Rates of scalar inferences beyond 'some' – A corpus study

Richard Breheny (University College London), Chao Sun (University College London), Ye Tian (Universite Paris Diderot)
chao.sun.13@ucl.ac.uk

In a large-scale corpus-based web study, Degen (2015) extracted 1363 utterances containing *some*-NPs from the Switchboard corpus. For each utterance, they measured the rate of the scalar inference (SI) from *some* to *some but not all* using a paraphrase task. Their findings showed that around half the time *some* is used, an SI reading is not judged to be available. Little is known about how frequently scalar expressions of different lexical categories give rise to SIs in *real use*. In an inference task, van Tiel et al. (2016) showed that different scalar terms give rise to SIs at different rates. Here, we adopt Degen's paraphrase task using a Twitter corpus we constructed to investigate whether the rates of SI derivation vary to the extent found in van Tiel et al.'s inference task. We do find variability in the rates of SIs across different scalar expressions, but not the same degree of variability found when items are presented out of context in an inference task. A modest amount of this variance could be explained by factors which van Tiel et al. found contributed to the variances in the experimental setting. Our study yields several interesting results, mentioned below.

Collecting a Twitter corpus: We selected 28 out of 43 scalar expressions found in van Tiel et al. (2016). There were 2 quantifiers (e.g. <some, all>), 1 adverb (<sometimes, always>) and 25 adjectives (e.g. <intelligent, brilliant>). For each scale, we extracted tweets containing the weak scalar term - with a minimal length of 30 characters. Then we conducted part-of-speech (POS) tagging on each tweet and used regular expressions to filter out tweets where scalar expressions appear in environments which the inferences are unavailable or less likely to arise (see Table 1).

environment	example
in the scope of negation	I'm not really hungry.
in the scope of conditional antecedents	If the weather was warm, we would have some people over for a small party in our backyard.
in the scope of wh-questions or polar questions	Do you get adequate vitamin D?

Table 1: Environments prohibit the scalar inference

To perform the final exclusion, we conducted a word sense disambiguation task on Amazon Mechanical Turk to obtain human annotation on tweets containing polysemous scalar expressions. Considering <old, ancient> for example, in (1a) the sense of *old* meaning "existing a long time" is on the same scale as the core meaning of *ancient*. However, in (1b) the sense of *old* meaning "previous" was not on the same scale as the strong term. Cases like (b) need to be excluded because in these cases the strong term is not contextually available which make it infelicitous to investigate the rate of SIs. We consulted the Merriam-Webster dictionary and found 20 out of 28 our scalar expressions have at least two meanings.

(1a) I'm in an ***old*** abandoned train station w/ a translator working on the script.

(1b) That means my ***old*** boss has been approaching a breakdown for the last 2 years.

80 M-Turk workers were recruited and each annotated 50 tweets of a particular scalar expression. In total, 4000 tweets were annotated, 200 tweets per scale. We presented workers with a tweet containing the scalar expression, e.g. *warm* (I guess he wants his home to feel ***warm*** and inviting.) and ask them to choose the meaning of *warm* from the following three sense labels: (if none are appropriate, workers can click 'none of the above' option) (a)

having a fairly high temperature; (b) friendly and affectionate; (c) light and bright colors. (a) is the sense that could be understood on the same dimension as the strong term, whereas (b-c) are the relatively common senses listed in the dictionary. Based on our results, we excluded tweets in which weak terms evoke senses that are not on the same scale as strong terms.

Corpus-based paraphrase task: We ran a paraphrase task based on Degen (2015) to measure the frequencies of SIs triggered by the 28 scalar expressions. After the final exclusion, we ended up with 3075 tweets in total. We randomly selected 50 tweets for each scale as the target sentences. On each trial, participants read an utterance containing a scalar expression *X* (the weak term, in red) and a nearly identical utterance, expect that the negation of the stronger term *not Y* (in green) was inserted (Figure 1). Participants were asked to rate on a seven point scale to indicate how similar is the statement with *X but not Y* to the statement with *X*. 550 participants each judged 28 items – one item per scale.

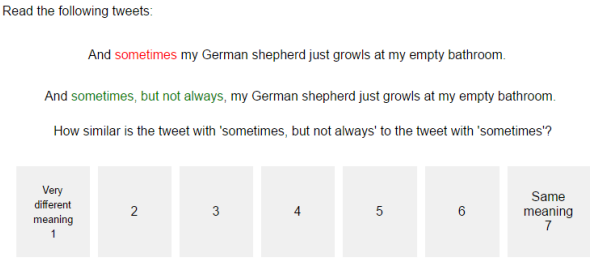

Figure 1 paraphrase task example item

Results: The responses were coded into three categories: low (ratings were 3 or lower), median (ratings were 4), and high (ratings were 5 or above). We considered high ratings as an indicator of SIs being drawn. Inspecting Figure 2, the frequency of SI varies across scalar expressions, from 27% for <adequate, good> to 86% for <sometimes, always>. These results correlated with the results of van Tiel et al. (2016) ($r=0.81$, $p<.001$), suggesting that, to some extent, the results yield from the inference task based on artificial examples could reflect frequencies of SIs triggered in *real use*. However, Levene's test for equality of variances showed that variances of two studies are not equal ($F(1,54)=14.69$, $p<.001$). Visual inspection of Figure 2 suggests that there is less variation on the paraphrase task. In particular, adjective scalar expressions give rise to SIs more frequently in real use. We replicate the result in Degen (2015) for 'some' and note that actual rates of SIs for this item and other terms like, 'possible' and 'allowed' are far lower than rates found on the inference task.

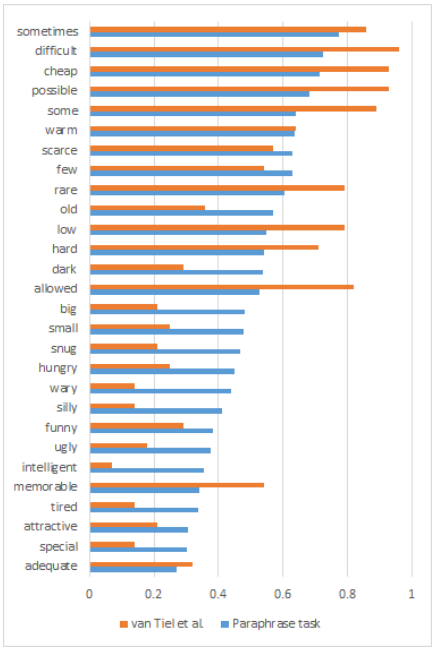

Figure 2 shows the percentage of 'High' ratings for 28 scalar expressions. Percentage of SI responses from van Tiel et al. (2016, Experiment 2) are shown in orange.

The variability displayed in the frequencies of SIs call for an explanation. Multiple linear regression analyses were conducted to predict the frequencies of SIs from possible factors explored in van Tiel et al. (2016), including association strength, grammatical class, word frequencies, semantic relatedness, semantic distance, and boundedness. As van Tiel et al., found with their inference task results, only semantic distance and boundedness are substantial factors. In this case, these factors together accounted for 43% of the variance. Future studies need to explain where the remaining variance comes from.

Reference: [1] Degen, Judith. 2015. *Semantics and Pragmatics* 8(11). 1–55. [2] van Tiel, B. van Miltenburg, E. Zevakhina N. & Geurts, B. (2016), *Journal of Semantics*, 33: 137-175.