Acquiring morphological generalisations

Evidence from artificial language learning

Maria Copot (work done with Olivier Bonami)

Séminaire expérimental

The big question

- What is the role of learnability and cognitive biases in shaping morphological systems?
  - How much meaning are humans willing to assign to a morphological variant based on incomplete evidence?
  - are typologically common morphologically-encoded meanings easier to learn than rarer ones?
  - Are some types of form distinctions easier to learn than others as a means of conveying meaning distinctions?
  - How does a speaker's native language influence the ease of processing different morphological distinctions?

The long term plan

1. Construct an experimental design that allows us to explore these questions through direct manipulation (an artificial language experiment)
2. Validate the design
3. Establish baseline behaviour of participants
4. Optimise analysis protocol
5. Decide on most promising new directions
6. Employ the protocol developed to answer questions about the impact of learnability and cognitive constraints on the shape of morphological systems.

Today's plan

- The big question: what's the role of learnability and cognitive constraints on the shape of morphological systems?
- Previous literature (or *"why we had to put in all this work"*)
- The experimental material and design
- Putting our experimental design to work
- Analyses, first results and puzzles
- Going forward

Some background

Morphological systems

- A (highly abstract, hand-wavey) take on morphology: words related in meaning are also related in form. The role of a morphological system is to systematically express these relationships.

	SG	PL
LINGUIST	linguist	linguists
CAT	cat	cats

- This one-to-one correspondence between form and meaning is the exception rather than the rule, and there is high variability inter- and intra-linguistically.

	SG	PL
MASC	ragazzo	ragazzi
FEM	ragazza	ragazze

	SG	PL
MASC	linguista	linguisti
FEM		linguiste

	SG	PL
MASC	uomo	uomini
FEM	donna	donne

Morphological systems

- More frequently, we observe a structured lexicon, replete with partial regularities.
- (We never find a lawless system, and it is rare to observe one of perfectly behaved form-to-meaning mappings.)

The literature

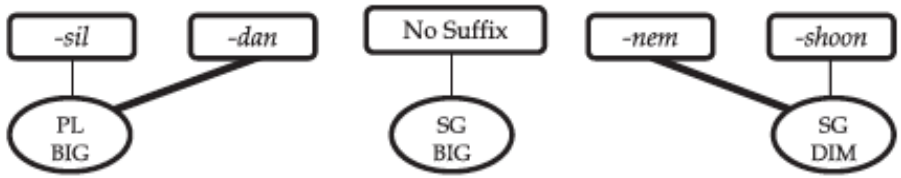
- Existing experimental work on learning morphology is of four main types.
  1. Acquisition of lexical semantics (à la Xu & Tenenmbaum (2007))
     - Informative on general biases in acquisition, but not specifically about morphology.
  2. Iterated learning (à la Kirby & al. (2008))
     - About iteratively assigning structure to random signal (occasionally culminating in morphology).
     - Must start from a small closed system, so that it is learnable in the first iteration (many of our questions target aspects morphological productivity, impossible with a closed system).
  3. Studies on the importance of implicative paradigmatic structure (Seyfarth, Ackerman & Malouf, 2014)
  4. Addressing the questions with suboptimal methodology (Harmon & Kapatsinski, 2017)

Harmon & Kapatsinski, 2017

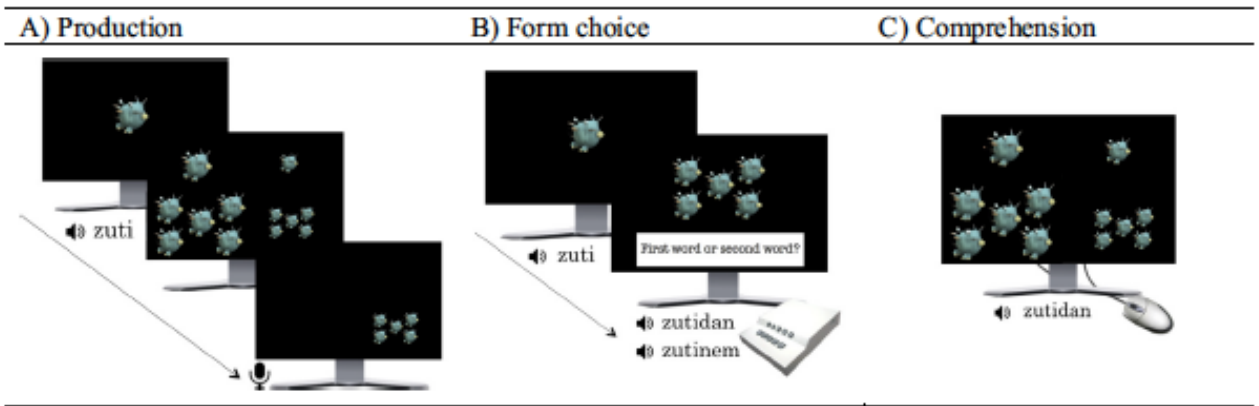
- Question: are speakers willing to assign more meaning to a morph than strictly necessary?
- Great!
  - But they also throw frequency in input and comprehension vs forced choice vs production in the mix straight away.
- Hypothesis
  - Frequent morphs will be extended to new uses in production (but not in comprehension)
    - If a morph is frequent, one can be confident about its meaning and purposefully extend it to new contexts (like in a metaphor) without confusing the listener
  - Infrequent morphs will be assigned new meanings in comprehension (but not in production)
    - If infrequent, less confident about its meaning, more willing to assign it new meaning when hearing it, but unwilling to accidentally misuse it.

Harmon & Kapatsinski, 2017

- Manipulating many variables at once…
- Artificial language experiment
  - Morphology of the language targets things which are routinely encoded in languages (size and number), more chance of bias from real languages that the participant knows.
  - Artificial language constructed in a morphologically naïve way
    - Use of bare stems
    - Complex structure without baseline for simpler cases



  - Biased testing protocol



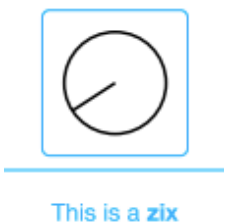
Our work

- We aim to improve upon Harmon & Kapatsinski's core idea
- Set up a more polished experimental design to answer the question they ask, as well as others related to The Big Question©

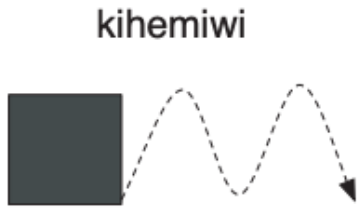
Our work

The materials

- Most artificial language experiments interested in meaning either choose real objects (e.g. Xu & Tenenbaum (2007)) or minimalist drawings



(a) from Carr et al. (2020)



(b) from Kirby et al. (2008)

The materials

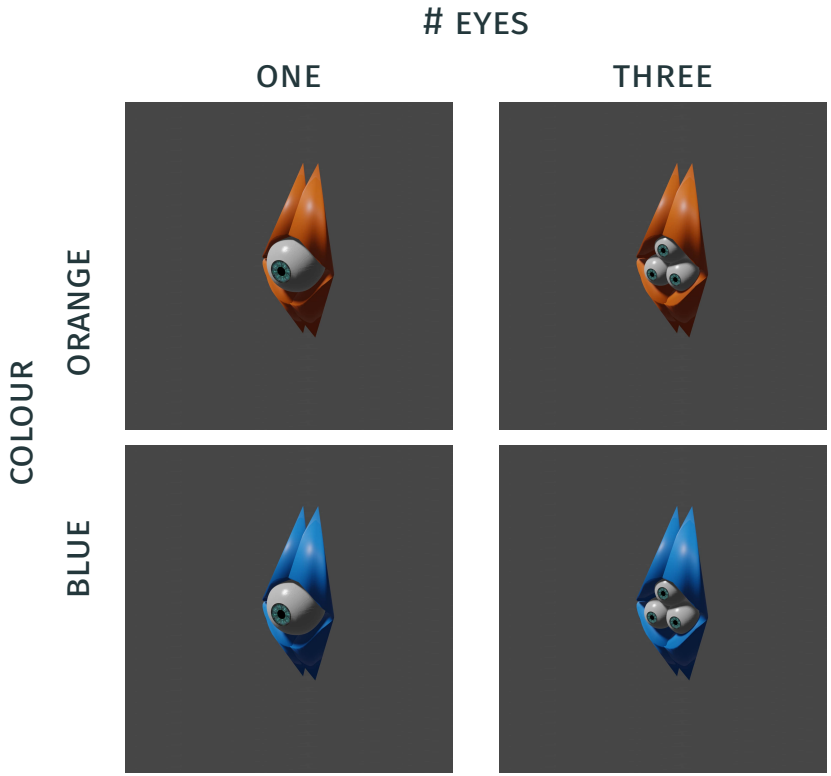
What do we want?

- A (seemingly) open-ended set of beings
- Beings unfamiliar to speakers
  - Less chance of bias from existing concepts and their morphological encoding
- Visually interesting and varied
  - improves participant attention
  - unsystematic variation alongside systematic variation
- Features of systematic variation must
  - Not be commonly inflectionally encoded in the world's languages
  - Be roughly equally salient for all value combinations
  - Be continuous but discretisable
  - Have no clearly unmarked value

The materials

What we came up with

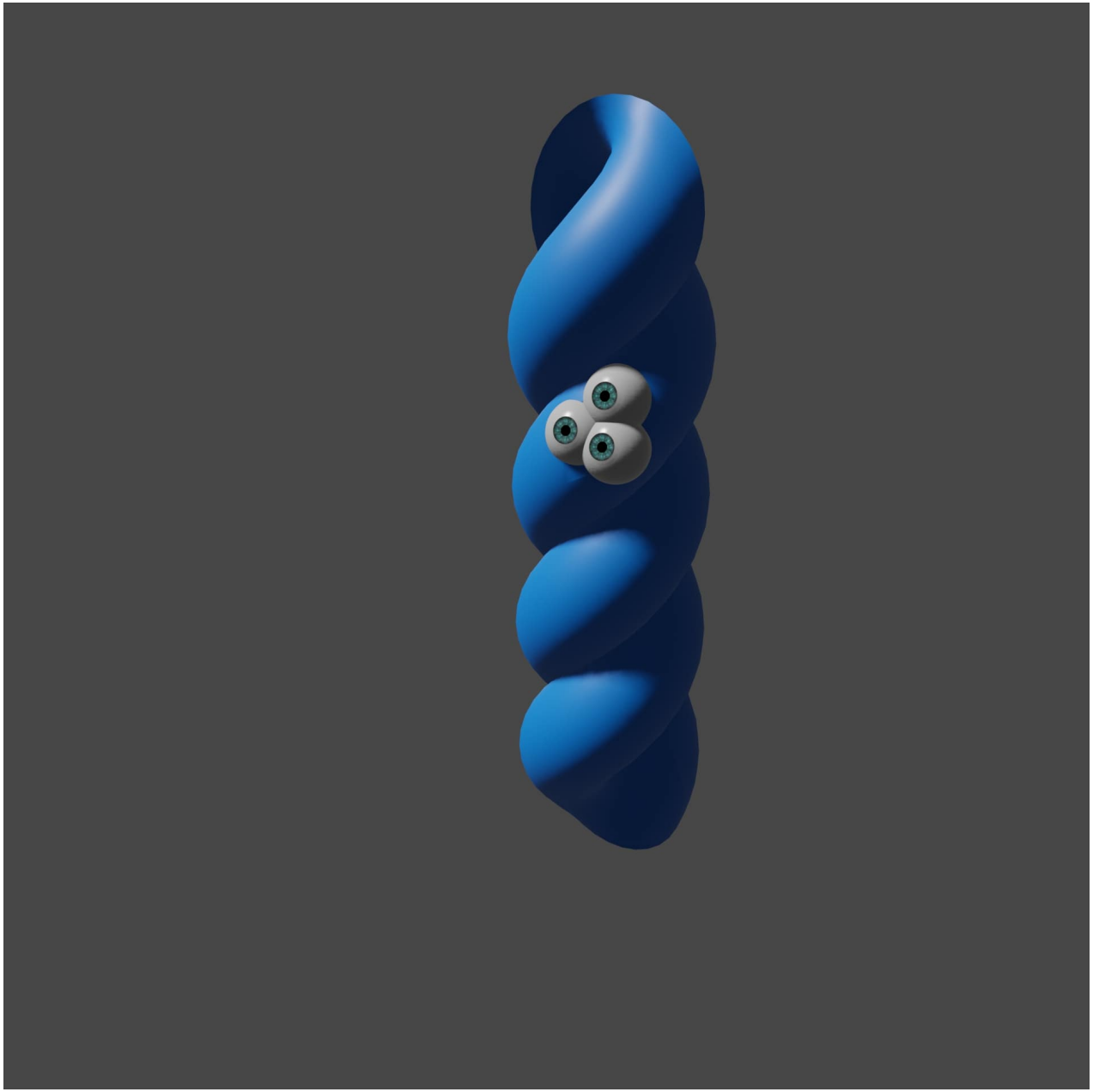
- 40 creatures
- 4 species of each (systematic variation along two dimensions)
  - 1 vs 3 eyes
  - blue vs orange



The materials

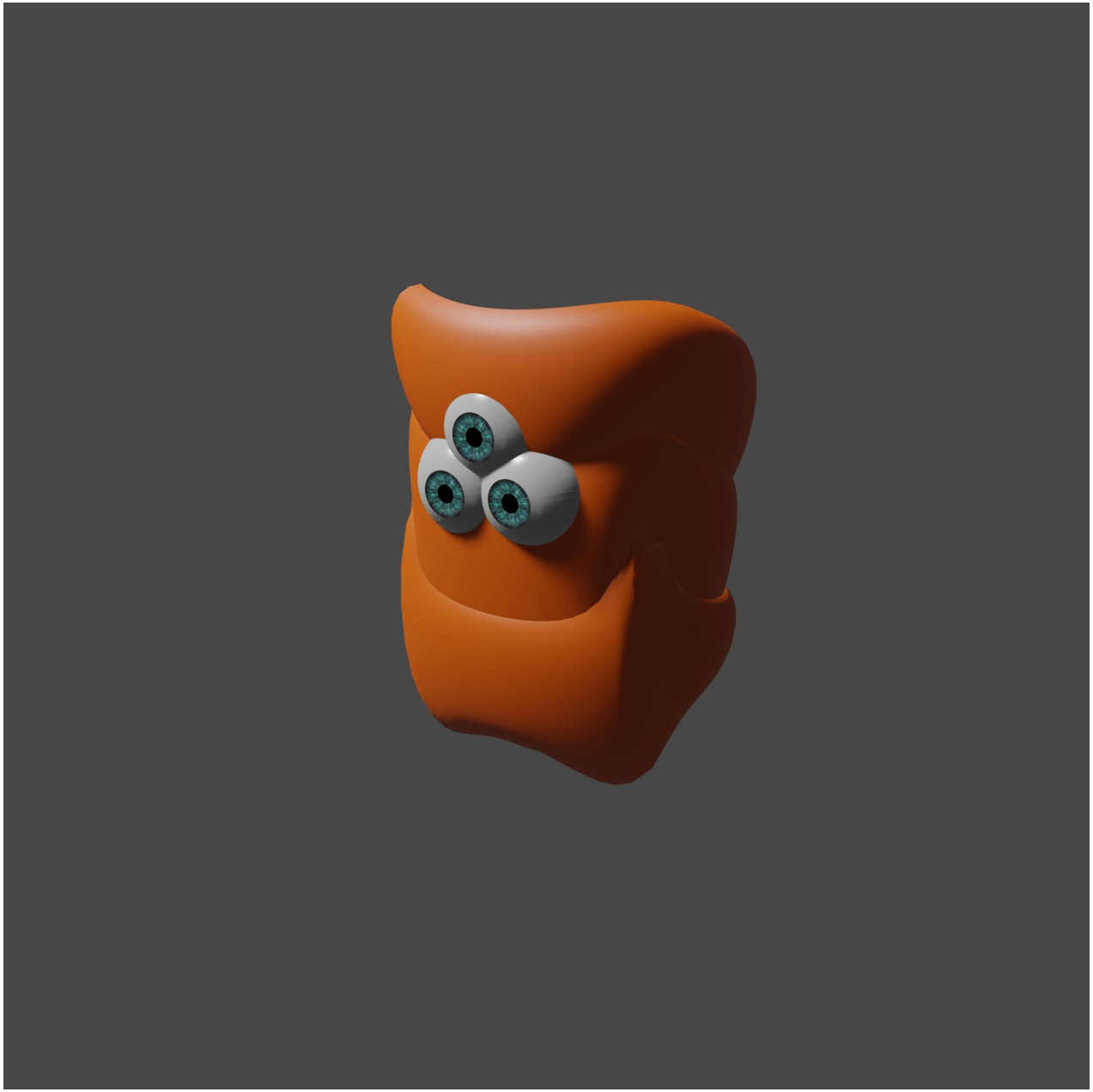
- More elaborate 3D objects are
  - More **interesting** for the participant (attention → memory → learning)
  - Fully **customisable** (can control everything)
- Why these two features?
  - **Neither is inflectionally encoded** in the large majority of languages
  - Both variables have values that **are or can be made discrete.**
  - Roughly equally **salient**
    - after SHAPE, COLOUR is the most salient feature of an object.
  - Addition of eyes makes a big difference in the extent to which participant pay attention to the creature (adds... **relateability?**)

The materials

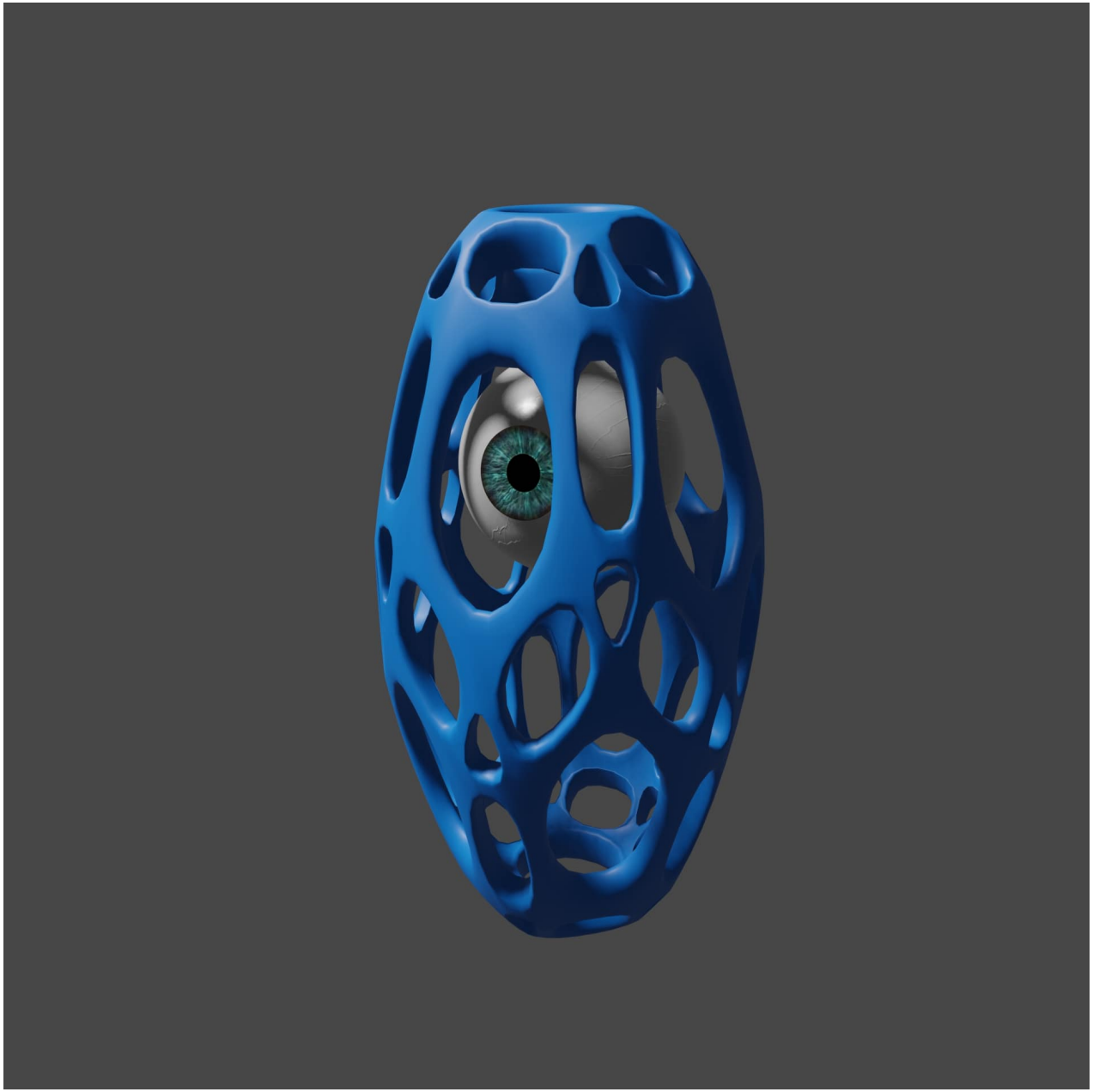


The materials

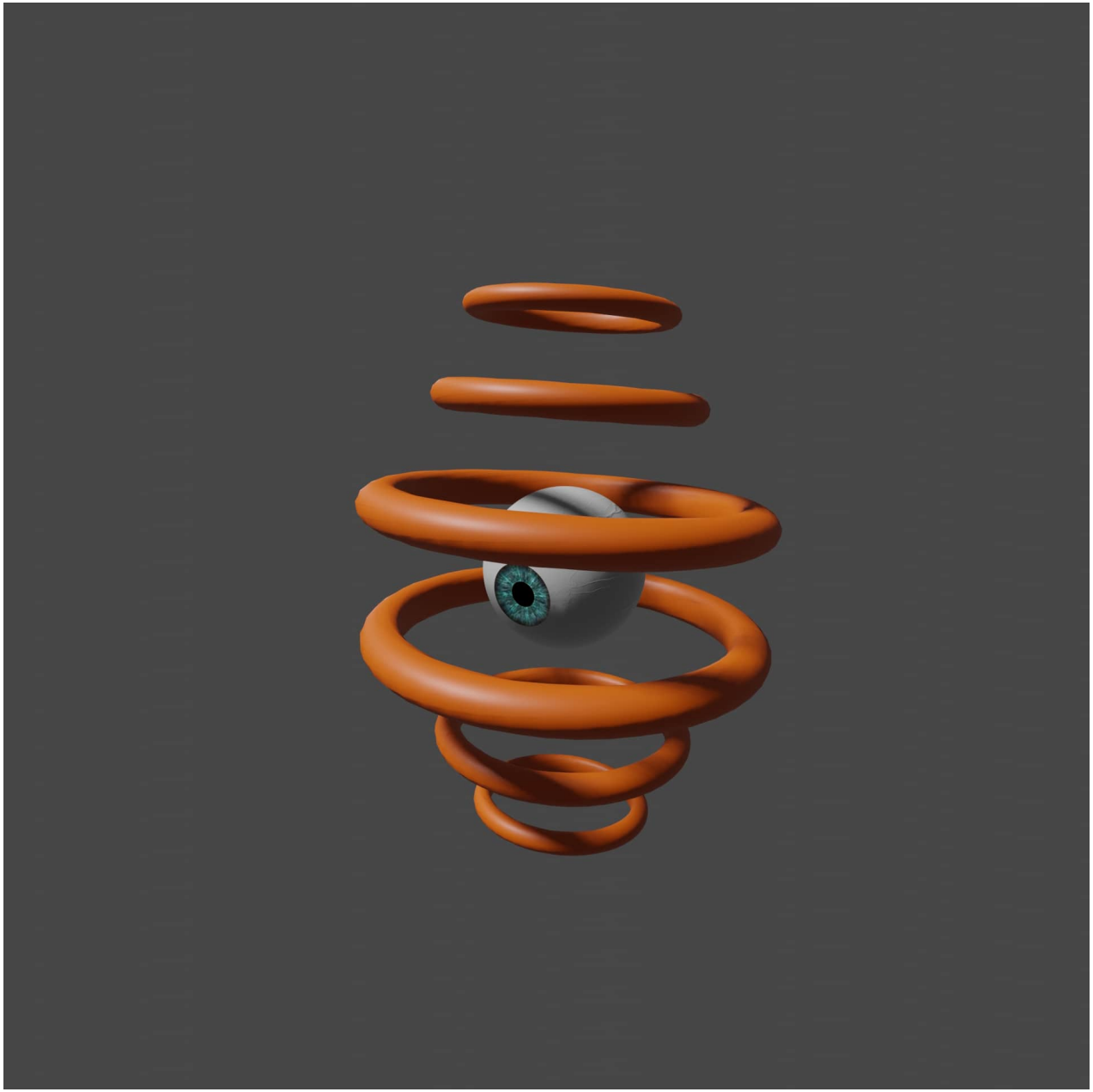
The materials



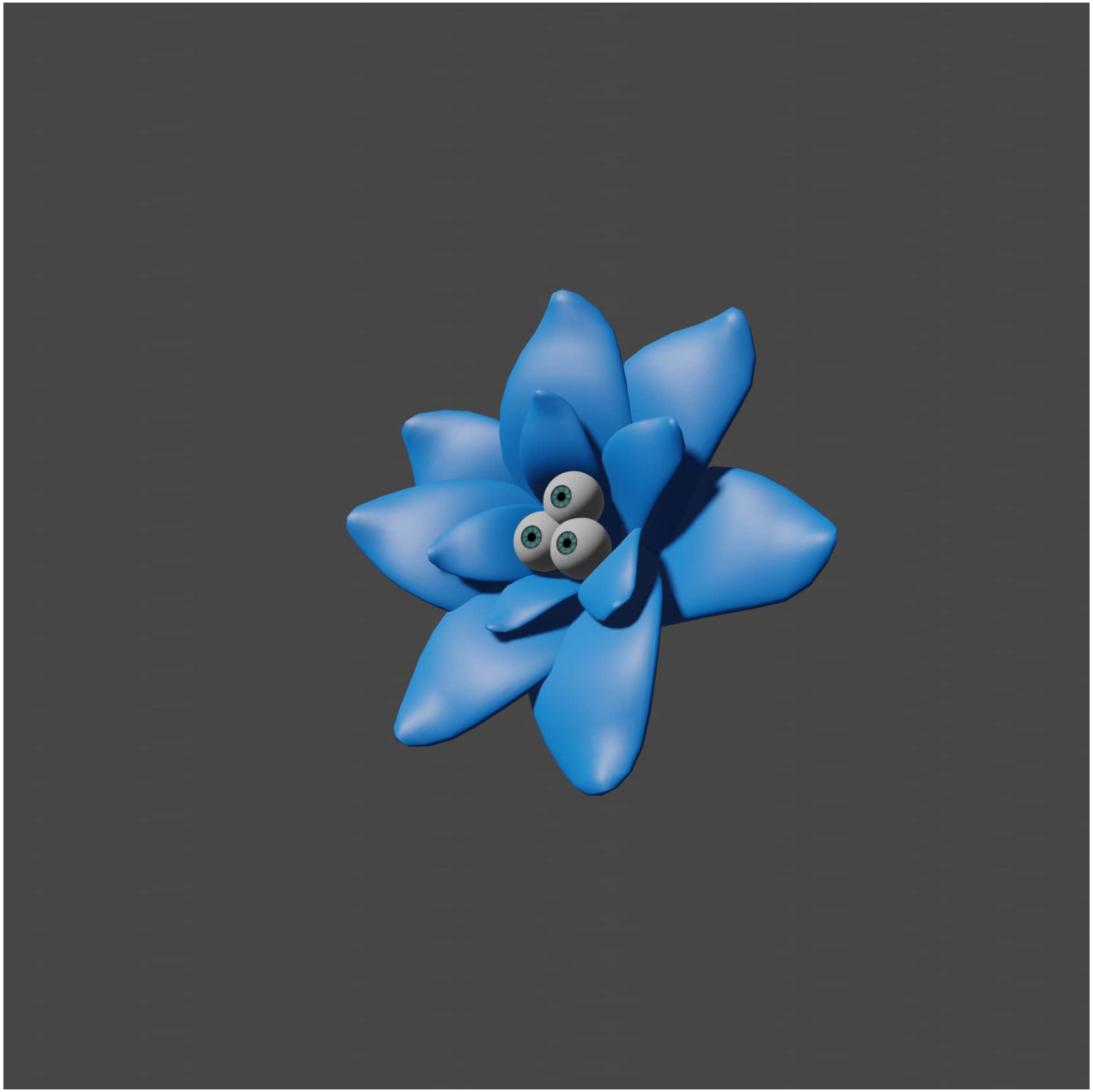
The materials



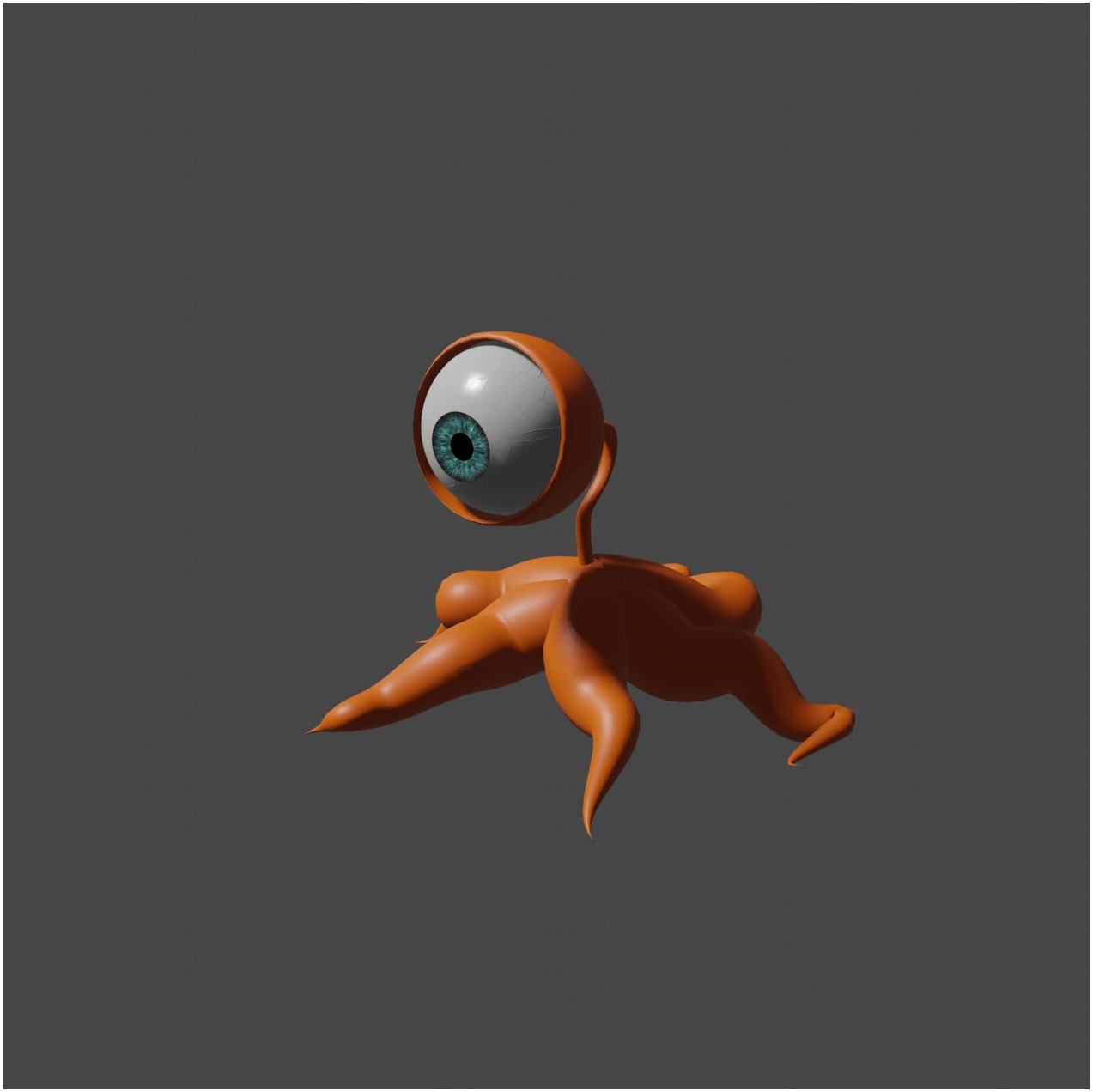
The materials



The materials

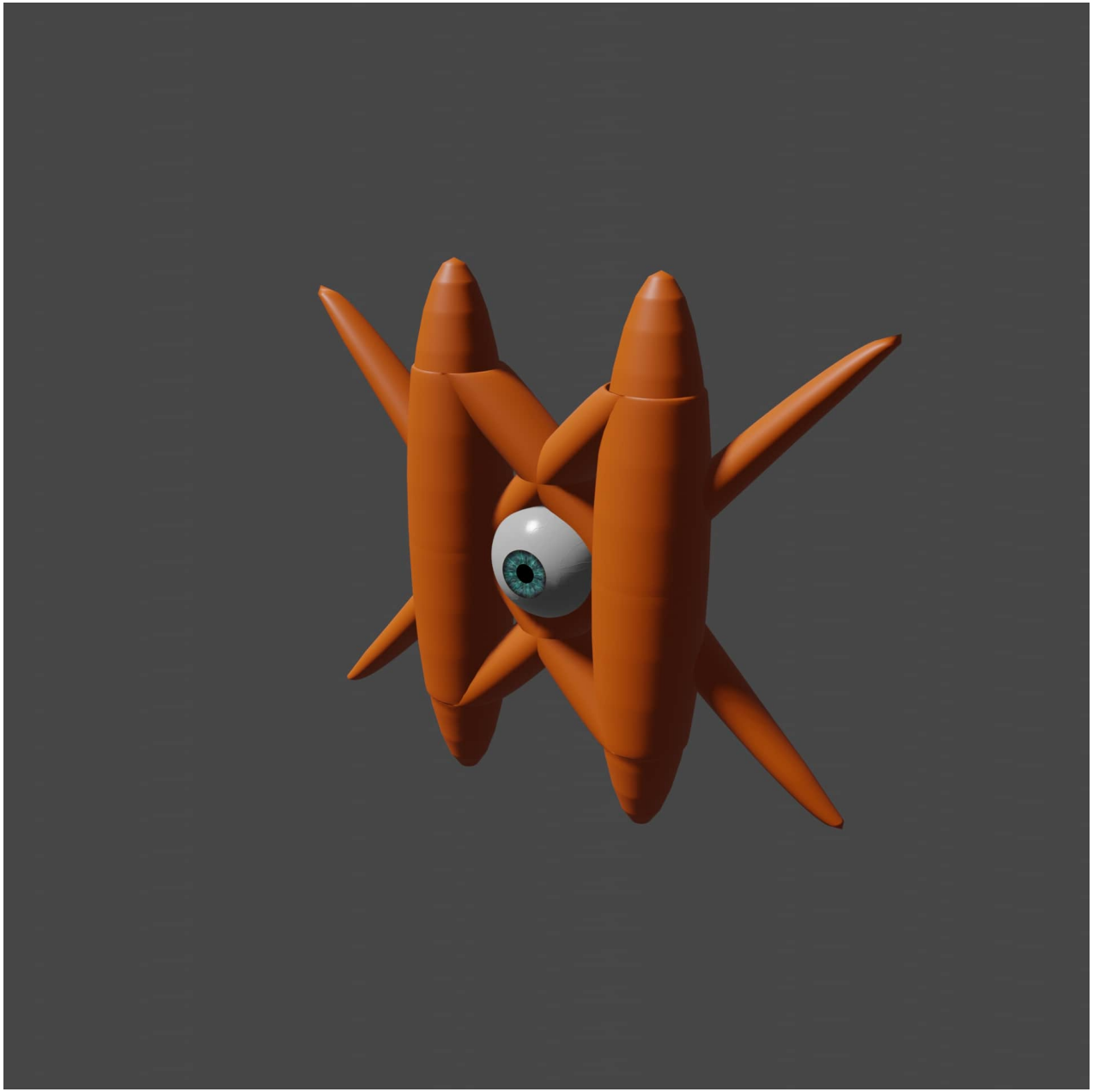


The materials

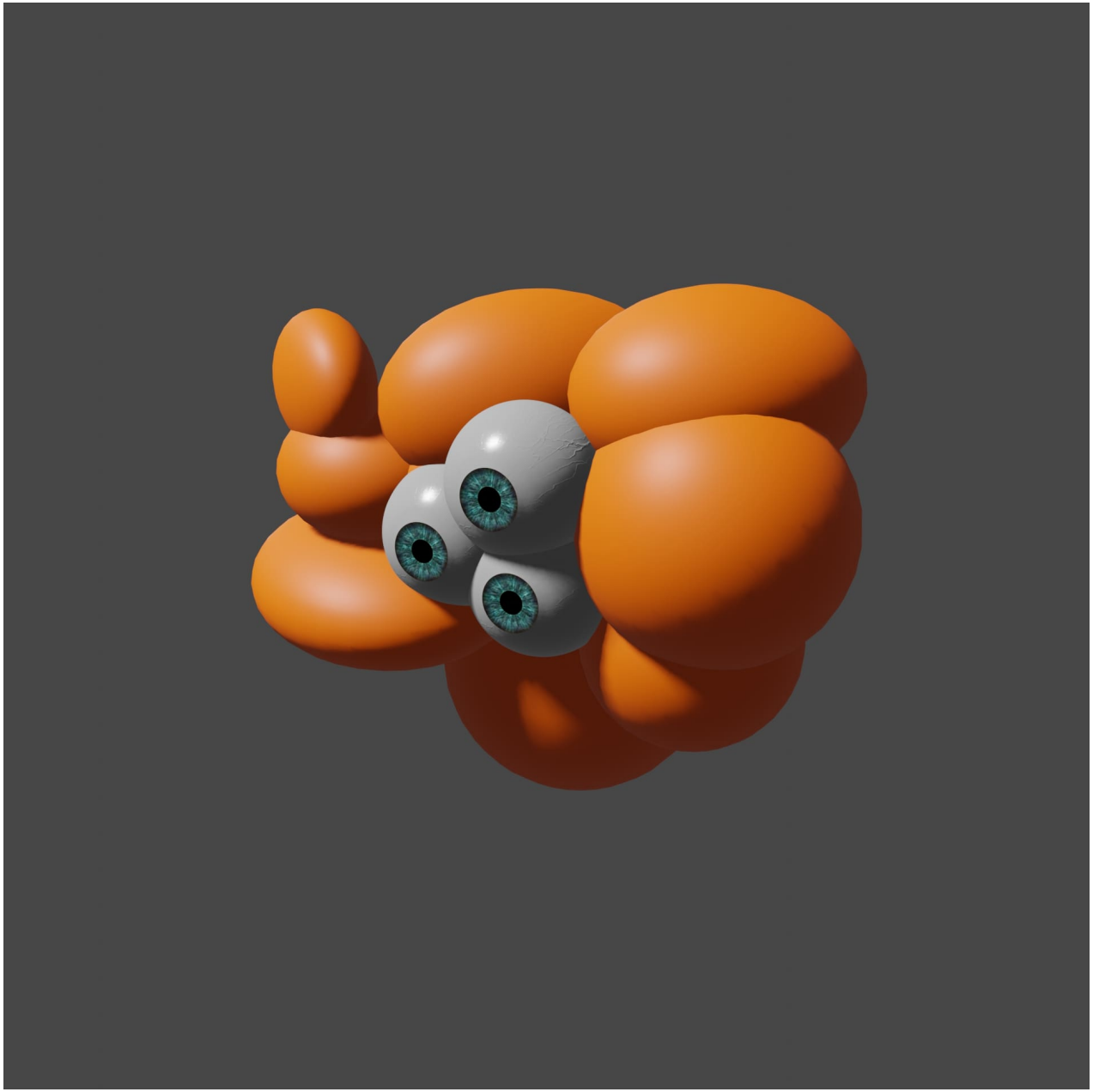


The materials

The materials



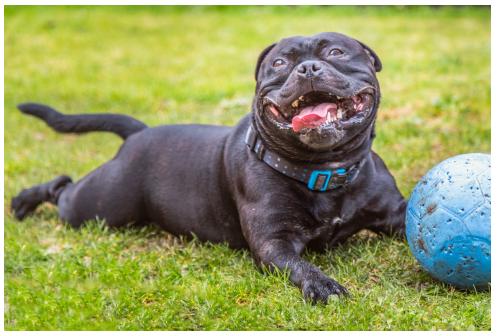
The materials



A test case

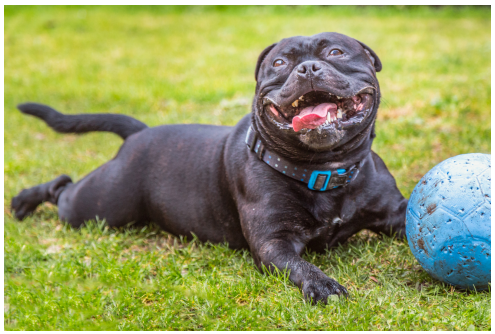
- To validate the materials, and establish a baseline, we chose a simple question: are speakers willing to extend an affix's meaning more than strictly justified by the data?
- This is the same question that Harmon & Kapatsinski (2017) had. Their conclusion:
  - in production, more likely for frequent morphs
  - in comprehension, more likely for infrequent morphs
- Xu & Tenenbaum (2007) asked this question for lexical semantics.

Xu & Tenenbaum (2007)



”These are **FEPS**”

Xu & Tenenbaum (2007)



”These are **FEPS**”



”These are **FEPS**”

Xu & Tenenbaum (2007)

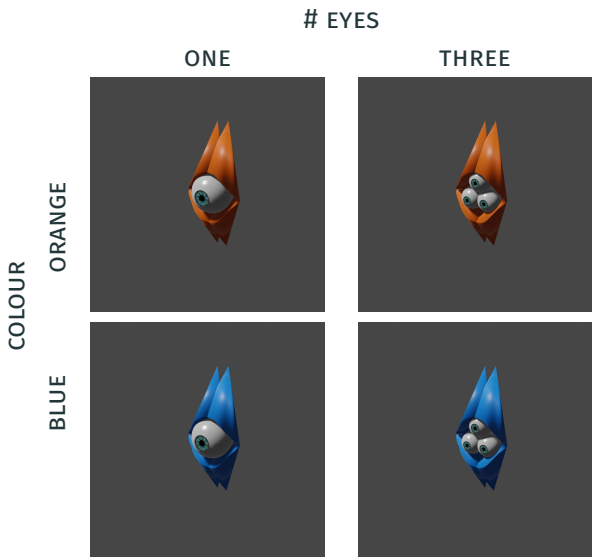


”These are FEPS”

- FEP is still compatible with the superordinate category of DOG
- Though, speakers find it a suspicious coincidence that out three randomly sampled FEPS, all tree happen to be of the subordinate category DALMATIAN.
- In the testing phase, speakers that are shown three dalmatian FEPS are reluctant to extend the word to mean DOG more broadly, even if that would be congruent with the input.

Xu & Tenenbaum (2007)

- Speakers will assign the most specific meaning possible that is not contradicted by the data (entrenchment)
  - The more data available, the more true this is
- They test this for a hierarchical organisation of concepts, but they predict this to be true for concepts with multidimensional organisation too.
- Morphological systems are a prime example of linguistic expression of multidimensional organisation.

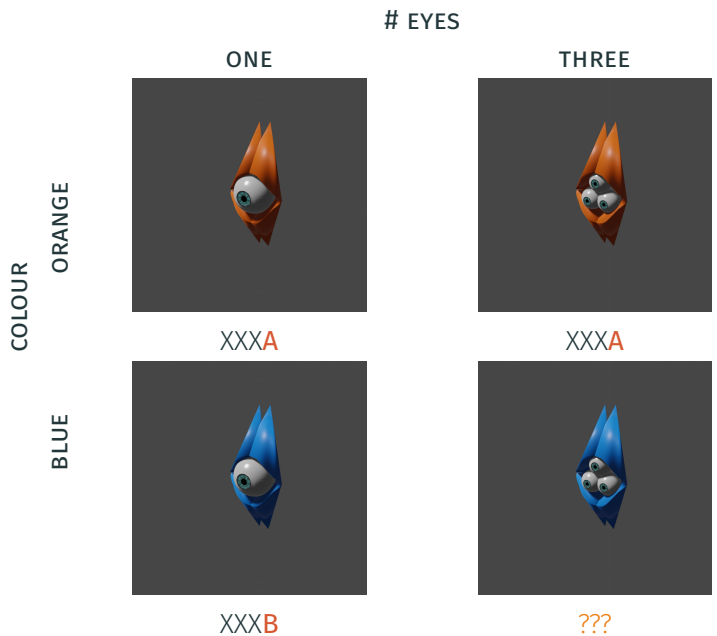


	SG	PL
MASC	ragazzo	ragazzi
FEM	ragazza	ragazze

Our experiment

Our experiment

- An artificial language experiment based on the previously presented materials.
- Does partial evidence lead to entrenchment (via the suspicious coincidence effect)?



The Language

- 40 stems (20 CV, 20 CVCV)
- Features:
  - number of eyes (1 vs 3)
  - colour (blue vs orange)
- Only one feature is expressed morphologically
- 2 suffixes (-ko, -ni)
- No bare stems

The Language

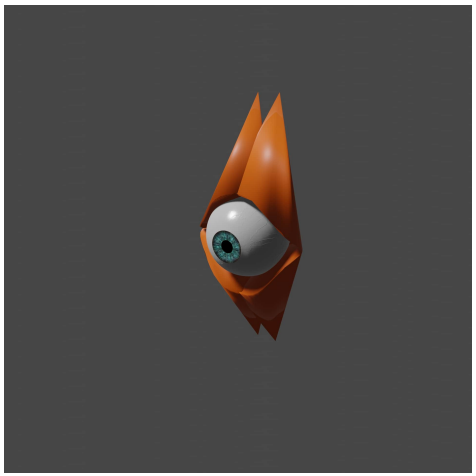
EYES

ONE

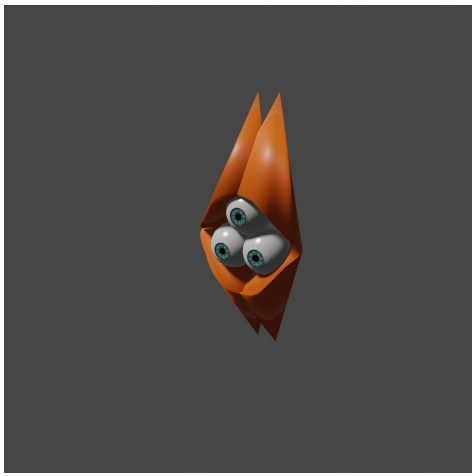
THREE

COLOUR

ORANGE

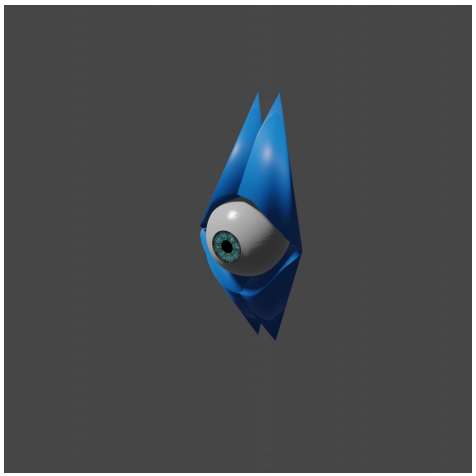


mateni

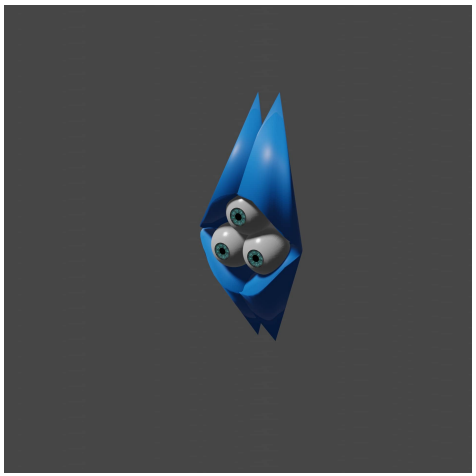


mateni

BLUE



mateko



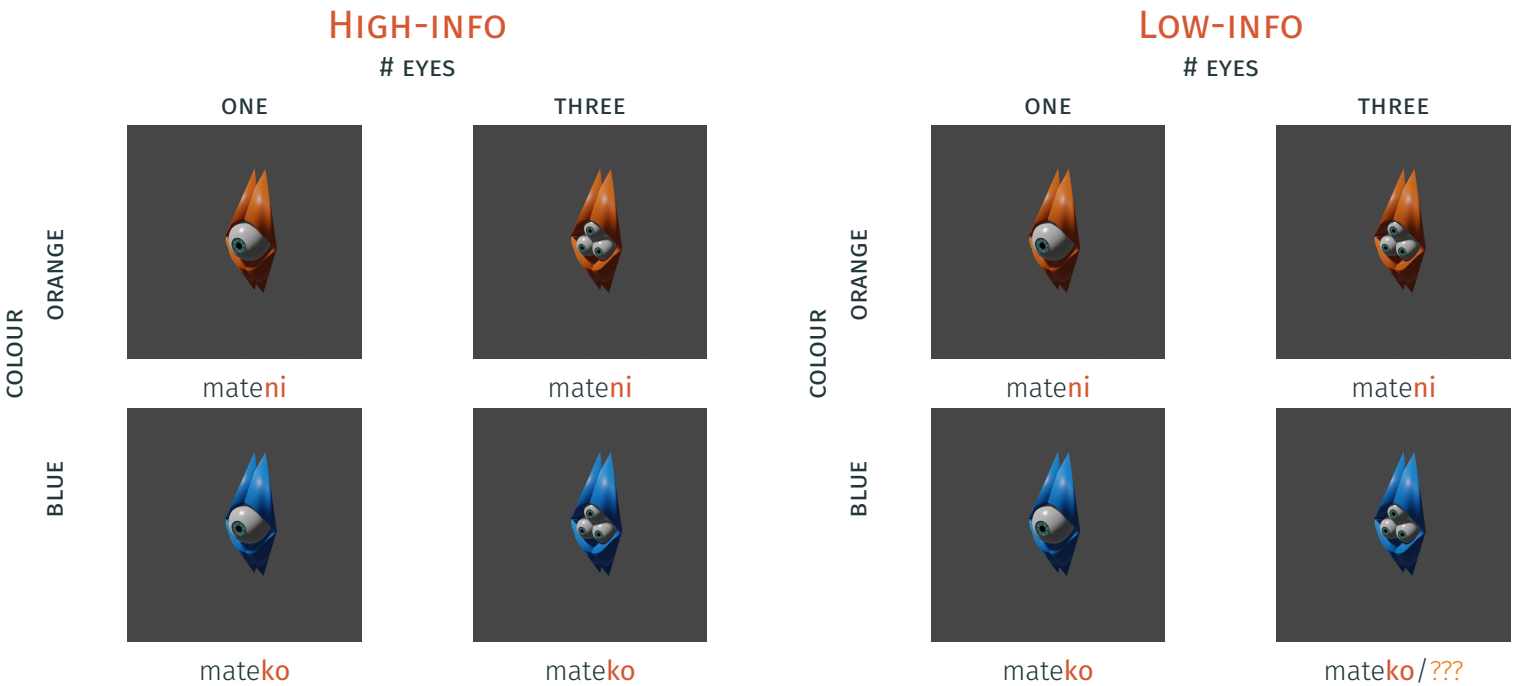
mateko

The language

- 4 possible versions of the language:

	FEATURE REALISED	
	COLOUR	EYES
SUFFIX VALUE	-ko = blue -ni = orange	-ko = 1eye -ni = 3eye
	-ko = orange -ni = blue	-ko = 3eye -ni = 1eye

Participant conditions



- Low-info participants only see -ko for [1EYE, BLUE]. Will they
  - make a minimal generalisation about meaning (-ko = [1EYE, BLUE])
  - make a maximal generalisation about meaning (-ko = [, BLUE], parallel to meaning of -ni = [, ORANGE])

Participant Conditions

- Crucial condition: the condition hidden from low info participants
- Participant is randomly assigned to one of 6 conditions
  - High info
    - Colour morphologically realised
    - #eyes morphologically realised
  - Low info
    - Colour morphologically realised [crucial condition is ORANGE]
    - Colour morphologically realised [crucial condition is BLUE]
    - #eyes morphologically realised [crucial condition is 1EYE]
    - #eyes morphologically realised [crucial condition is 3EYE]
  - What suffix expressed which value of the morphologically realised feature was randomised
  - For low-info participants, which of the two species was withheld out of the two picked out by the crucial condition was randomised.
- 120 participants, from Prolific.co. £9/h (length: 45 min). PCIbex as the experimental platform.

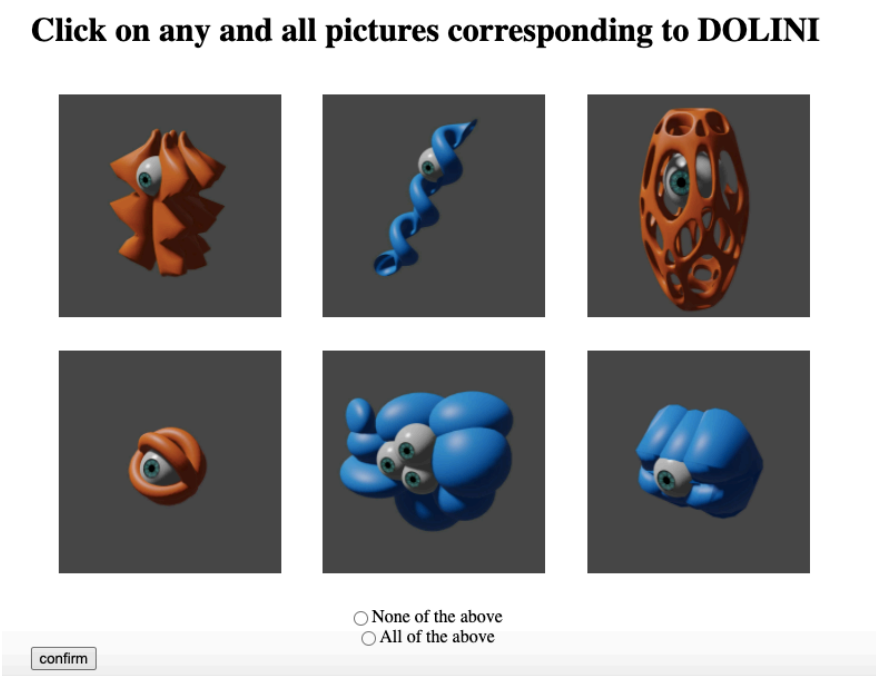
Experiment Structure - Training

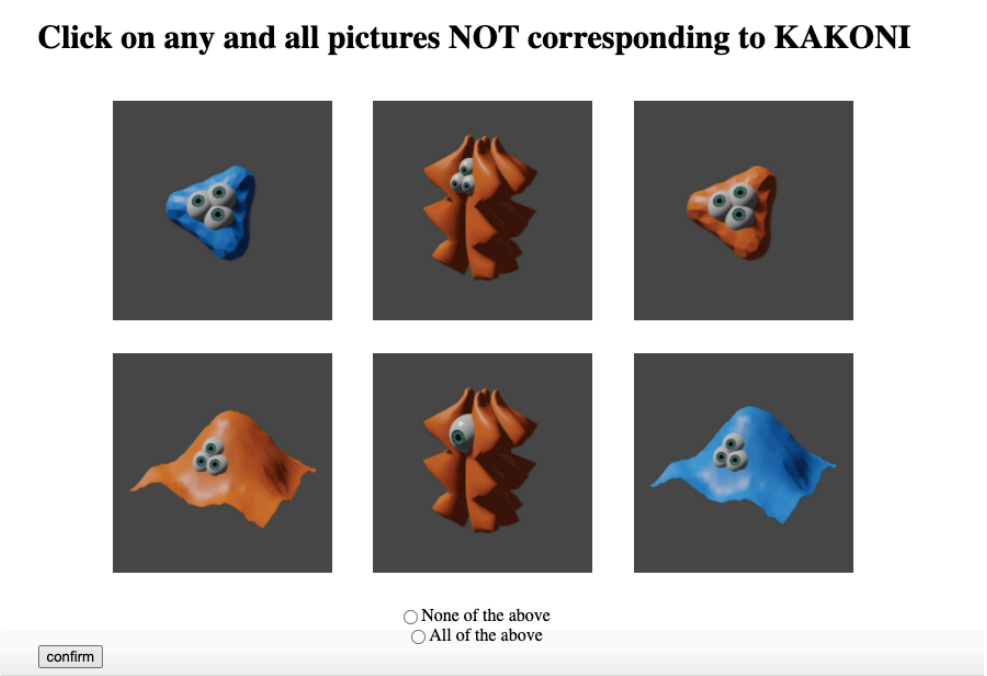
- Learning by exposure: an image and the corresponding word appear together on the screen for 5s.
- Every 10 exposures, participants are shown an image they have seen during the last exposure chunk and asked to input its name.
  - If wrong, the correct name is shown, and they are asked to input it in the text field to continue.
- What is kept constant across participant conditions?
  - Total number of exposures: 180 (90 unique items)
  - The number of stems that each participant sees.
  - The distribution of the conditions is uniform (for each of the N conditions, items from that condition will be seen $1/N$ of the time).
  - What changes: the number of conditions that each stem is presented in (but the average number of times a given stem is shown will be the same across participant conditions)

Experiment Structure - Testing

- Multiple choice questions with no fixed number of possible responses (40 test trials).

# VALID	POLARITY	
	1valid+pos	1valid+neg
	2valid+pos	2valid+neg





Experiment Structure - Testing

- Why?
  - More conducive to gathering accurate information about the speaker generalisation than
    - forced comprehension/production
    - free text input
    - yes/no comprehension
  - Best way we could find to allow participant to select all relevant options without being forced or overwhelmed.
  - In addition, a subset of items are directly informative about whether participants make maximal or minimal generalisations.
  - Negative questions help differentiate between *uncertain about X* and *certain that not X*
  - Based on the results of H&K, comprehension is the type of question most likely to result in participants making minimal generalisations/entrenching.

Recap

- The question: When incomplete information is available, are speakers willing to extend available means to new contexts (maximal generalisation) or are they going to be conservative (minimal generalisation)?
- The setup
  - The creatures:
    - 40 beings, 4 item conditions
  - The language:
    - Only one feature is realised morphologically. Each of its two values takes a different suffix.
  - Participant conditions:
    - high-info: see all 4 species, exposed to all feature combinations and their morphological realisation
    - low-info: see only 3 species and exposed to their morphological realisation (withheld item condition = crucial condition)
- The experiment
  - 180 training trials (differ between participant conditions)
  - 40 testing trials (comparable across participant conditions)

Analysis

Analysis

- The question: do low-info participants make the maximal generalisation or the minimal generalisation?
  - In other words: are low-info and high-info responses meaningfully different?
- Multiple ways of characterising the status of participant responses, each is enlightening from a particular perspective.
  1. Validity of choice based on stem only
  2. Validity of choice based on stem+affix
  3. Validity of choice based on affix alone

  1. Single images as data points

     $$\text{is\_image\_selected?} \sim \text{participant\_condition*is\_crucial}$$

  2. Trials as data points

     $$\text{max\_vs\_min?} \sim \text{participant\_condition}$$

Results

- By all available measures, low-info participants behave like high-info participants: both seem to consistently make the maximal generalisation.
- If anything, it seems that low-info participants have slightly better performance overall than high-info participants
  - the stimulating effect of unseen items on attention?
  - performance was slightly worse on negative questions
  - performance slightly worse on eyes as morphologically realised variable.
  - stem associations are challenging to learn

Lessons learned

- The current design was chosen to be maximally comparable with previous literature, and to maximise the chance of seeing minimally generalising behaviour, but it is not the optimal design to answer this specific question
  - A more straightforward option (for participants and for analysis): single image shown - ”is this an X?” + confidence rating.
- More emphasis should be placed on communicative function
- Stems are really hard to learn for participants (50% error rate). Testing and analysis should try to rely less on this.

Analysis

- Models discussed in this section are bayesian, fitted with brms.
- Priors set are weakly informative (questions welcome, but won't go into depth today).
- For the purposes of analysis, selection values for negative questions have been reversed.

Single images as data points

- PROS
  - Can exploit every single participant choice
- CONS
  - Treats items in the same trial as independent
  - The main question is answered somewhat indirectly (the interaction coefficient of participant_condition:is_crucial)
    - To make this possible, high-info participants are assigned at random one of the four conditions as crucial.

Model 1A: images as data points + only stem matters

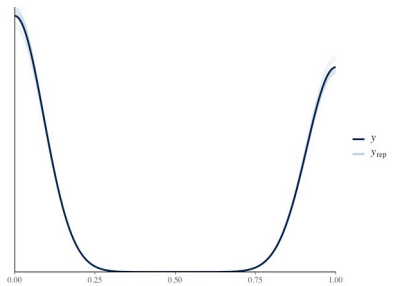
How hard is the task of learning stem associations?

- Data:
  - only trials for which a single stem has been selected
  - only images selected by participant

$$\text{intended\_stem} \sim \text{pos\_neg} + \text{n\_valid} + \text{order} + \text{part\_cond} + \text{morph\_real} + (\text{pos\_neg} + \text{n\_valid} + \text{order} + \text{part\_cond} + \text{morph\_real}|\text{partID})$$

intended_stem	is the image associated with the stem in the task demand?
pos_neg	question polarity
n_valid	how many valid choices available?
order	trial is nth in experiment
part_cond	high-info vs low-info
morph_real	which of the two features does the language realise on words?
partID	ID of participant

- Data points: ~3000
- Family: bernoulli (log link)



Model 1A: images as data points + only stem matters

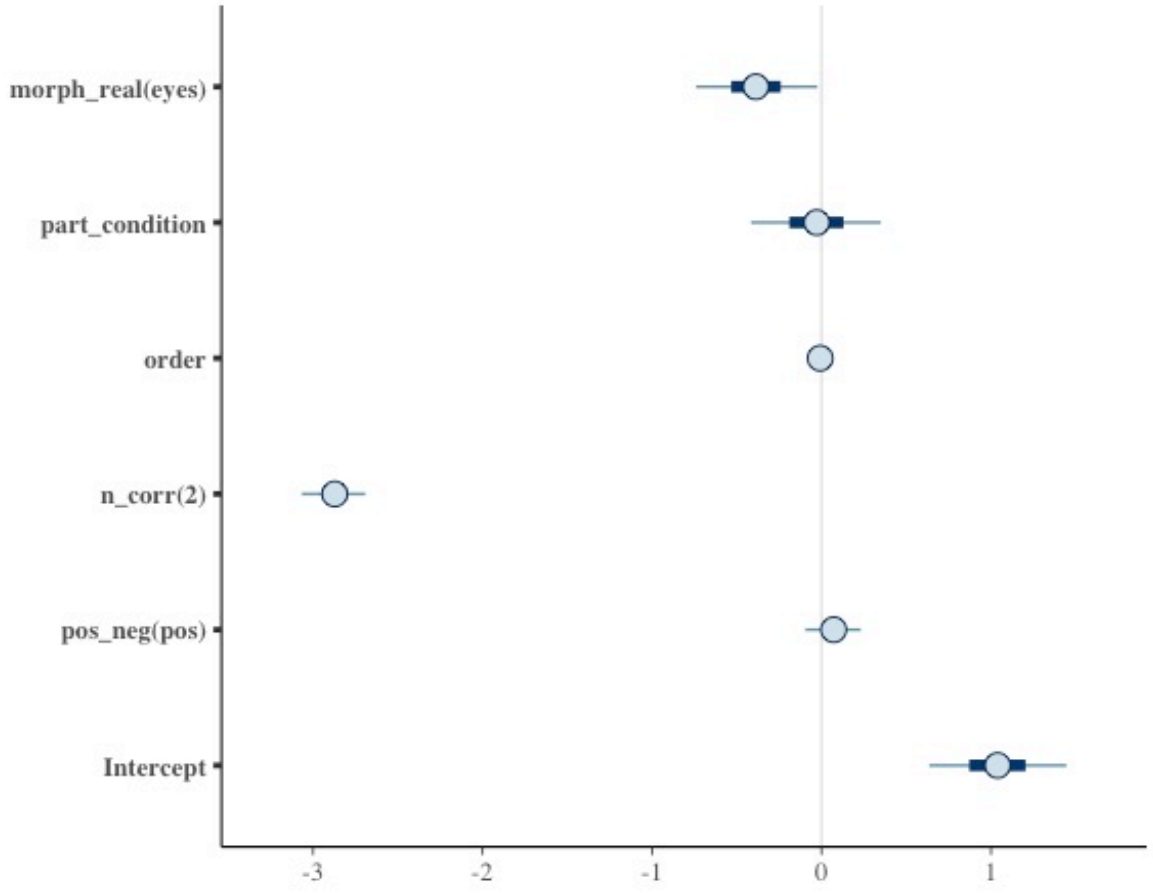


Figure 8: fixed effects

Model 1A: images as data points + only stem matters

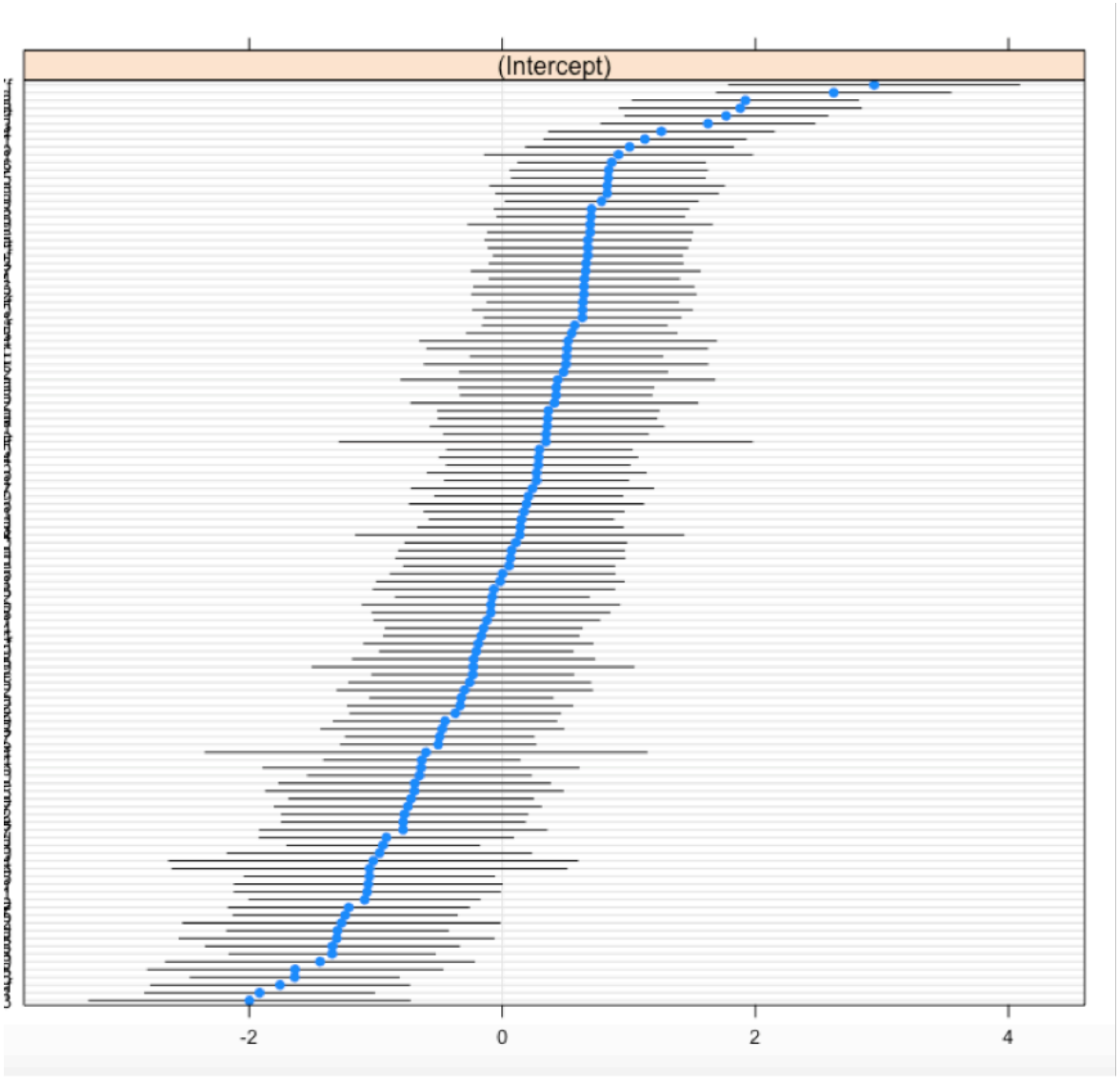


Figure 9: by-participant random intercept

Model 1A: images as data points + only stem matters

- Remembering the stem is not trivial
  - The analysis should rely less strongly on the correct stem being chosen
- Unclear whether to exclude participants based on this: some superperformers, but no clear underperformers

Model 1B: images as data points + stem & affix matter

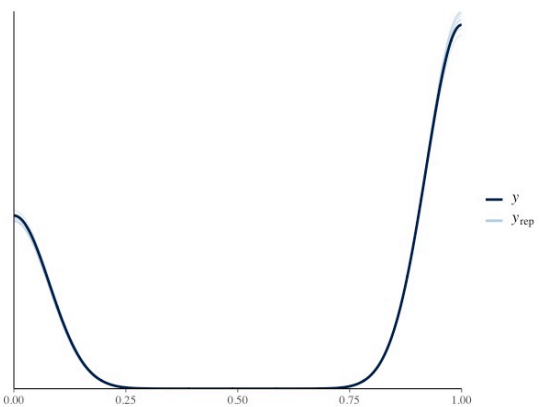
Do participants in the low-info condition select the crucial condition less frequently than those in the high-info condition?

- Data:
  - only trials for which one single stem had been selected
  - only data points which constitute a valid selection based on the meaning of stem+affix under a maximal generalisation.

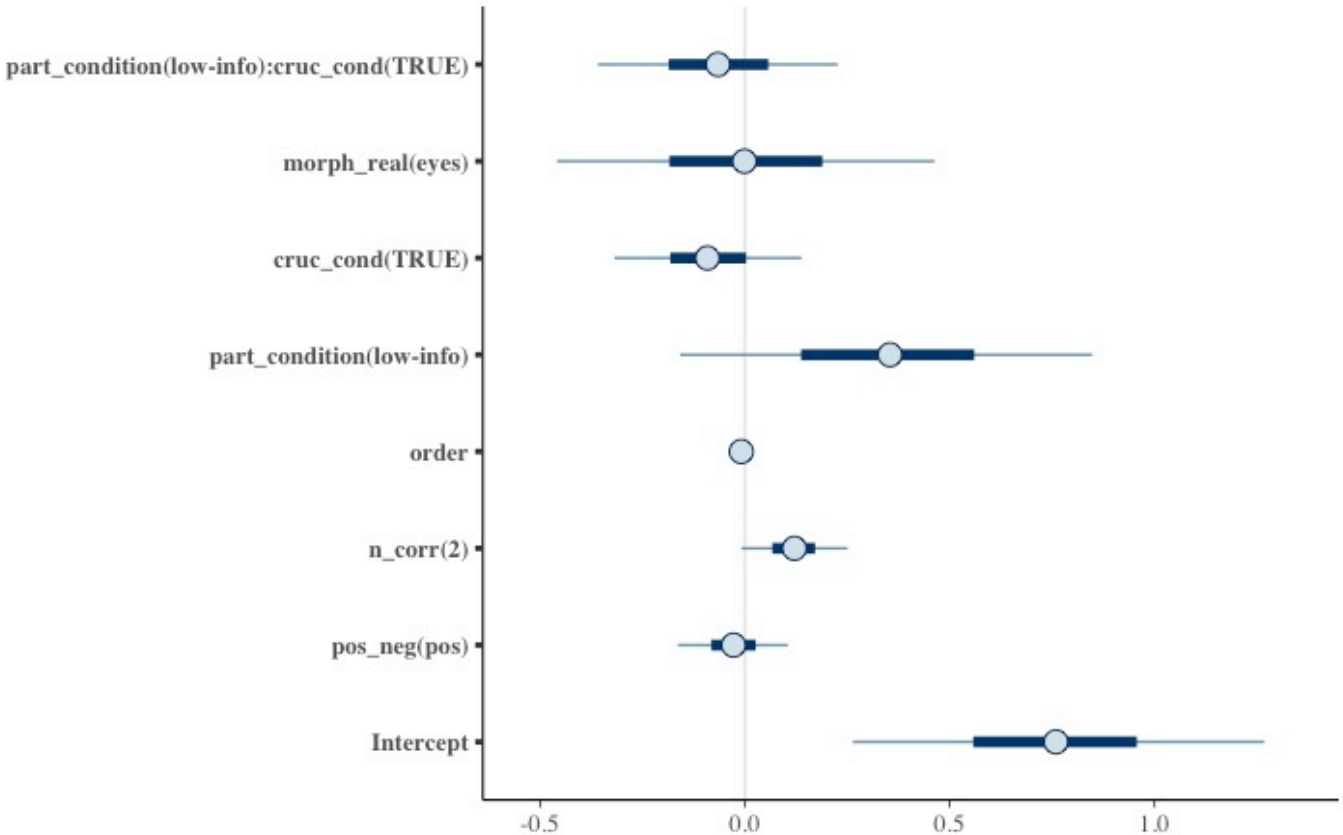
$$\text{sel} \sim \text{pos\_neg} + \text{n\_valid} + \text{order} + \text{part\_cond*cruc\_cond} + \text{morph\_real} + (1|\text{partID})$$

sel	has the image been selected?
cruc_cond	does the image belong to the crucial condition?

- Data points: ~3000
- Family: bernoulli (log link)



Model 1B: images as data points + stem & affix matter



Model 1B: images as data points + stem & affix matter

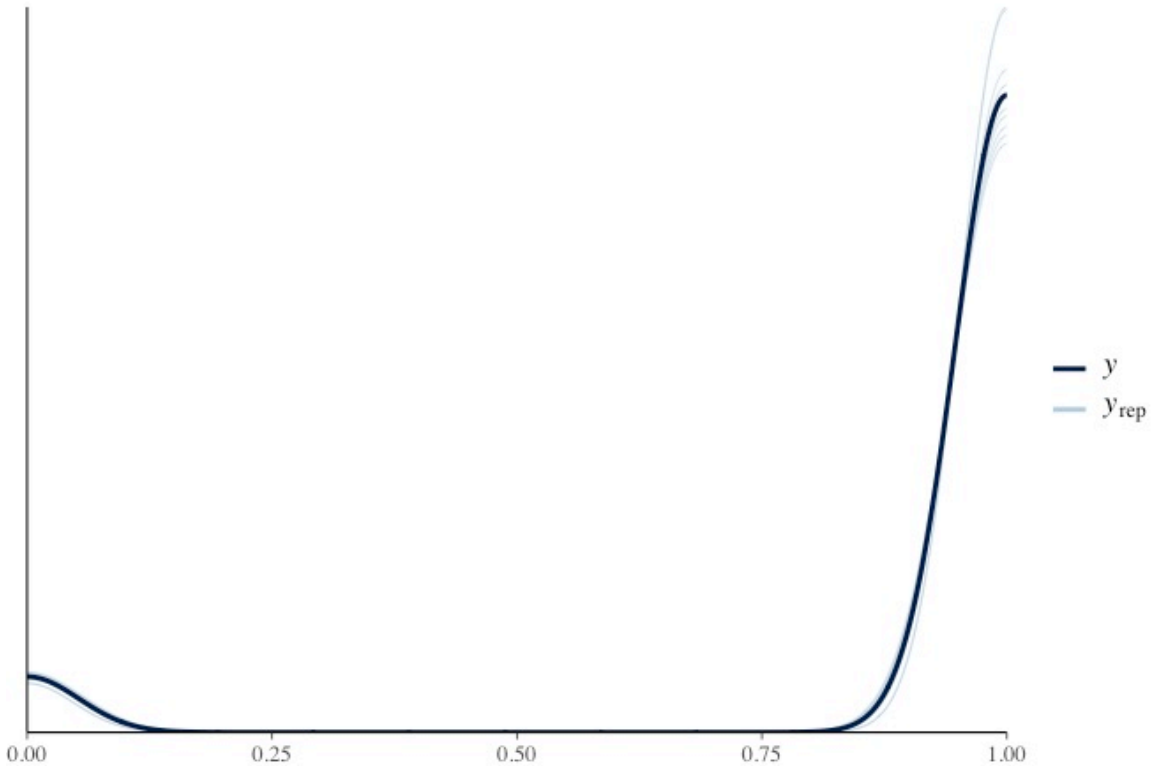
- No meaningful difference in how the two groups react to the crucial condition.
- A puzzle: overall, low-info participants seem to do somewhat better than high-info ones. Two possible explanations
  1. There are twice the number of low-info participants than high-info ones. Could more confidence in one of the means account for the difference?
  2. Are low-info participants simply more attentive because they are encountering new things in testing?

Model 1C: images as data points + only affix matters

- What if the participant has the right feature-affix association but gets the wrong stem?
- For trials in which only one stem is selected, recode possible answers based on the stem choice of the participant.
- Exact same model structure as 1B
- Data: only valid choices under a maximal generalisation, based on the stem selected by the participant

Model 1C: images as data points + only affix matters

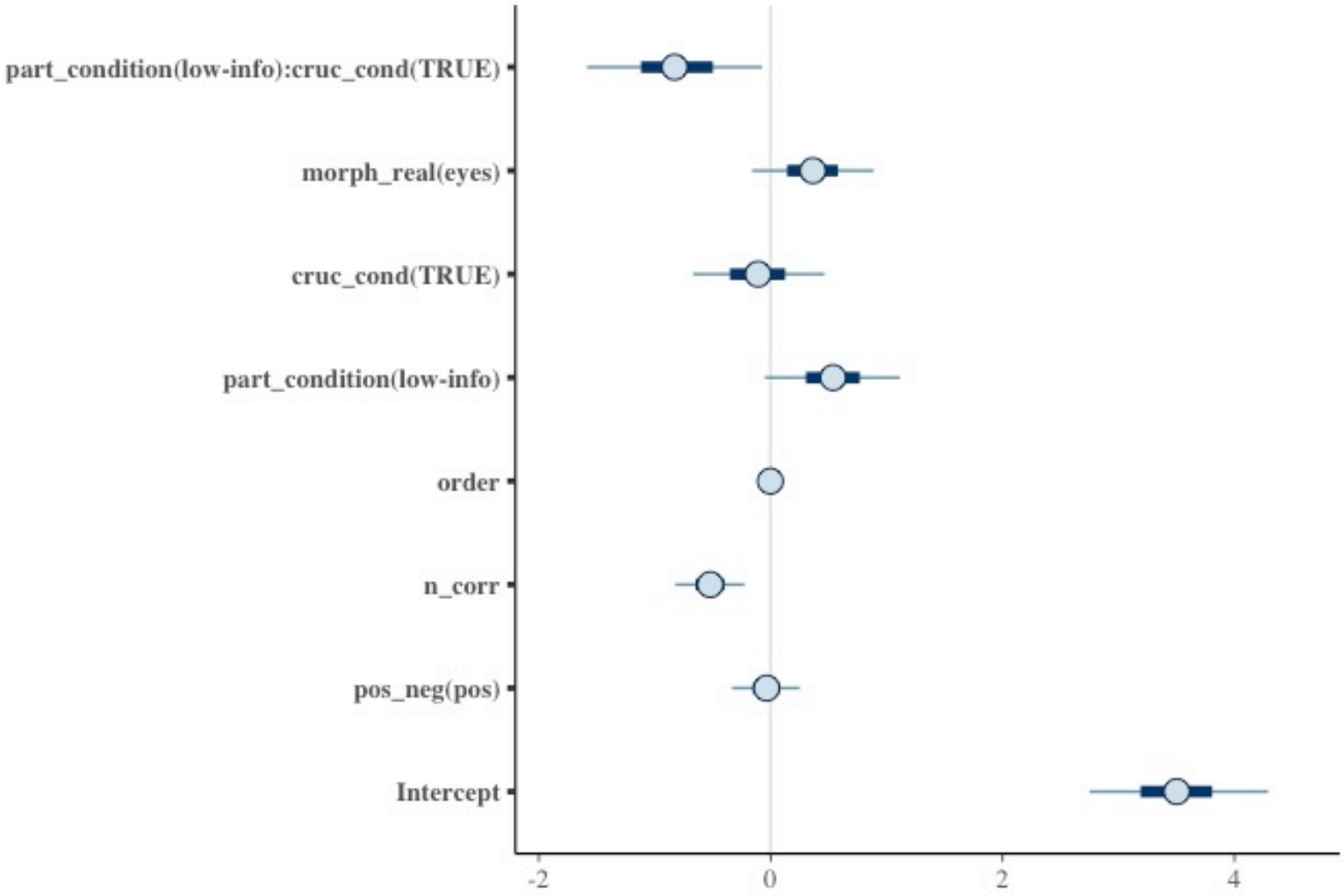
- Data points: ~4000
- Family: bernoulli (log link)



- Under these criteria it appears everyone is a maximal generaliser

Model 1C: images as data points + only affix matters

- What's going on with the interaction?



Model 1C: images as data points + only affix matters

- Difference not because low-info:TRUE is low, but because low-info:FALSE is high
- Also, absolute values are basically at ceiling, so this is not the difference we are looking for

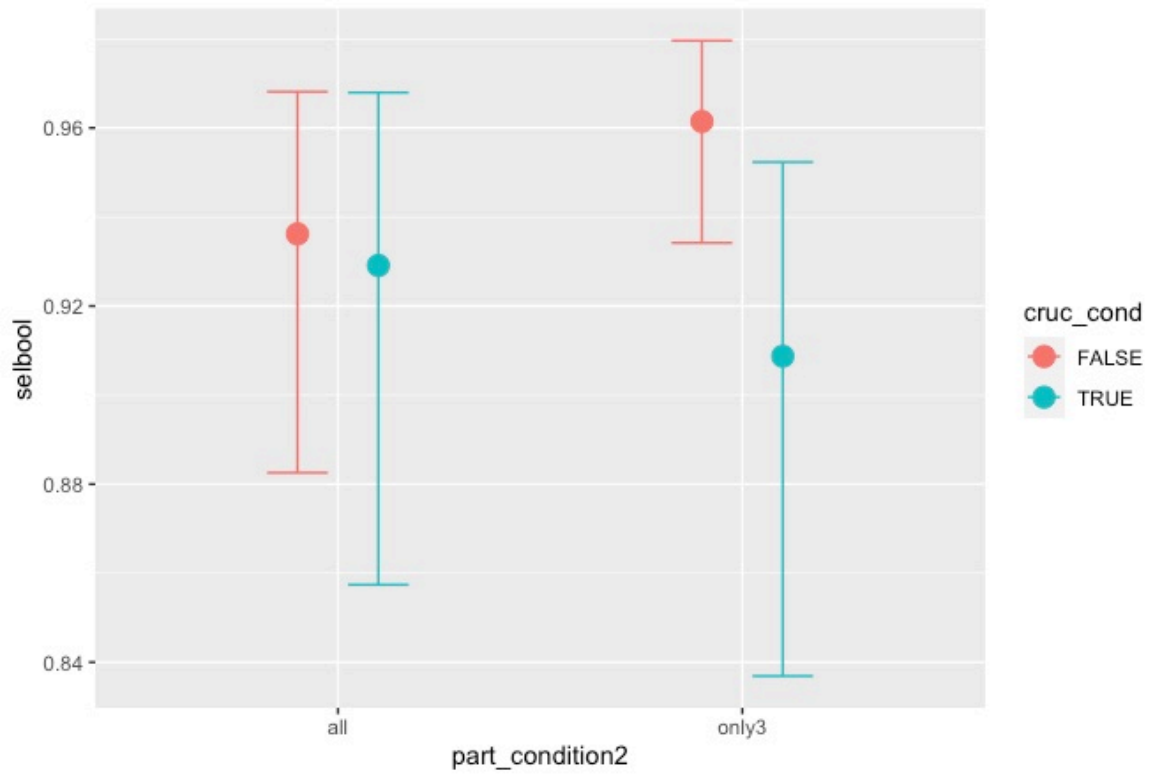


Figure 10: conditional effects plot of fitted values. Random effects excluded.

Interim Conclusions

Interim Conclusions

- So far, everyone seems to have adopted the maximal generalisation, by all metrics.
  - People seem to be willing to extend learned morphological associations to new objects (without being forced!), if a distinction is ignored elsewhere in the system
  - This is directly contra Harmon & Kapatsinski (2017) who, especially in choice tasks, found entrenchment, and contra Xu & Tenenbaum (2007)

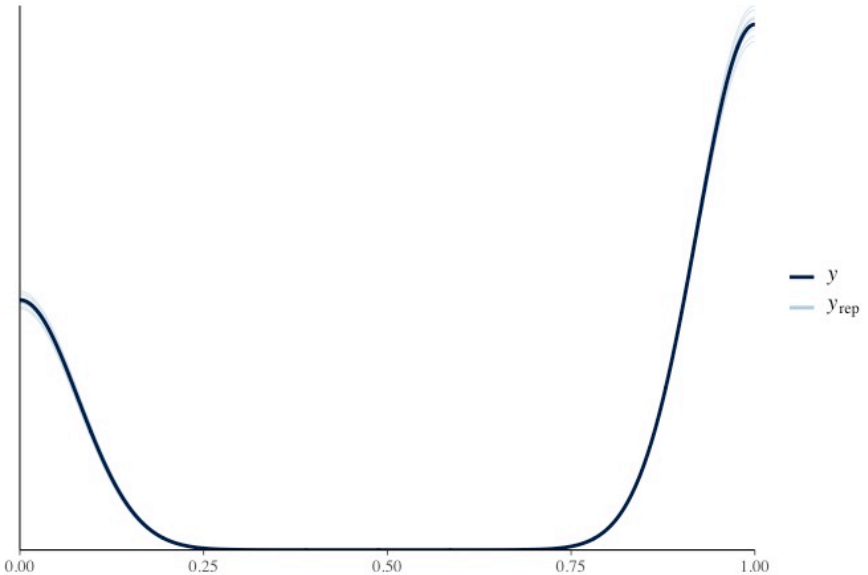
- But problems with models using images as data points
  - high-info participants are assigned a random crucial condition (=meaningless)
  - can't capture well the interdependency of data points in the same trial
  - answers the question indirectly

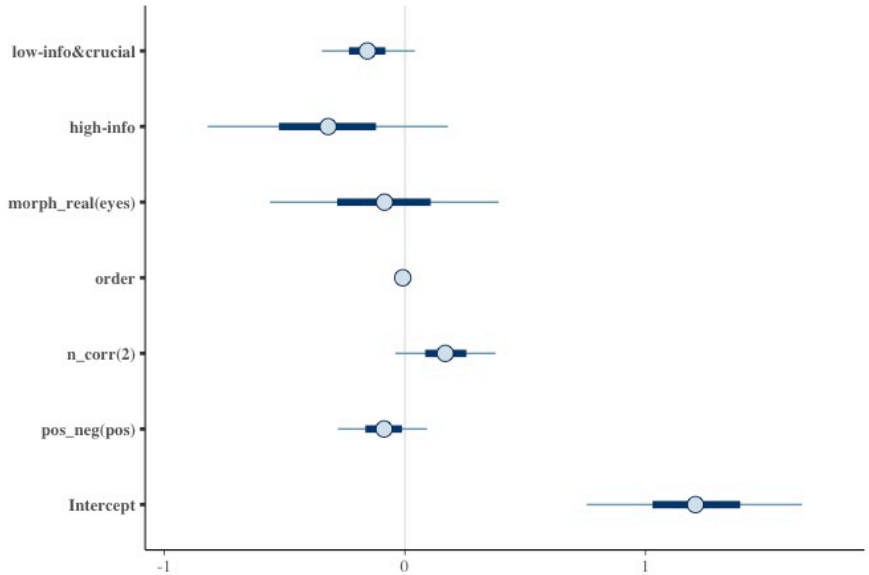
Recoding participant condition

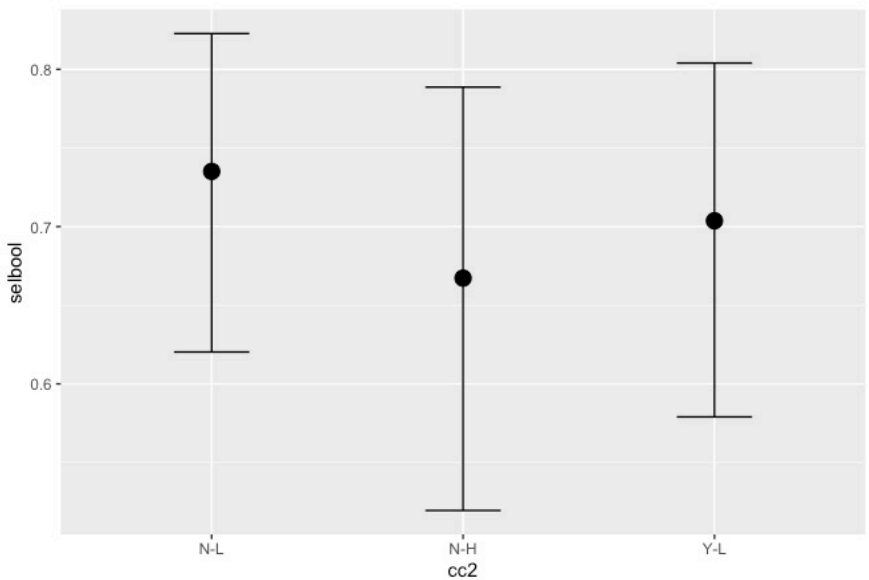
- Let's rethink part_condition*cruc_condition.
- instead of assigning a random condition as crucial for high-info participants, we can recode part_condition*cruc_condition as a three-way variable.
  - high-info participant
  - low-info participant, data point is crucial
  - low-info participant, data point is not crucial
- Let's redo models 1B and 1C with this variable instead of part_condition*cruc_condition
- we set[low-info, not crucial] as the reference level.
  - If low-info participants are maximal generalisers, we expect no meaningful difference between this and either of the other two conditions
  - If low-info participants are minimal generalisers, we expect a difference with [low-info, crucial], but not with [high-info,]

Mod2B: thee-way participant condition - affix+stem matter

$$\text{sel} \sim \text{pos\_neg} + \text{n\_valid} + \text{new\_part\_cond} + \text{morph\_real} + (1|\text{partID})$$

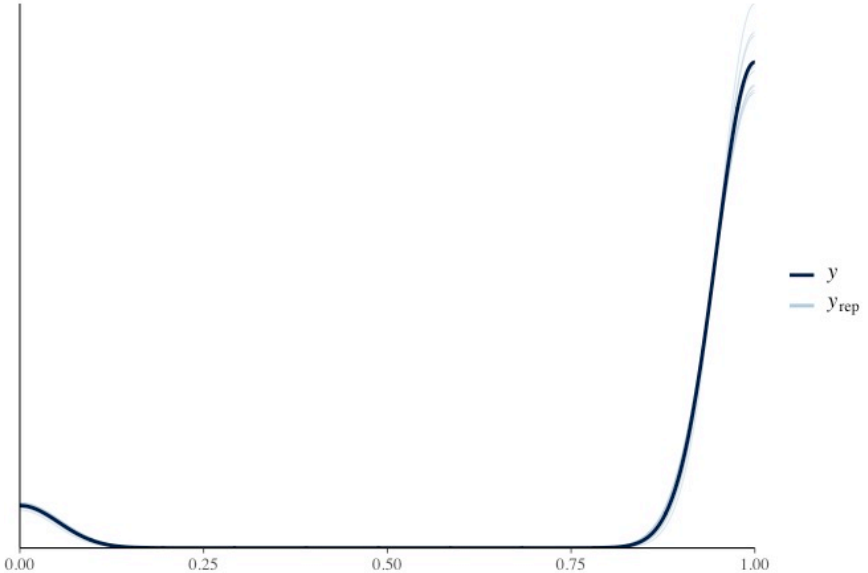


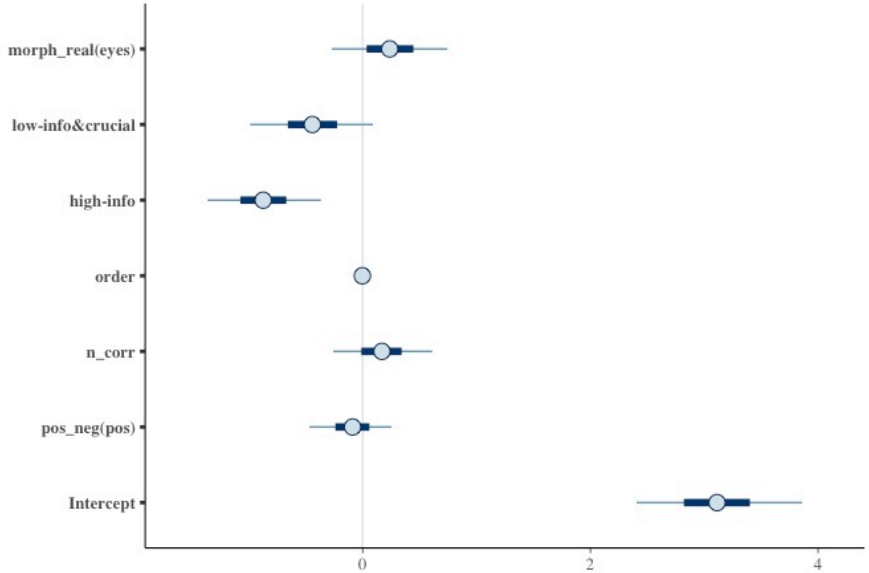


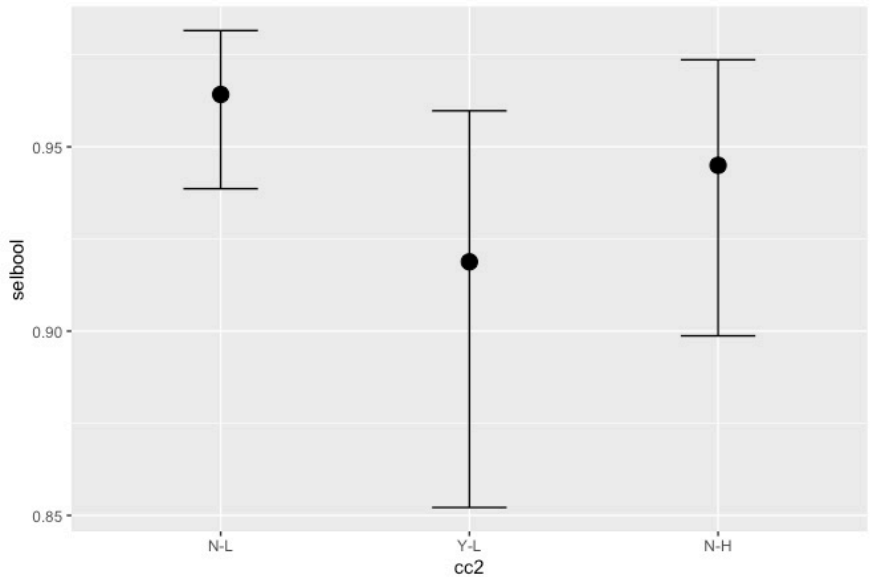


Mod2C: thee-way participant condition - only affix matters

$$\text{sel} \sim \text{pos\_neg} + \text{n\_valid} + \text{new\_part\_cond} + \text{morph\_real} + (1|\text{partID})$$







Three-way participant condition

- We still do not see a meaningful difference that would indicate low-info participants are minimal generalisers
- The puzzle of why low-info participants seem to do better than high-info ones

Trials as data points

- Using entire trials as data points allows us to ask the question more directly, by predicting minimal vs maximal generalising behaviour for each trial based on participant condition.

- Unfortunately, it appears impossible to combine the benefits of trials as data-points with an analysis that doesn't assign a meaningless crucial condition to high-info participants

- Select only trials for which the crucial condition is a valid choice, predict whether generalising behaviour is minimal vs maximal based on participant condition.
  - But again runs into the problem of meaninglessly assigning a crucial condition to high-info participants

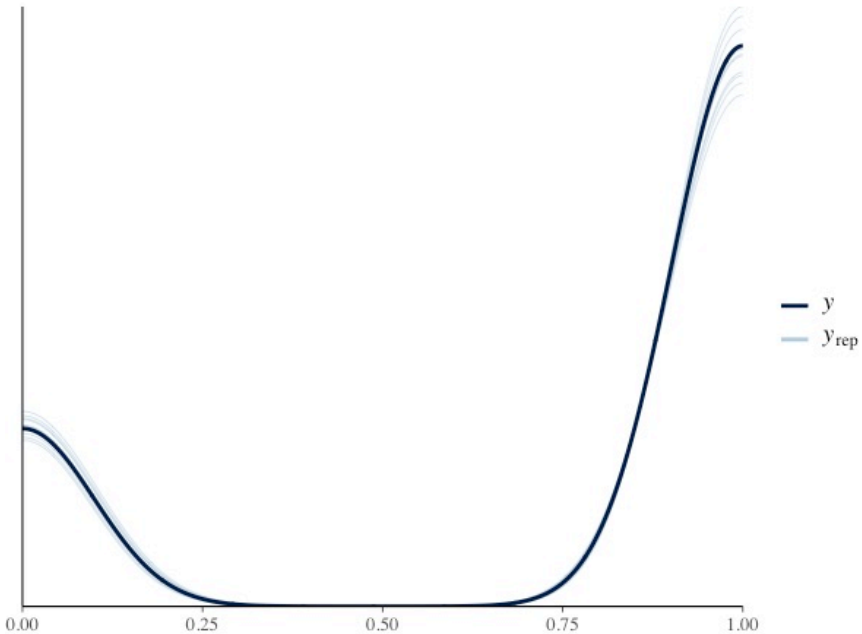
Model 3B: trials as data points - affix+stem matters

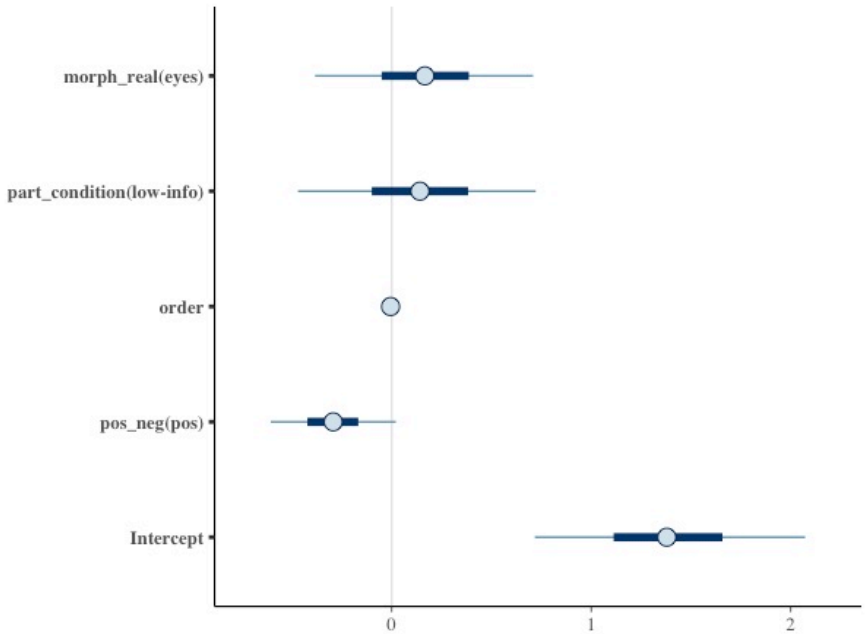
- Only trials containing crucial condition as a valid answer are considered
- Trials are coded as maximal, minimal (if crucial condition could have been selected, has it?) or inconsistent (inconsistent ones are discarded)

$$\text{max\_min} \sim \text{pos\_neg} + \text{n\_valid} + \text{part\_condition} + \text{morph\_real} + (1|\text{partID})$$

- Data points: 880

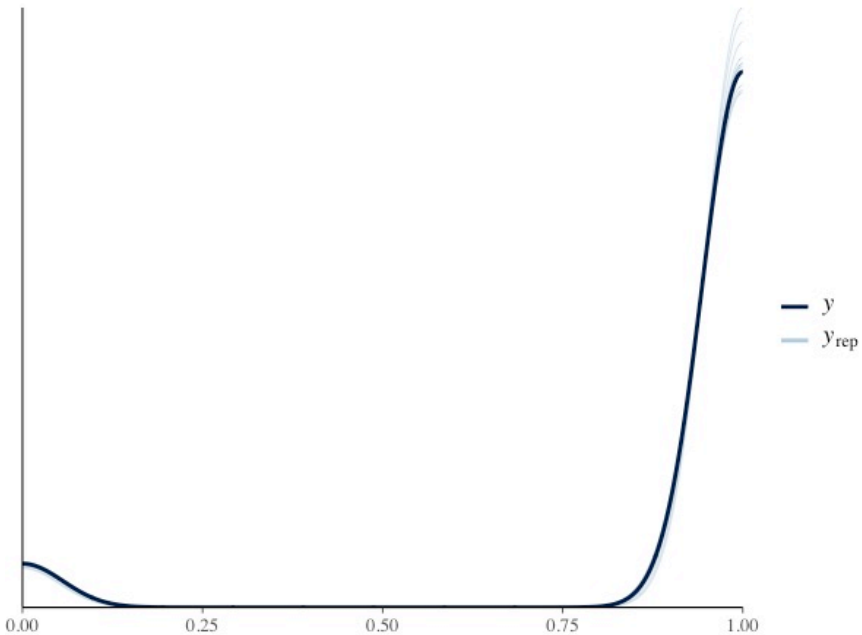
Model 3B: trials as data points - affix+stem matters

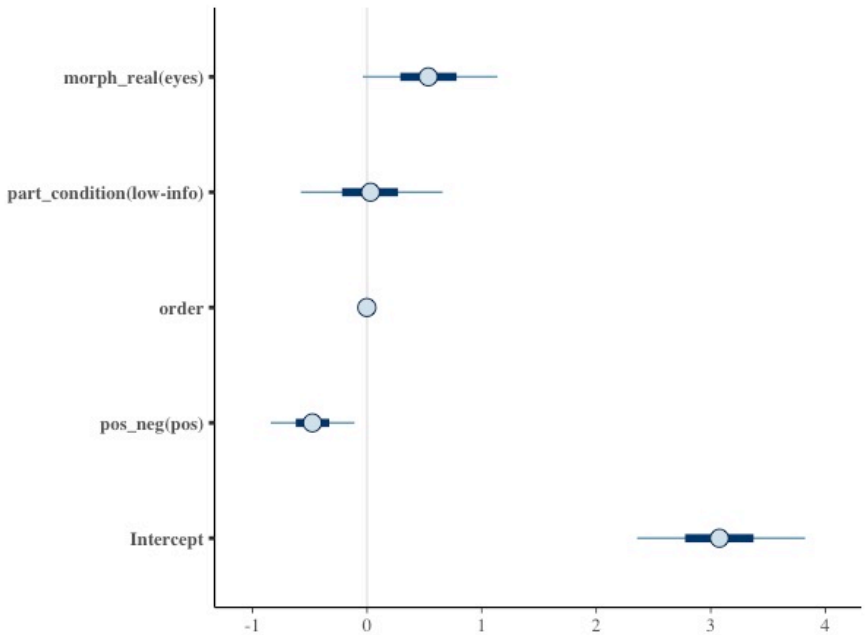




Model 3C: trials as data points - only affix matters

- Data points: 1300





Trials as Datapoints

- The same picture is reaffirmed: participant condition doesn't matter, everyone is a maximal generaliser.
- Participants readily generalise a form-meaning mapping in a way that parallels the behaviour of other elements in the system (we like our paradigmatic systems symmetric and orthogonal)

Conclusion

Conclusion

- The methodology
  - The artificial language material designed has a number of upsides.
    - Interesting to participants
    - Aware of morphological typology
    - Fully customisable across multiple dimensions
  - The testing paradigm is different from predecessors in that it doesn't force a choice. This appears to yield different results from the literature.
  - At least for this specific question, multiple models are needed to give a full picture.
- The results
  - Counter to previous literature, it appears speakers are willing to freely extend the meaning of an affix, even when not strictly necessary, if it parallels the distribution of another affix in the system.

Next steps

- Vary frequency of exposure to different conditions
- Vary the way morphological information is encoded in the form
  - Does the type of formal distinction matter?
  - Does the organisation of the system (e.g. asymmetry) matter?
- What is the role of production vs comprehension?
- What is the role of individual differences?