Generics are puzzling.
Can language models find the missing piece?

Gustavo Cilleruelo Calderón **Emily Allaway** **Barry Haddow** **Alexandra Birch**
School of Informatics, University of Edinburgh
g.cilleruelo-calderon@sms.ed.ac.uk {emily.allaway, bhaddow, a.birch}@ed.ac.uk

Humans use generalisations to abstract away from particular objects, events or facts and convey regularities about the world. One way of expressing generalisations in language is through quantified sentences, such as *most insects are nocturnal* or *some mosquitoes have white stripes*, which explicitely express statistical claims about the members of a kind that share the predicated property.

In this work we study generic sentences, such as *insects have six legs* or *mosquitoes carry malaria*, which express generalisations without explicit quantification. These two generic sentences are acceptable in many contexts, but the quantifications they convey are widely different: almost all insects have six legs, but fewer than 1% of mosquitoes carry malaria.

Even as generics seem to express inconsistent quantifications, they are at the heart of communication and dissemination in science (DeJesus et al., 2019; Bowker, 2022), medical research (Peters et al., 2024), and politics (Novoa et al., 2023). Furthermore, in the social realm generics serve as linguistic vehicle for social essentialism (Rhodes et al., 2012) and stereotyping (Leslie, 2017; Bosse, 2022).

The nature of generics and their importance in communication has led to extensive literature on the semantics of generic sentences (e.g., Carlson, 1977b; Cohen, 1999; Leslie, 2008; Liebesman, 2011; Sterken, 2015; Nickel, 2016; Tessler and Goodman, 2016; Stovall, 2019; Nguyen, 2020; Bosse, 2021; Kirkpatrick, 2023). However, many open questions remain, such as how they relate to quantifiers and the degree to which generics are context sensitive.

In this work we introduce a dataset (ConGen) and a metric (p-acceptability) to explore the implicit quantification, context-sensitivity and stereotypes in relation to generics using language models.

The dataset: ConGen. Most existing datasets of generics are synthetic, often derived from knowledge bases or generated by language models (Bhakthavatsalam et al., 2020; Allaway et al., 2024). Since the examples in these datasets are machine-generated and/or lack a context in which they might be uttered, there is no guarantee that they represent how speakers actually use generics.

To solve this problem, we build ConGen, a manually annotated dataset of 2873 naturally occurring generic and quantified sentences with their respective contexts.

The metric: p-acceptability. We build a metric to judge whether the generic or the quantified uttering of a sentence is more acceptable. To do this, we use the surprisal of a language model: given a set of quantifiers (and the generic), we take the most acceptable for a given sentence (i.e. *tigers have stripes*) as the quantifier that makes the tokens after the verb have less surprisal. Intuitively, the p-acceptable quantifier is the one that makes it easier to predict the property of the generalization for the language model.

More formally, let Q be a set of candidate quantifiers, s a bare plural generic and θ a language model. We construct $\{q + s \mid q \in Q\}$ the set of variations of s. We call q the p-acceptable quantifier for s if $q+s$ is the sentence with the lowest surprisal of the property tokens (i.e. tokens after the verb):

$$\text{p-acceptable}(s; Q, \theta) := \text{argmin}_{q\in Q} S_p(q + s; \theta)$$

where S_p is the surprisal of the property tokens

$$S_p(s; \theta) := -\frac{1}{|P|}\sum_{i\in P} \log p_\theta(t_i|t_{<i})$$

with P is the set of indices of the property tokens and t_i the tokens in sentence s.

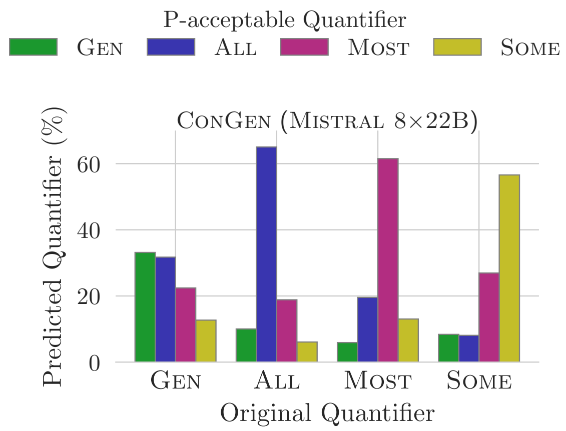

Figure 1: P-acceptability reconstructs semantic intuitions on quantification.

Experiments. We validate this metric by showing that it recovers the expected dynamics of quantifiers (*all*, *most*, *some*) and the generic in ConGen. Note how in Figure 1, *some* is confused by *most*, *most* is confused by *all* and *some* and *all* is confused with *most* but not *some*. For the generic, it gets confused with all quantifiers, as we would expect since generics can have widely different quantificational strengths.

We argue that this metric combined with a strong language model is expressive enough to investigate the implicit quantification of generic sentences, which could be understood as what quantifications speakers convey pragmatically when they use generics instead of explicit quantifiers. We design three experiments: on implicit quantification, on context-sensitivity and on stereotypes and striking generics.

Implicit quantification. We take the generic sentences in ConGen and find the p-acceptable quantifier for them, excluding the generic itself as an option ($Q = \{\text{all}, \text{most}, \text{some}\}$). Intuitively, we select the quantifier that makes it easier for the model, given the subject and the verb, to predict the object.

We find that around 80% of generics are implicitly quantified by *all* and *most*, that is, they are *majority generics*. Nevertheless, across two datasets and for all the models we test, between 18% and 23% of sentences get implicitly quantified with *some*: they are *weak generics*, they express a weak or existential quantificational strength.

Context-sensitivity. For generics and quantified sentences in ConGen, we examine how the accuracy of p-acceptability (the percentage of correct predictions) changes as we calculate the p-acceptable quantifiers by adding left-context tokens. We observe that while for quantified sentences the accuracy is quite stable as the context increases, in generics we see a big improvement, from 35% to 50%, as we gradually add the first 20 context tokens. This suggests generics are context sensitive (the model uses information in the context to better predict the generic) in some distinctive sense to determiner quantifiers.

Stereotypes. We compile a small collection of stereotypes, both positive and negative. We find that for negative (striking) stereotypes the most predicted quantifier is *all*, whereas for positive stereotypes is *most*. This is congruent with studies in psycholinguistics that show that when dangerous properties are predicated, humans perceive them as more prevalent (Cimpian et al., 2010).

We also test the hypothesis that the way the stereotype is phrased affects the quantificational strength it expresses (Leslie, 2017). Each sentence gets 3 paraphrases: bare plural (Ks are F), singular + 'people' (K people are F) and 'people who are' + singular (people who are K are F). We find that the 'people who are' + singular paraphrase changes the most predicted quantifier by p-acceptability from *all* to *some* for negative stereotypes, suggesting the expressed quantification is weaker.

References

Emily Allaway, Chandra Bhagavatula, Jena D. Hwang, Kathleen McKeown, and Sarah-Jane Leslie. 2024. Exceptions, Instantiations, and Overgeneralization: Insights into How Language Models Process Generics. *Computational Linguistics*, pages 1–60.

Sumithra Bhakthavatsalam, Chloe Anastasiades, and Peter Clark. 2020. Genericskb: A knowledge base of generic statements. *CoRR*, abs/2005.00660.

Anne Bosse. 2021. Generics: Some (non) specifics. *Synthese*, (5-6):14383–14401.

Anne Bosse. 2022. Stereotyping and generics. *Inquiry: An Interdisciplinary Journal of Philosophy*, pages 1–17.

Mark Bowker. 2022. A problem for generic generalisations in scientific communication. *Journal of Applied Philosophy*, 39(5):1002–1017.

Greg N. Carlson, editor. 1977b. *Reference to Kinds in English*.

Andrei Cimpian, Amanda C Brandone, and Susan A Gelman. 2010. Generic statements require little evidence for acceptance but have powerful implications. *Cognitive Science*, 34(8):1452–1482.

Ariel Cohen. 1999. Generics, frequency adverbs, and probability. *Linguistics and Philosophy*, 22(3):221–253.

Jasmine M. DeJesus, Maureen A. Callanan, Graciela Solis, and Susan A. Gelman. 2019. Generic language in scientific communication. *Proceedings of the National Academy of Sciences*, 116(37):18370–18377. Contributed by Susan A. Gelman, July 20, 2019 (sent for review October 15, 2018; reviewed by Ellen M. Markman and Douglas L. Medin).

James Ravi Kirkpatrick. 2023. The dynamics of generics. *Journal of Semantics*, 40(4):523–548.

Sarah-Jane Leslie. 2008. Generics: Cognition and acquisition. *Philosophical Review*, 117(1).

Sarah-Jane Leslie. 2017. The original sin of cognition: Fear prejudice, and generalization. *Journal of Philosophy*, 114(8):393–421.

David Liebesman. 2011. Simple generics. *Noûs*, 45(3):409–442.

Anthony Nguyen. 2020. The radical account of bare plural generics. *Philosophical Studies*, 177(5):1303–1331.

Bernhard Nickel. 2016. *Between Logic and the World: An Integrated Theory of Generics*. Oxford University Press UK, Oxford, GB.

Gustavo Novoa, Margaret Echelbarger, Andrew Gelman, and Susan A. Gelman. 2023. Generically partisan: Polarization in political communication. *Proceedings of the National Academy of Sciences*, 120(47):e2309361120. Contributed by Susan A. Gelman; received June 3, 2023; accepted September 25, 2023; reviewed by Yphtach Lelkes and Gregory L. Murphy.

Uwe Peters, Henrik Sherling, and Benjamin Chin-Yee. 2024. Hasty generalizations and generics in medical research: A systematic review. *PLOS ONE*, 19.

Marjorie Rhodes, Sarah-Jane Leslie, and Christina M. Tworek. 2012. Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, 109(34):13526–13531. Edited by Douglas L. Medin, Northwestern University, Evanston, IL, and approved July 2, 2012 (received for review May 25, 2012).

Rachel Sterken. 2015. Generics in context. *Philosophers' Imprint*, 15:1–30.

Preston Stovall. 2019. Characterizing generics are material inference tickets: A proof-theoretic analysis. *Inquiry: An Interdisciplinary Journal of Philosophy*.

Michael Henry Tessler and Noah D. Goodman. 2016. The language of generalization. *CoRR*, abs/1608.02926.