Some but not all **speakers** ***sometimes but not always*** **derive scalar implicatures**

Sonia Ramotowska (University of Amsterdam), Paul Marty (University of Malta), Leendert Van Maanen (Utrecht University), Yasutada Sudo (University College London)

Background. The tendency to derive Scalar Implicatures (SIs) drastically varies between individuals [1]. Results from truth-value judgement (TVJ) tasks demonstrate this fact: sentences like (1) *Some eagles are birds* – true on their literal reading (i.e., *at least some*) but false on their pragmatic, with-SI reading (i.e., *some but not all*) – are systematically accepted by some participants but rejected by others. The question of what factors drive these differences is considered crucial to understanding the mechanisms involved in SIs and currently at the center of numerous discussions. To date, there is no agreement on how to quantify individual differences in SI rates and what approach should be used when distinguishing literal (LR) from pragmatic responders (PR). Recently, the findings that different scalar terms have different SI potentials [2], which can be modulated by certain linguistic features of the context [3], have increased this long-standing challenge. Here, we report on a TVJ task investigating variability in rates of upper-bounding (UB, negating *all*) and lower-bounding (UB, introducing *some*) SIs associated with the ⟨some, all⟩ scale [4] and show how a hierarchical Bayesian modelling approach can be used to quantify subjects' preferences across linguistic contexts and test different hypotheses about the distributions of LR and PR groups.

Experiment. The material, task and design are built on [5]. Subjects read categorical English sentences like (1), presented one word at a time on the screen, and provided TVJs ('True' or 'False' responses). They were not given specific instructions on how to interpret the sentences. Test sentences were positive SOME sentences and negative SOME NOT and NOT ALL sentences with the potential to give rise to LB and UB SIs, respectively (see Fig.1). Each survey included 15 trials per sentence type (5 targets and 10 unambiguously true/false controls) and 54 filler trials. 95 adult native speakers of English (63% female; mean age: 30.6 years) participated in the study.

Modelling. We developed a series of beta-binomial hierarchical Bayesian models with classifications to two latent groups (LG: LR and PR) to TVJ responses. Among other things, this modelling framework allows us to explicitly model different prior response distributions of the LR and PR groups and to account for probabilistic modulations of SI rates depending on the linguistic environment [3]. The model assumes $a_{lit} \sim$ Unif$(n/2, n)$ distribution for LR and $a_{prag} \sim$ Unif$(0, n/2)$ for PR (n = 5, see Fig.2). The models Basic 1 and 2 (Fig.2, dashed lines) assume that all participants belong to either LR or PR group for all SIs types. Thus, in the Basic models, we have $a_{ji} = a_{lit}$ or a_{prag}. The LG Mix models with latent group classification (Fig.2, solid lines) assume that each participant i belongs to one or the other group for each implicature type with a probability $z_{ik} \sim$ Bernoulli(q_k). We fit the three LG Mix models with different numbers of classification into groups. For $k = 1$, participants were classified once for both SI types (q_1). For $k = 2$ (LG SI), they were classified separately for UB (q_{ub}) and LB (q_{lb}) SIs. For $k = 3$, they were classified separately for each SI type (q_{3j}). The number of literal 'True'-responses to test sentences in n trials was predicted via the Binomial model with probability p_{ji} for the j-ty implicature and i-ty participant: $p_{ji} = 0 + \beta_{0i} \times x_{some} + \beta_{1i} \times x_{somenot} + \beta_{2i} \times x_{notall}$. Each β_{ji} was sampled from Beta(a_{zji}, b_{zji}) distribution, where a_{zji} depends on group classification and $b_{zji} = n - a_{zji}$.

Results. The LG SI model had the lowest DIC = 925, hence the best fit to the data (see the model fit in Fig.3). For this model, almost all participants were more likely classified as LR for UB SIs (LR group mean posterior $q_{ub} = 0.86$) and as PR for LB SI (PR group mean posterior $q_{lb} = 0.34$; compare blue vs. pink diamond shapes in Fig.4). Our results also show inter-individual differences in the choice of reading (Fig.4 blue vs. pink points): while a vast majority of the subjects differentiated UB and LB SIs, a few of them consistently gave literal responses for both SI types.

Discussion. First, our results show that the LB SI associated with the ⟨some, all⟩-scale were more likely to be derived than their UB counterparts. This finding shows a novel aspect of linguistic variability in SI derivation and provide further evidence that SI strength is modulated within individuals by certain linguistic features, here the absence vs. presence of negation. Second, we presented a rigorous approach to quantify individual differences in SI derivation. The differences observed in the probability of being classified as LR or PR speaks in favour of a probabilistic approach to SIs of the sort advocated in [3]. We will discuss how this approach can be used in future work to test novel hypotheses about processing differences between LR and PR groups.

Condition	Example sentence with their SI	Type of SI
SOME	Some eagles are birds. *SI: Not all eagles are birds*	Upper-bounding (UB)
SOME NOT	Some mosquitoes are not mammals. *SI: Some mosquitoes are mammals*	Lower-bounding (LB)
NOT ALL	Not all woodpeckers are insects. *SI: Some woodpeckers are insects*	Lower-bounding (LB)

Figure 1: Example test sentences in the SOME, SOME NOT and NOT ALL conditions with their corresponding SI. A 'False' response is indicative of a pragmatic (with-SI) response.

Table 1: Model comparison. Resp indicates the assumed responder group and No. q_ks the number of classifications.

Model	Resp.	No. q_ks	DIC	pD	PP
Basic 1	LR	0	1059	262	0.53
Basic 2	PR	0	1076	271	0.02
LG	Mix	1	958	242	0.49
LG SI	Mix	2	**925**	242	0.49
LG	Mix	3	1024	318	0.50

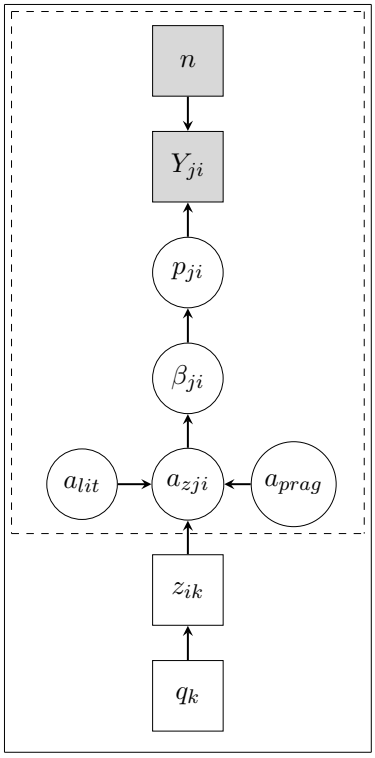

Figure 2: Graphical representation of the Basic models (solid lines) and Latent Group (LG) models (dashed lines).

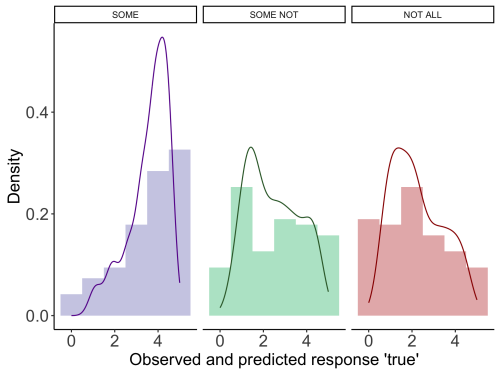

Figure 3: By-subject means of literal responses observed (histogram) vs. predicted by the best model (line)

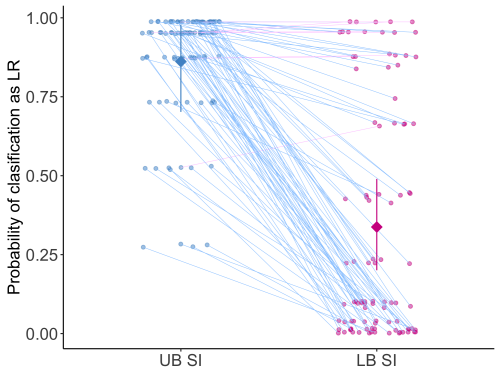

Figure 4: By-subject probability of being classified as LR ($z_{ik=2}$) with mean posteriors q_{lb} and q_{ub} and their 95% CIs. Lines show within-subject consistency in response for UB and LB SIs (less consistent subjects have steeper lines).

Selected references [1] Khorsheed, A. & Gotzner, N. (2023). *Front. Commun., 8:1187970*. [2] van Tiel, B., van Miltenburg, E. Zevakhina, N. & Geurts, B. (2016). *J. Semant., 33(1)*. [3] Degen, J. (2015). *Semantics and Pragmatics, 8:11* [4] Horn, L. (1989). *A natural history of negation*. [5] Bott, L. & Noveck, I.A. (2004). *J. Mem. Lang., 51(1)*