How do context and countability modulate scalar implicature strength?

VENKATA S GOVINDARAJAN

1 INTRODUCTION

A core component of the Gricean programme is that language use ought to be treated as a form of rational activity subject to norms and expectations (Grice 1975). This framework has been immensely useful in computing *pragmatic inferences* where listeners infer information beyond what is the literal meaning of an utterance. A classical example of this is the scalar implicature:

(1) a. Mary ate some of the cookies.
b. Mary ate some, but not all, of the cookies.
c. Mary ate all of the cookies.

A speaker who utters (1-a) is generally taken to conversationally implicate (1-b). Most theories treat this inference (and other inferences that follow from a scale) as a form of generalized *default* inference that is a consequence of Gricean maxims of Quantity and Quality(Grice 1975, Horn 1984). The speaker wants to convey as much true information as possible — if Mary had eaten all of the cookies, the speaker would have uttered (1-c) instead of (1-a). The listener in turn, can reason that since the speaker chose not to utter (1-c), Mary didn't eat all of the cookies, giving rise to the scalar inference.

This default form of reasoning, while attractive, makes many simplifying assumptions. Under this reasoning, scalar implicatures are generalized conversation implicatures (**GCI**s) that arise because of the presence of lexicalized items like *some* (rather than special contextual features). Contextual features of the linguistic signal is deemed to play a minor role in cancellation of an implicature. This CONTEXTUAL INVARIANCE assumption follows directly from the HOMOGENEITY assumption - listeners can either derive the inference or not.

By crowdsourcing human judgements on a corpus, Degen (2015) showed that neither assumption holds. She suggests that the strength of scalar inferences from

some to *not all* are highly variable and systematically context-dependent. This raises interesting implications about the natur and theory behind scalar implicature. I use her paper as a starting point for this report, and try to answer two broad questions:

i. What is the relationship between the strength of a scalar implicature and the **frequency of its occurence in varying contexts**? This question tries to tackle the issue of whether annotators might be biased to rate a scalar implicature as stronger if they are more familiar with the form of the utterance.

ii. How does the **countability profile** of *some*-NPs influence the strength of the scalar implicature? Similar to how the partitive construction is shown to lead to higher implicature strength ratings, do count nouns exhibit higher strength ratings than mass nouns?

§2 goes over some of the background literature on scalar implicatures, as well as describing the motivation behind the probabilistic studies of pragmatic phenomena, including scalar implicatures. In §3, I examine the methodology Degen uses to derive a score for scalar implicature strength, and investigate if this strength is related to the contextual diversity of the *some*-NPs. Adding to Degen's work showing how the partitive, determiner strength and discourse accesibility influence scalar implicature strength, I examine if the countability profile of some-NPs influence implicature strength in §4. I conclude this report with remarks on my findings and possible future avenues for research (§5)

2 BACKGROUND

2.1 CLASSICAL THEORETIC APPROACHES

Since the focus of this report is on the lexicalized scalar implicature from *some* to *not all*, let us focus on how this particular scalar implicature has been studied in literature. Most linguistic theorizing of this scalar implicatures analyses it as a form of **Gricean Conversational Implicature** (Grice 1975, Horn 1984, Levinson 2000). This distinguishes them from Particularized Conversational Implicatures, which as Grice (1975) states, are carried through because of special features of their context. GCIs on the other hand, are calculated because of the use of certain forms of words, and assumptions of rational behavior on the part of the speaker and listener.

A walkthrough of the classical reasoning behind the *some* to *not all* implicature was illustrated in (§1), which I will not repeat here. Let's focus on the assumptions that this style of reasoning makes regarding the scalar inference?

1 The scalar inference is treated as a *categorical* phenomenon. The listener can either derive the inference based on the presumption of co-operative behavior and the lexical item *some*, or not. There is no room for the listener to hold degrees of belief in the scalar inference. This is the HOMOGENEITY assumption

2 The only contextual information that can trigger or act as a cue towards the scalar inference is the lexical item *some* itself. This is the CONTEXT INDEPENDENCE assumption.

While there has been acknowledgment of the ability to cancel implicatures (Horn 1984, Levinson 2000), this form of reasoning doesn't predict that elements of the context play a *systematic* role towards the generation or cancellation of the scalar implicature.

2.2 GAME THEORETIC APPROACHES

Recent work in Bayesian pragmatics (Frank & Goodman 2012, Lassiter & Goodman 2013, Rothschild 2013, Goodman & Frank 2016) have provided evidence for treating pragmatic phenomena, such as the scalar implicature from *some* to *not all*, as a form of **Bayesian probabilistic reasoning** on the part of speakers and listeners. This framework has been encapsulated under the heading of *rational speech-act* (RSA) theory.

In the RSA model, a pragmatic listener updates his beliefs about the state of the world using Bayes' rule, given that the speaker chose an utterance over the alternatives. The pragmatic speaker in turn is assumed to be approximately rational and chooses their utterances proportional to their expected utility (based on how much epistemic utility a *literal* listener would derive from an utterance). Thus, the pragmatic listener can infer a degree of belief over the various states of the world conditioned on the speaker's chosen utterance over the alternatives (Goodman & Stuhlmüller 2013, Goodman & Frank 2016).

How does this affect our understanding of scalar inferences, and what predictions does it makes? As described by Degen, the revised assumptions and claims are as follows:

1 Scalar implicatures are **probabilistic** rather than categorical. The probability reflects the hearer's belief that an alternative is more probable than others. This is the strength of a scalar implicature.

2 Scalar implicatures are determined by their context — if hearers update their beliefs based on their reasoning over possible utterances, it stands to reason that **contextual cues** in the linguistic signal would influence their beliefs.

These revised assumptions make a crucial claim — scalar implicatures can now be studied using human judgements over corpora on a continuous scale. This is the main contribution of Degen (2015).

2.3 STRENGTH AND CONTEXTUAL VARIANCE

Degen (2015) performed a corpus and web-based study to test the two assumptions in the classical model described previously — is the *some* to *not all* scalar implicature HOMOGENOUS, and is it CONTEXTUALLY INDEPENDENT? In her study, participants were asked to perform a paraphrase similarity task, which she takes as a proxy for the scalar inference strength. For example, participants were asked to rate the similarity between the following two statements on a 7-point scale (from not at all similar to exactly similar):

(2) a. Mary ate some of the cookies.
b. Mary ate some, but not all, of the cookies.

To investigate the contextual independence assumption, she analyses the effects of 3 contextual features (or *cues*) on the implicature strength — syntactic partitivity of the *some*-NP, determiner strength, and discourse accesibility of the *some*-NP. She finds that not only is there a wide variance in scalar implicature strength contrary to the homogeneity assumption, this variance in the strength of the scalar implicature is highly systematic and probabilistically determined by features in the context of the utterance. Implicature strength is greater on average when *some* occurs with the partitive, when its usage as a determiner is strong, and when the *some*-NP is relatively discourse accessible.

Using her dataset and models as a starting point, I will start by investigating the relationship between implicature strength and the **contextual diversity** of the *some*-NP.

3 CONTEXTUAL DIVERSITY AND STRENGTH

Degen's (2015) annotation protocol asked annotators to rate the similarity between a *some* utterance with a paraphrase that substituted *some* with *some but not all*, like in (3). If two statements are judged to be highly similar (on the 7-point Likert scale), she concludes that the scalar implicature is encoded strongly in the *some* utterance.

(3) a. I like, I like to read *some* of the philosophy stuff.
b. I like, I like to read *some, but not all*, of the philosophy stuff.

This paraphrase task is a novel measurement of scalar implicature strength, and Degen discussed how certain effects of context(monotonicity), shifts in Question Under Discussion, and annotator variability could affect its interpretability. However, this raises the question — if an annotator says that the *some* utterance and its *some, but not all* paraphrase are highly similar (or dissimilar), does that mean that the scalar implicature is necessarily strong (or weak)? For example, might annotators choose to rate two statements as dissimilar because of their unfamiliarity with the *some*-NP and its usage in a *some, but not all* construction? Narrowing down the question to one of **contextual diversity**, does a *some* to *some, but not all* scalar implicature need to occur in a wide variety of contexts to be considered a strong implicature?

Data Analysis To investigate whether the contextual diversity of the object of a *some-not all* implicature modulates its strength, I performed a simple analysis of the *some*-NPs from Degen (2015). The *some*-NPs were extracted from the dependency parse of each sentence (Qi et al. 2020) by finding those NPs that were immediately governed by (or that governed) *some* (White et al. 2016, Zhang, Rudinger & Van Durme 2017). To simplify the analysis, only one-word NPs were chosen for further analysis. Each NP was lemmatized, and the frequency of occurence of the lemma in the corpus was calculated. In total, **509** unique lemmas were found in **1291** unique sentences. The frequency of a lemma is defined as the number of unique sentences which contain the lemma in question in our corpus. I take this to be a proxy for the **contextual diversity** of a lemma.

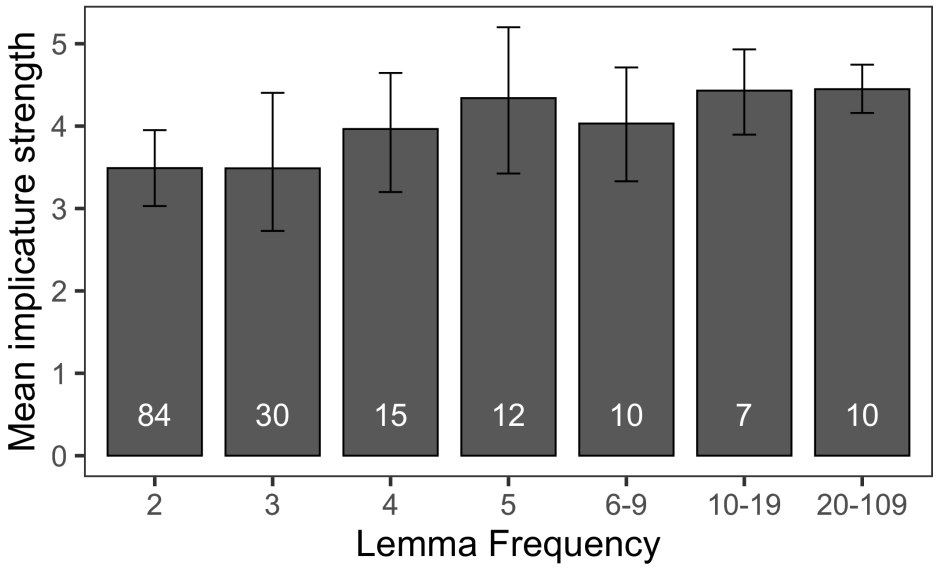

FIGURE 1 Mean implicature strength versus lemma frequency. The numbers inside each bar indicate the number of unique lemmas.

Results Each sentence was rated by 10 different annotators in the dataset — all results henceforth report the average implicature strength assigned by all annotators. Figure 1 shows the average implicature strength rating against the lemma frequency (the number inside each bar indicates how many unique lemmas had that frequency — for example 15 unique lemmas occured in 4 different sentences and had an average implicature strength of around 4). While there is a disparity in distribution of the lemmas which is to be expected (highly frequent lemmas occur are fewer), we also notice that there is a slight tendency for highly frequent lemmas to have a higher mean implicature strength.

For an objective measure for whether lemma frequency is correlated with implicature strength, we use the Spearman Rank correlation test. A non-parametric test was chosen since Figure 2 shows that the mean by-item implicature strength distribution is not normal, and a Shaprio-Wilk's test (Shapiro & Wilk 1965) verifies this observation(W=0.96, p<2.2e-16). The Spearman's correlation coefficient between lemma frequency and implicature strength (ρ) is 0.51 (S=651, p<0.05). The Pearson's correlation coefficient was also calculated, and found to be 0.44 (t=2.09, df=18, p=0.051).

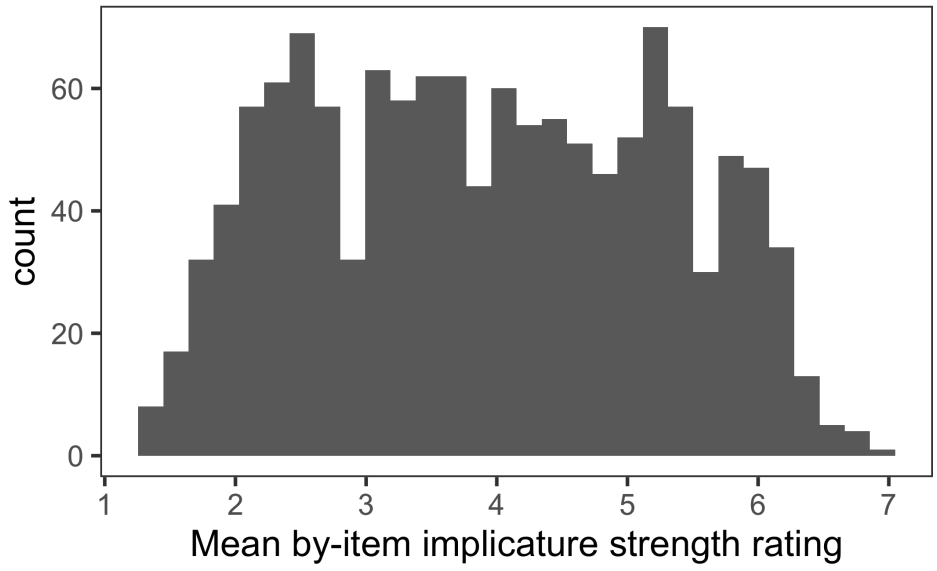

FIGURE 2 Average by-item implicature strength ratings in our restricted dataset.

Discussion We started with the question of whether paraphrase similarity was a reliable indicator of scalar implicature strength, or whether external factors might influence the similarity rating **independent** of the scalar implicature strength. I chose to investigate one facet of this interaction — Is the implicature strength of the lemma of the *some*-NP correlated with its frequency of occurence in unique contexts? The results above show that a significant correlation does exist. **Highly frequent lemmas are correlated with a higher implicature strength rating**. For example, given below are sentences which have the 2 most frequent lemmas (*people* and *thing*) along with their mean implicature raing and their corresponding paraphrase:

(4) a. I mean, **some people** sing, **5.5**
b. I mean, **some, but not all**, people sing,

(5) a. i, i'm glad, to see **some of these things**, i think . **4.7**
b. i, i'm glad, to see **some, but not all of these** things, i think .

Now consider some ratings for the lemmas *noise* whose frequency was 2, and *information*, whose frequency was 7:

(6) a. and i was picking up, uh, **some background noise**,. **2.1**
b. ?and i was picking up, uh, **some, but not all, background noise**.

(7) a. i got **some interesting information** about crawfish **2.5**
b. ?i got **some, but not all, interesting information** about crawfish

One potential reason for less frequent lemmas having lower implicature strength ratings could be that the paraphrases tend to be more **unacceptable** like in (6-b) and (7-b). While it might appear that these unacceptable sentences are more likely to occur with low-strength determiner uses of *some*, determiner strength by itself wasn't found to be indicative of the unacceptability of certain statements (as discussed in §2.3.2 of Degen (2015)), and there isn't a significant correlation between lemma frequency and determiner strength either (ρ=-0.35, p=0.134).

It might be tempting to attribute the lower ratings for less frequent lemmas to annotators being more likely to give high ratings towards lemmas they are more familiar with. However, the lemma frequencies used here are not indicative of real-word distributions, or the entire range of contexts in which they may occur. For instance, the lemma *music* occurs only twice in the dataset, and yet has high implicature strength ratings:

(8) e-, even in some families some people talk a little bit different. **5.3**

Further analysis would be necessary to verify if the observed correlation is merely an artifact of the datset under study, or indicative of a possible factor to consider when deriving scalar implicature strength based on a similarity score.

4 MASS-COUNT DISTINCTION

As already discussed, Degen (2015) found that the partitive, determiner strength, and discourse accessibility provide reliable cues for scalar implicature strength. In this section, I wish to investigate if the **countability** profile of *some*-NPs are also a significant contextual cue in determining scalar implicature strength.

The classical mass-count distinction is a morphosyntactic distinction that is generally taken to have semantic significance (Moltmann 2020). Mass nouns (such as rice, wood and water) are generally found to differ from count nouns (like apples, chairs and carpets) in several syntactic and and semantic dimensions. For instance mass nouns generally lack a plural(**I ate rices*), cannot take cardinal or ordinal numerals(**two waters*), and also disallow singular quantifiers (**every water*).

The semantics of mass and count nouns are also different, although there are differing views on how to interpret the distinction between the two (Moltmann 1998, Link 2008, Champollion & Krifka 2016, Champollion 2010, Krifka 1989, Moltmann 2020).

Recent work has questioned the assumption underlying the countable - non-countable contrast (Grimm & Wahlang 2020, Grimm 2018) — is it ontologically based or is there a scale of countability between prototypically countable nouns and non-countable (mass) nouns? These questions are relevant for studying scalar implicatures in light of the varying semantics of mass nouns:

(9) a. I ate some rice for lunch.
b. ?I ate some, but not all, rice for lunch.

(9-a) is not similar in meaning to (9-b). Moreover, (9-b) as a sentence is, if not ungrammatical, possibly unacceptable (there is, however, a possible reading where the speaker is saying that they didn't eat *only* rice for lunch. I ignore this reading). The incongruity of the *some but not all* utterance disappears when the partitive is used:

(10) a. I ate some of the rice for lunch.
b. I ate some, but not all, of the rice for lunch.

This brings up an interesting question — could the lower implicature strength ratings with sentences that have low strength determiners or non-partitive clauses be explained by the countability profile of their *some*-NPs? Thus, the questions I wish to ask are as follows:

1 Does the countability profile of the object NP under *some* have an impact on the implicature strength ratings?

2 If it does, is it wholly independent from the effects of determiner strength and partitivity?

Data Analysis Similar to the context analysis in §3, the relevant NP lemmas were extracted from the dependency parse of sentences in the dataset. Once again, this analysis is restricted to one word NPs. The CELEX database (R.H., Piepenbrock & Gulikers 1996) is a lexical database of English that uses a classical countability classification — countable, uncountable or both. Using this classification method, we add two cues corresponding to each lemma in the dataset — MASS and COUNT, each

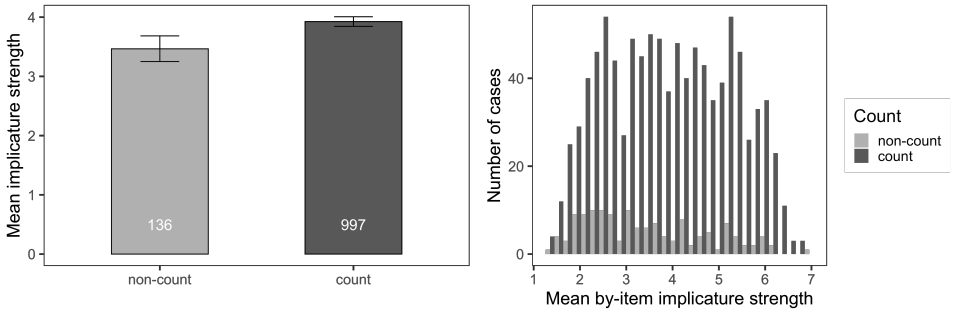

FIGURE 3 Mean implicature strength ratings (left) and distribution of mean by-item ratings (right) for non-count and count some-NPs.

of which has either the values *yes* or *no*. COUNT and MASS are treated separately as there are some lemmas that can function as mass or count depending on the context and their word sense, and because mass and count nouns are known to undergo shifts to their counterparts in taxonomic and packaged readings (Moltmann 2020). Sentences for which the lemmas were not found in the CELEX database were discarded, thus providing us with **1133 sentences** to perform further analysis on.

To verify if MASS and COUNT are reliable cues that influence the scalar implicature strength determined by annotators, I report the results of the same linear mixed effects regression models (Baayen, Davidson & Bates 2008) used in Degen (2015), with the new countability columns included as fixed effects in the regression. To recap, the mixed-effects linear regression model predicts the implicature strength from fixed effects of interest (the cues under investigation) along with random by-item intercepts, random by-participant intercepts, and random by-participant slopes for all fixed effects. MASS and COUNT were added as fixed effects (and allowed to interact with the PARTITIVE and DETERMINER fixed effects), and random by-participant slopes were added for each as well. Three types of models were considered — the basic model only considered random effects, the intermediate model only considered fixed effects, while the final model considered both fixed and random effects. The final model from Degen (2015) was used as a baseline for comparing the impact of adding COUNT and MASS as fixed effects and random by-participant effects.

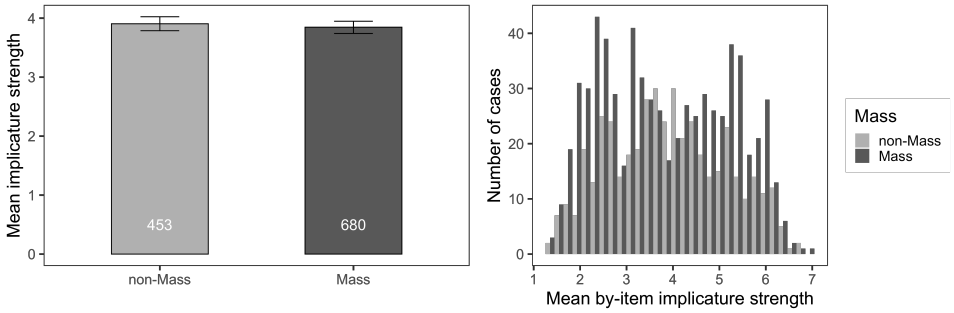

FIGURE 4 Mean implicature strength ratings (left) and distribution of mean by-item ratings (right) for non-mass and mass some-NPs.

	Basic Model	Intermediate Model	Final Model
Degen 2015	0.099	0.264	0.457
Degen 2015+Mass+Count	0.099	0.271	0.461

TABLE 1 Proportion of variance (Conditional R^2) explained by each model.

Results *count* items have an average strength rating of 3.46, and they are higher than *non-count* items which have an average rating of 3.92 (Figure 3). Similarly, *non-mass* items (which could be considered analogous to count items) have a slightly higher average implicature strength rating (3.90) than *mass* items (3.85) (Figure 4). However, to conclude that either MASS or COUNT are reliable cues for scalar implicature strength, we need to turn to the performance and coefficients of our linear mixed-effects model.

Table 1 reports the conditional R^2 of various models on our restricted dataset. The conditional R^2 measures the percentage of variance explained by the fixed and random effects of our model. We observe that while the MASS and COUNT fixed effects do help in boosting the performance of our model, the improvement is very small. This could mean that the mass-count distinction does not play a major role in determining scalar implicature strength. However this could also be an indication that our method for incorporating the countability profile of NPs for analysis needs to be improved. The model coefficients for the MASS and COUNT factors in Table 2 indicate that the cues as incorporated currently do not add any relevant information towards determining implicature strength. The slopes of the individual fixed effects

	Coef β	SE(β)	**t**	p
Intercept	3.94	0.06	63.2	**<.0001**
Partitive	0.90	0.12	7.4	**<.0001**
Strength	−0.53	0.06	−9.1	**<.0001**
Mass	0.09	0.08	1.2	>0.25
Count	0.08	0.14	0.6	>0.55
Partitive:Mass	0.19	0.20	1.0	>0.33
Strength:Mass	−0.14	0.10	−1.5	>0.15
Partitive:Count	−0.50	0.44	−1.1	>0.26
Strength:Count	0.16	0.18	0.9	>0.35
Mass:Count	0.49	0.28	1.8	>0.08

TABLE 2 Model coefficients for our final model with MASS and COUNT added as fixed effects

are small and insignificant (as suggested by the p-values). While there does seem to be some interaction between our new fixed effects with the previous ones, none of them are potentially significant.

Discussion Incorporating MASS and COUNT as factors from the CELEX database did not help in improving the performance of our model significantly, nor do they seem to be reliable cues towards scalar implicature strength one way or another. While I had hypothesized that one reason could be that the partitive or determiner strength could interact with the countability profile of NPs under *some*, our model coefficients show this not to be the case.

Future work needs to look into whether expanding the definition of countability to include more information from CELEX, such as *singularia tantum*, which is encoded separately from *uncountable* nouns in the database (presumably to distinguish between uncountable nouns that do have a plural form versus those that don't). In addition, *pluralia tantum* information from the database needs to be incorporated as well. These lemmas are the most interesting for further study since they don't fall neatly into classical countability categories. Consider the following examples from the dataset of *pluralia tantum* NPs:

(11) a. and then with that i try and handle, you know, **some of the clothes** that the girls need and things like that, **4.4**

b. i got **some interesting information** about crawfish **2.5**

c. uh, she can't eat **some cream cheese**. **3.2**

d. i like **some country music**. **6.9**

The *tantum* NPs in the examples above showcase a wide range of implicature strengths. Future work needs to look into whether incorporating countability profile of NPs as a real-valued score between prototypical count NPs, and prototypical mass NPs provide more reliable cues towards the scalar implicature (Grimm 2018).

5 Conclusion

This report adds to the growing body of work showing that probabilistic models and theories of scalar implicatures (and other pragmatic phenomena) are fruitful areas of research which translate well to human judgement and statistical studies. I have tried to answer two major questions — is the measure of scalar implicature strength related to the contextual diversity of the *some*-NP, and does the countability profile of the *some*-NP provide a reliable cue towards scalar implicature strength?

A significant positive correlation was found between the frequency of the *some*-lemma in the dataset and its implicature strength.One plausible explanation for the lower ratings of less frequent lemmas in the dataset is the general unacceptability/weirdness of the paraphrase, which may cue annotators to rate them as less similar. Future work can look into whether the addition of an additional question where annotators are asked if the paraphrase is ungrammatical and/or unacceptable influences their similarity ratings (White & Rawlins 2016).

No significant interaction was found between the countability profile of *some*-NPs and implicature strength. The lack of a reliable effect suggests that future work needs to look into better methods of calculating and incorporating countability profiles into a probabilistic model of scalar implicatures.

REFERENCES

Baayen, R Harald, Douglas J Davidson & Douglas M Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language* 59(4). 390–412. DOI: 10.1016/j.jml.2007.12.005.

Champollion, Lucas. 2010. *Parts of a whole: Distributivity as a bridge between aspect and measurement*. University of Pennsylvania dissertation. URL: https://repository.upenn.edu/edissertations/958.

Champollion, Lucas & Manfred Krifka. 2016. Mereology. In Maria Aloni & PaulEditors Dekker (eds.), *The Cambridge Handbook of Formal Semantics* (Cambridge Handbooks in Language and Linguistics), 369–388. Cambridge University Press. DOI: 10.1017/CBO9781139236157.014.

Degen, Judith. 2015. Investigating the distribution of some but not all implicatures using corpora and web-based methods. *Semantics and Pragmatics* 8(11). 1–55. DOI: 10.3765/sp.8.11.

Frank, Michael C. & Noah D. Goodman. 2012. Predicting Pragmatic Reasoning in Language Games. *Science* 336(6084). Publisher: American Association for the Advancement of Science Section: Brevia, 998–998. DOI: 10.1126/science.1218633. URL: https://science.sciencemag.org/content/336/6084/998.

Goodman, Noah D & Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science* 5(1). Publisher: Wiley Online Library, 173–184. DOI: 10.1111/tops.12007.

Goodman, Noah D. & Michael C. Frank. 2016. Pragmatic Language Interpretation as Probabilistic Inference. *Trends in Cognitive Sciences* 20. 818–829. DOI: 10.1016/j.tics.2016.08.005.

Grice, Herbert P. 1975. Logic and conversation. In *Speech acts*, 41–58. Brill. URL: https://doi.org/10.1163/9789004368811_003.

Grimm, Scott. 2018. Grammatical Number and the Scale of Individuation. *Language* 94(3). 527–574. DOI: 10.1353/lan.2018.0035. URL: https://muse.jhu.edu/article/702685.

Grimm, Scott & Aeshaan Wahlang. 2020. Determining Countability Classes.

Horn, Laurence. 1984. Towards a new taxonomy for pragmatic inference: Q-and R-based implicature. *Meaning, form and use in context*. Publisher: Georgetown University Press.

Krifka, Manfred. 1989. Nominal reference, temporal constitution and quantification in event semantics. *Semantics and contextual expression* 75. 115.

Lassiter, Daniel & Noah D. Goodman. 2013. Context, scale structure, and statistics in the interpretation of positive-form adjectives. *Semantics and Linguistic Theory* 23(0). Number: 0, 587–610. DOI: 10.3765/salt.v23i0.2658. URL: http://journals.linguisticsociety.org/proceedings/index.php/SALT/article/view/2658.

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

Link, Godehard. 2008. The Logical Analysis of Plurals and Mass Terms: A Lattice□theoretical Approach. In Barbara H. Partee & Paul Portner (eds.), *Formal Semantics*, 127–146. John Wiley & Sons, Ltd. DOI: 10.1002/9780470758335.ch4. URL: https://onlinelibrary.wiley.com/doi/10.1002/9780470758335.ch4.

Moltmann, Friederike et al. 1997. *Parts and wholes in semantics*. OUP USA.

Moltmann, Friederike. 1998. Part structures, integrity, and the mass-count distinction. *Synthese*. 75–111. DOI: 10.1023/A:1005046308299.

Moltmann, Friederike. 2020. Introduction to Mass and Count in Linguistics, Philosophy, and Cognitive Science. LingBuzz: 005164.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton & Christopher D. Manning. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. URL: https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

R.H., Baayen, R. Piepenbrock & L. Gulikers. 1996. CELEX2. Linguistic Data Consortium, Philadelphia, PA. URL: https://catalog.ldc.upenn.edu/LDC96L14.

Rothschild, Daniel. 2013. Game Theory and Scalar Implicatures. *Philosophical Perspectives* 27. 438–478. DOI: 10.1111/phpe.12024.

Schuster, Sebastian, Yuxing Chen & Judith Degen. 2020. Harnessing the linguistic signal to predict scalar inferences. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5387–5403. Online: Association for Computational Linguistics. URL: https://www.aclweb.org/anthology/2020.acl-main.479.

Shapiro, S. S. & M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52(3). 591–611. DOI: 10.1093/biomet/52.3-4.591. URL: https://academic.oup.com/biomet/article-pdf/52/3-4/591/962907/52-3-4-591.pdf.

White, Aaron Steven & Kyle Rawlins. 2016. A computational model of S-selection. *Semantics and Linguistic Theory* 26. 641–663. DOI: 10.3765/SALT.V26I0.3819.

White, Aaron Steven, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins & Benjamin Van Durme. 2016. Universal Decompositional Semantics on Universal Dependencies. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 1713–1723. Austin, Texas: Association for Computational Linguistics. URL: https://aclweb.org/anthology/D16-1177.

Zhang, Sheng, Rachel Rudinger & Ben Van Durme. 2017. An Evaluation of PredPatt and Open IE via Stage 1 Semantic Role Labeling. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*. Montpellier, France. URL: https://www.aclweb.org/anthology/W17-6944/.