

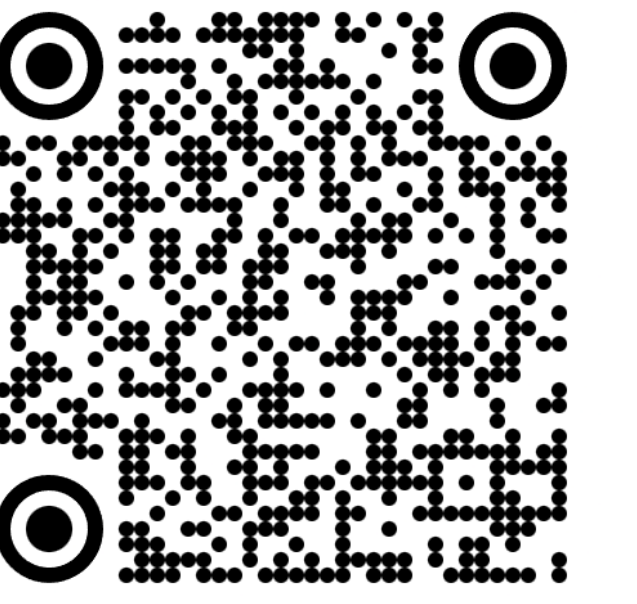


# GENERICIS ARE PUZZLING.

## CAN LANGUAGE MODELS FIND THE MISSING PIECE?

Gustavo Cilleruelo Calderón, Emily Allaway, Barry Haddow, Alexandra Birch

University of Edinburgh; Institute for Language, Cognition and Computation



all? most? some?

tigers

have

stripes

How many individuals need to have the property for speakers to use a **generic** (as opposed to explicit quantification)?

The subject is a **kind**, not a particular set of individuals.

The verb is in the third person plural, present indicative.

The object is a **property** that applies to individuals (not kinds).

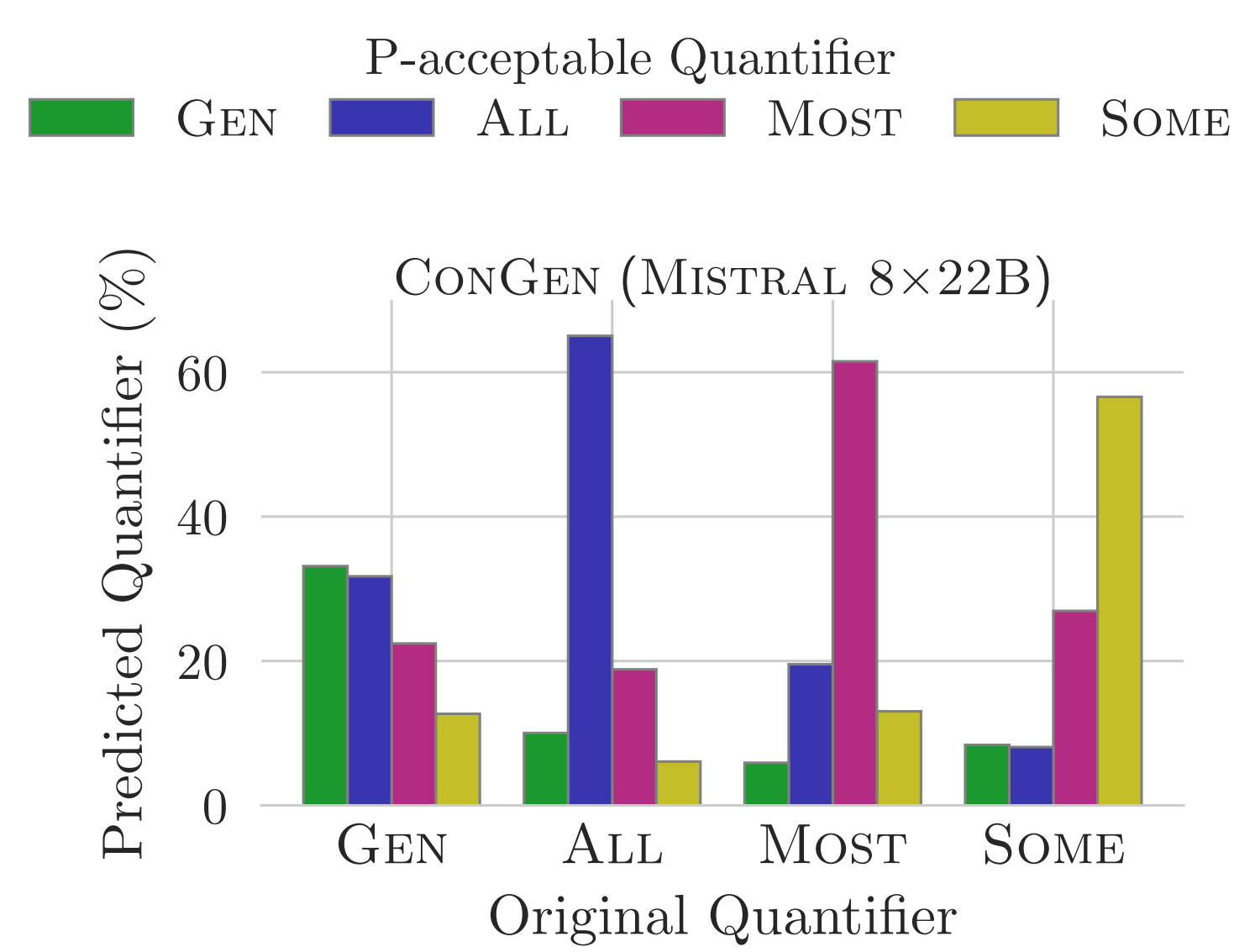
(all) aquatic crustaceans have gills for breathing [100%] · (most) parents serve healthy food [~ 70%] · (some) plastics bind to heavy metals [< 5%]

### Dataset: CONGEN

Manually annotated high quality naturally occurring generic (970) and quantified (1903) sentences in context.

... Runners in general have long legs. **Swimmers have long torsos.** Michael Phelps, who is 6'4", has shorter legs than the Olympic runner Hicham El Guerrouj, who is 5'9" ...

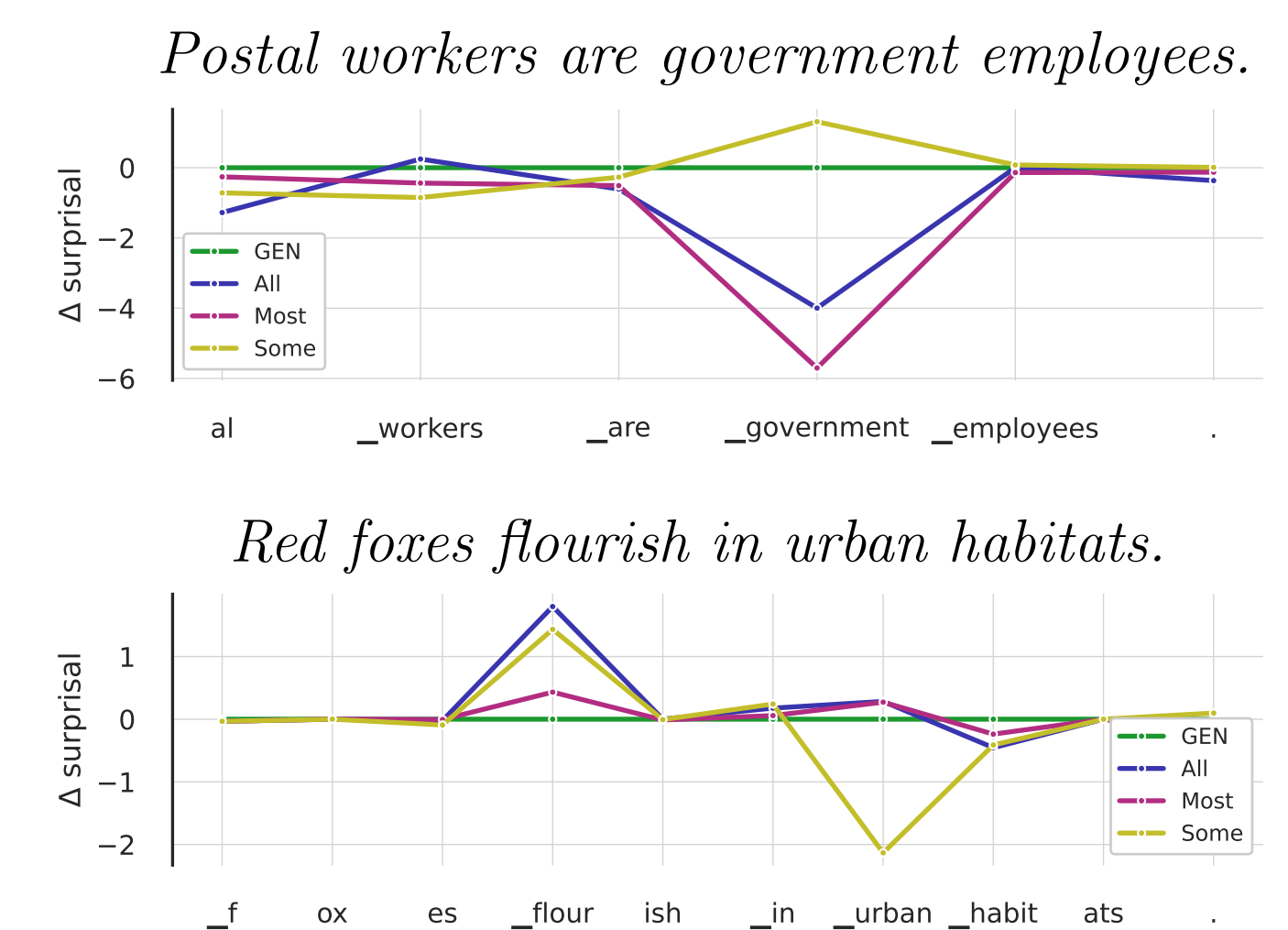
... The one big problem with the strength of the material would be the contraction and cracking during cooling. **Most natural lava tubes collapse because the basalt cracks as it cools.** I'm not sure how to address that issue. ...



Wasps eat | the bad beetles →  $S_\theta = 3.22$   
All wasps eat | the bad beetles →  $S_\theta = 3.2$   
Most wasps eat | the bad beetles →  $S_\theta = 3.13$   
Some wasps eat | the bad beetles →  $S_\theta = 3.01$

### Metric: p-acceptability

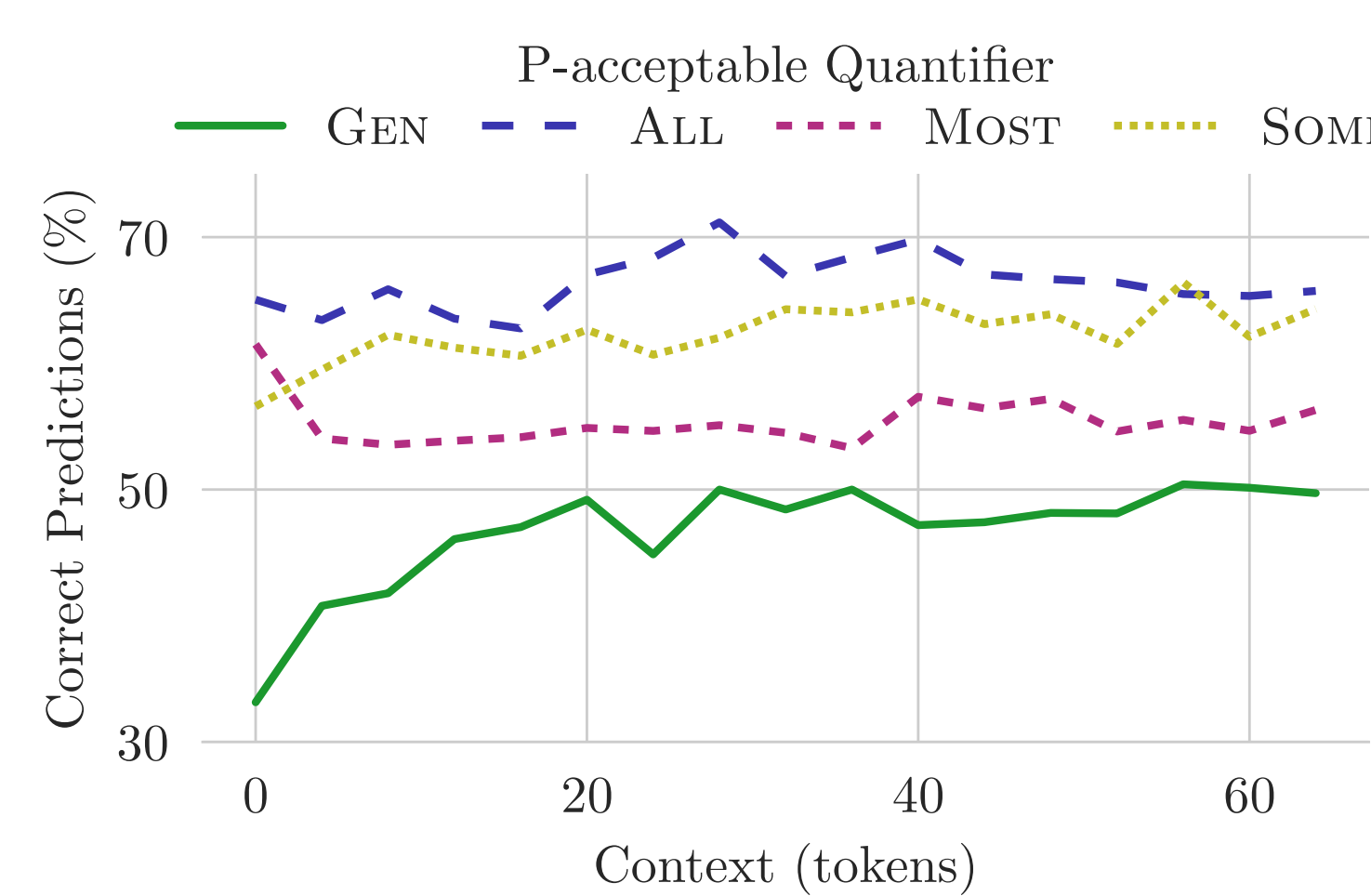
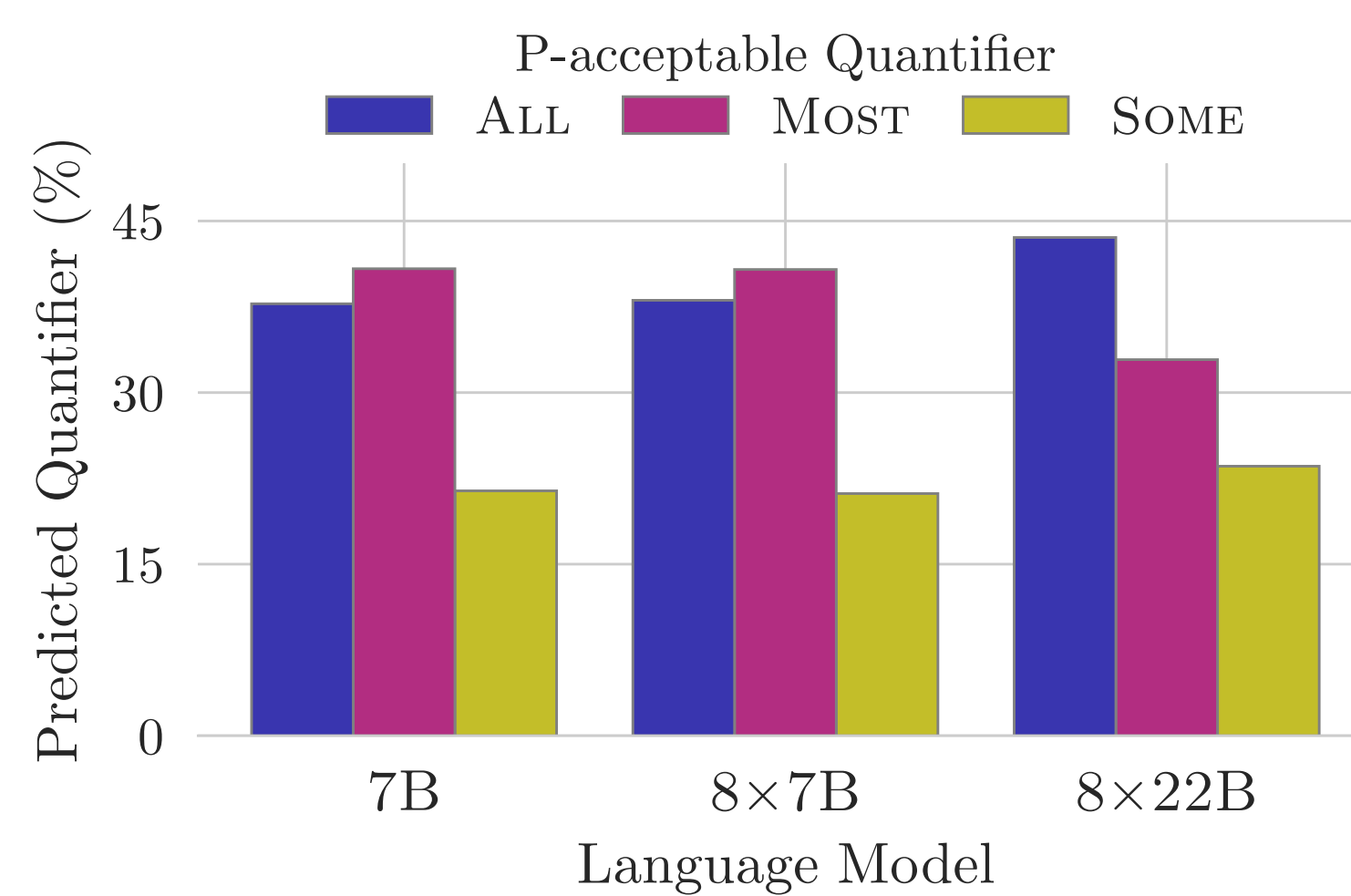
Idea: differentiate generics and quantified sentences with a notion of acceptability based on the surprisal (perplexity) of a language model.



### Implicit Quantification

What determiner quantifier is more p-acceptable in generic sentences?

- Most generics are majority generics (all, most)
- Around 20% are non-striking weak generics (some).



### Context Sensitivity

Does context affect when speakers use generics?

- Measure the accuracy of p-acceptability as we increase context tokens.
- The biggest effect is in generic sentences, suggesting they are distinctively more context-sensitive.

### Strikingness & Stereotype

#### Real Positive

- spaniards love gazpacho in the summer
- germans celebrate with lederhosen and dirndls

#### Invented Positive

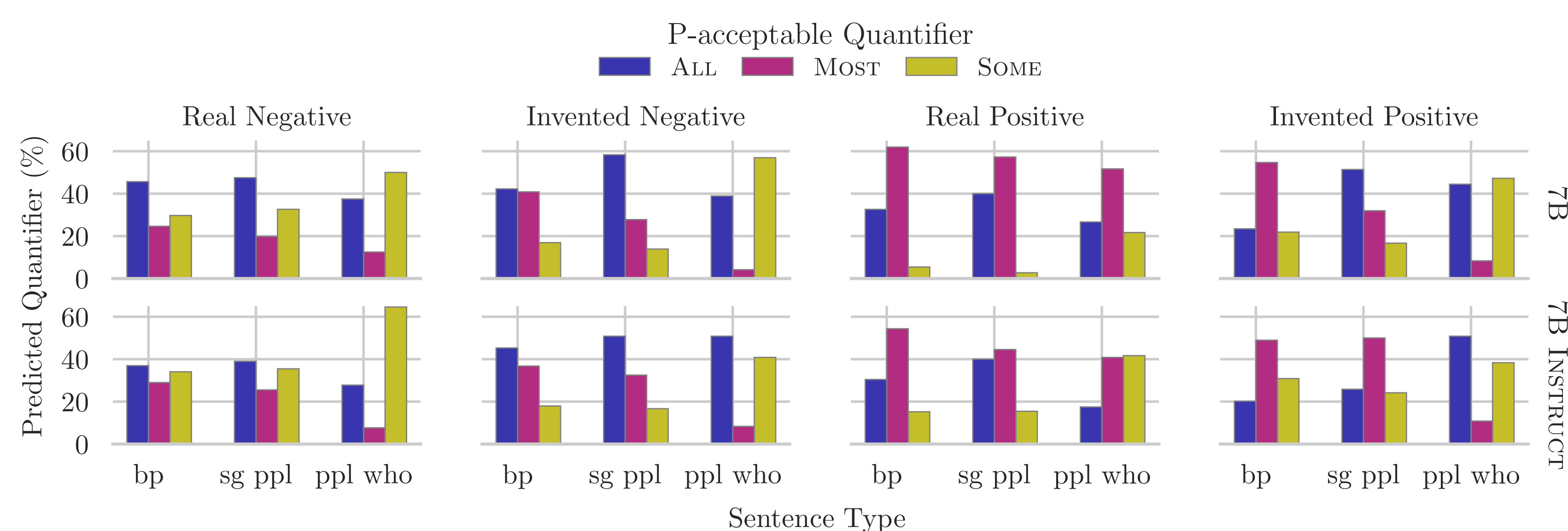
- flirels are smart
- corriards are warm and hospitable

#### Real Negative

- liberals are corrupt
- democrats are cheaters

#### Invented Negative

- slubracks smell bad
- craguils are murderers



#### Are stereotyping generics different?

- Small dataset of stereotypes (N = 504) divided into real (the subject-property is a real-world stereotype) and invented sentences (the subject is an invented word that morphologically resembles a demonym).

- Each sentence has 3 paraphrases: bare plural (Ks are F), singular + 'people' (K people are F) and 'people who are' + singular (people who are K are F).
- Paraphrases affect implicit quantification and negative stereotypes get quantified as all, while positive as most.

@ G.Cilleruelo-Calderon@sms.ed.ac.uk  
gustavocilleruelo.com/generics\_are\_puzzling  
github.com/ilyocoris/generics\_are\_puzzling

