with a constraint-based account [16] on which hearers integrate multiple sources of information to determine meaning. These results also have methodological implications: data hand-selected during theory-building may be biased and not reflect a realistic distribution of meanings [17].

Paraphrase	Example	Mean Rating
every (MA)	*Where have you skied?*	.66
	Where's it all going?	.59
a (MS)	*Where do you like to eat?*	.57
	How would you achieve that?	.51
the (MS=MA)	*Where you going to school?*	.99
	Where do you work?	.99
something else	*Who knows?*	.61
	How can you watch that?	.53

Table 1: For each paraphrase, examples of questions that resulted in high ratings on that paraphrase.

Wh	Modal?	% of Total
who	Yes	2.4%
	No	13.7%
where	Yes	1.2%
	No	27.8%
how	Yes	8.4%
	No	46.6%

Table 2: Joint distribution of *wh*-words and modals in database of 335 root questions.

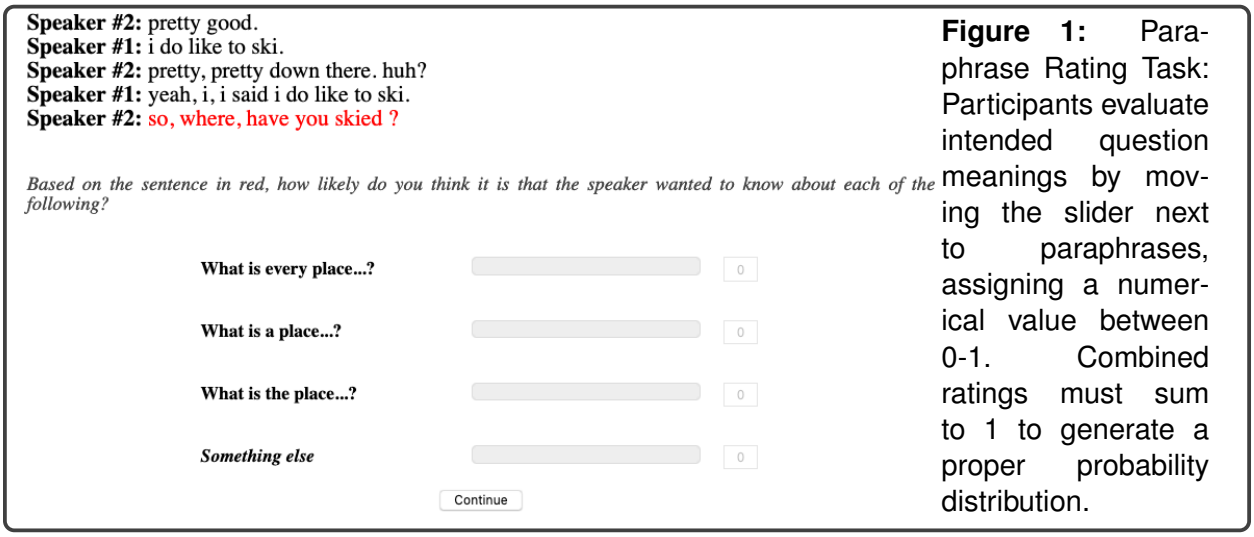

Figure 1: Paraphrase Rating Task: Participants evaluate intended question meanings by moving the slider next to paraphrases, assigning a numerical value between 0-1. Combined ratings must sum to 1 to generate a proper probability distribution.

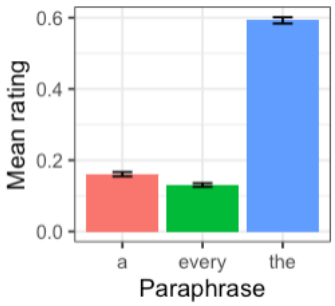

Figure 2: Surprisingly, *every* paraphrases were not preferred over *a* paraphrases.

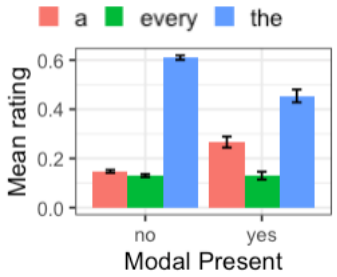

Figure 3: For modal questions, *a* received higher ratings than *every* (in line with [5-9]), but suprisingly not lower in non-modal questions (in contrast to [5-9]).

Figure 4: *A*-paraphrases received higher ratings than *every* for *how* (in line with [3-4]), but surprisingly not lower for *who* (in contrast to [3-4]).

References. [1] Karttunen (1977), [2] Groenendijk & Stokhof (1984), [3] Ginzburg (1995), [4] Asher & Lascarides (1998), [5] George (2011), [6] Nicolae (2013), [7] Fox (2014), [8] Dayal (2016), [9] Xiang (2016), [10] Moyer & Syrett (2019), [11] van Rooij (2003), [12] Rohde (2005), [13] Jaeger (2006), [14] Degen & Jaeger (2011), [15] Godfrey et al. (1992), [16] Degen & Tanenhaus (2019), [17] Degen (2015)