An efficiency-based effect of frequency on lexicalization

A dyadic experiment

Wataru Uegaki, Anne Mucha & Ciyang Qing

Waseda University Language Science Colloquium
February 12, 2025

Semantic universals, efficient communication, and dyadic language-learning experiments

Universal properties of human languages

Human languages exhibit certain universal properties.

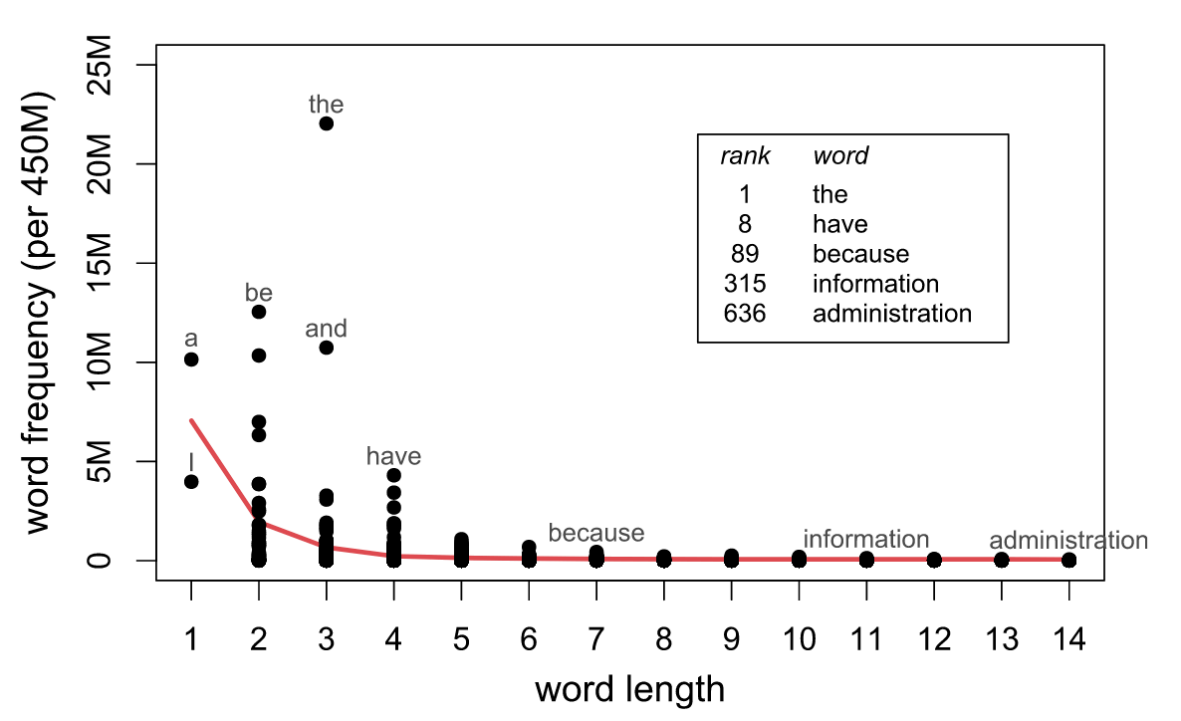
Overarching question: To what extent can the universal properties be accounted for by the idea that languages support efficient communication?

Alternatively: to what extent can they be attributed to innate biases independent from communication?

Zipf's law of abbreviation

Zipf's law of abbreviation: The more frequent a word is, the shorter it tends to be.

- Confirmed in a broad range of human langs Ferrer-i-Cancho & Hernández-Fernández, 2013; Sigurd et al. 2004; Strauss et al. 2007; Teahan et a. 2000); and
- Animal communication systems (Ferrer-i Cancho et al., 2013)



Zipf's law of abbrev. and communicative efficiency

Zipf's Law of Abbreviation (ZLA): The more frequent a word is, the shorter it tends to be.

Efficiency based explanation—Principle of Least Effort (PLE): The law is derived from the trade-off between pressures in terms of cognitive effort and communicative accuracy.

- Shorter forms incur less cognitive effort.
- Using longer forms allows for more accurate messages.
- How to use the available short forms optimally?
  Solution: allocate more **frequent** meanings to shorter forms (so the average token length will be shorter)

Dyadic language-learning experiments

Dyadic/interactive language-learning experiments can help test the explanation of universals in terms of efficient communication.

- Cognitive effort $\simeq$ processing/production cost associated with the task of learning and using a linguistic form
- Communicative accuracy $\simeq$ pressure from accurate communication in the interaction component

We can systematically manipulate these two pressures.
$\rightsquigarrow$ Only if the two pressures co-exist, the relevant universals emerge.

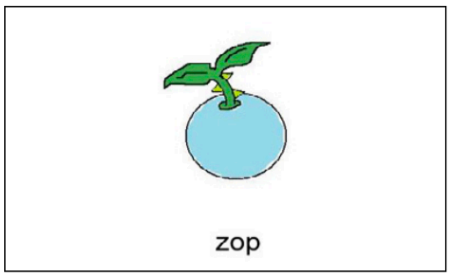
Kanwal et al. 2017: design

Kanwal et al. 2017: ZLA emerges only if we combine pressure from time and interaction.

a) input frequencies

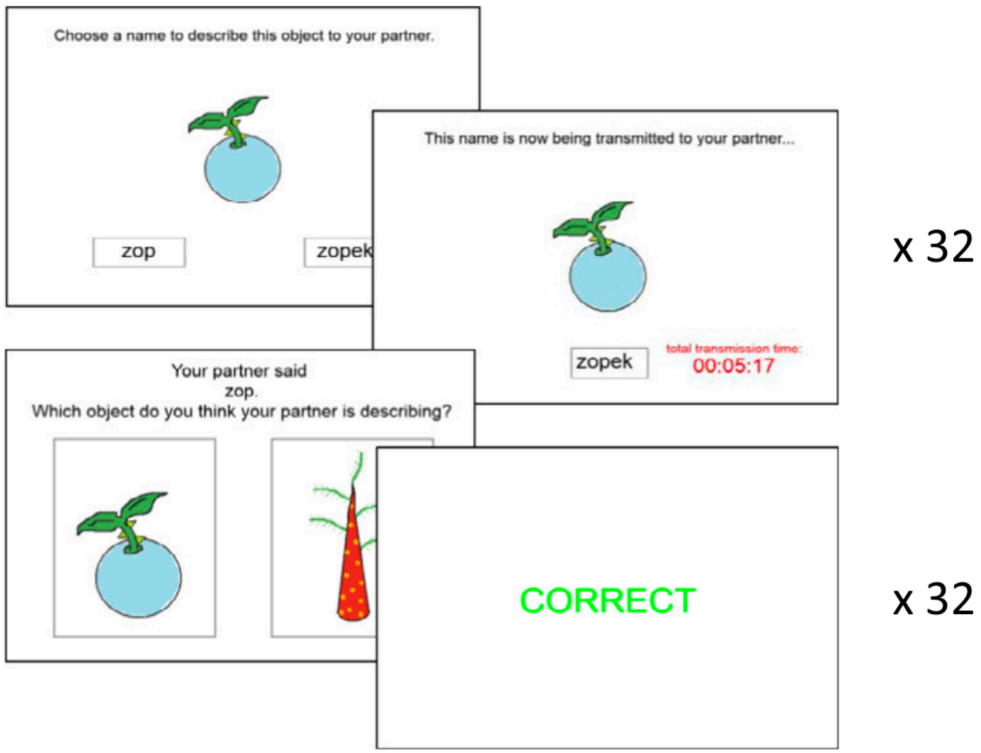
x 4 zopekil	x 4 zop
x 12 zopudon	x 12 zop

b) training trial format



x 32

c) testing trial format

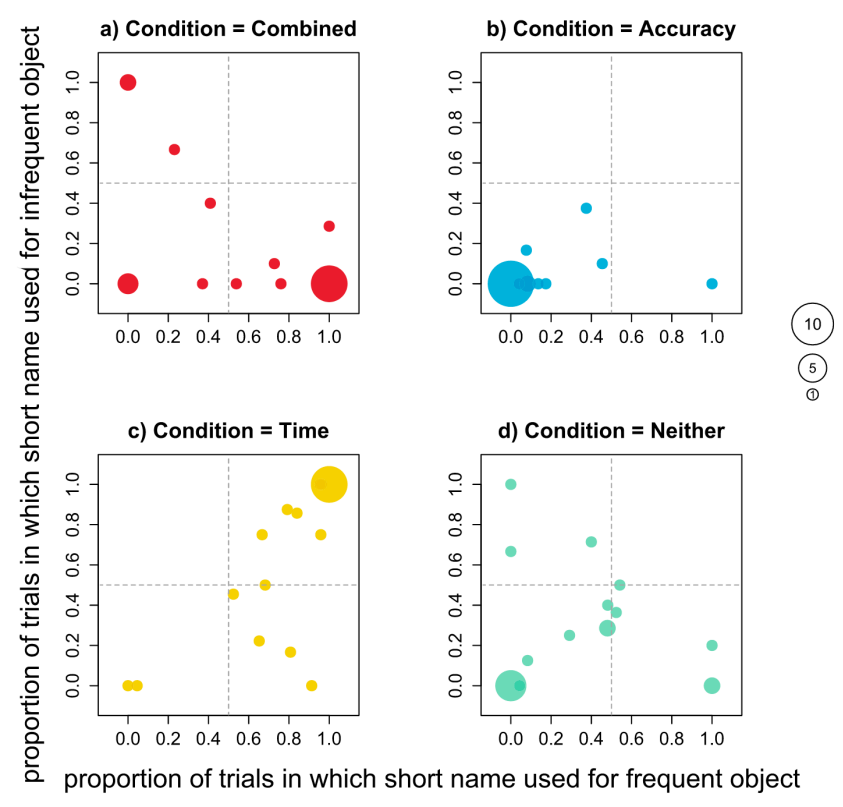


x 32

x 32

Kanwal et al. 2017: results

Kanwal et al. 2017: ZLA emerges only if we combine pressure from time and interaction.



Direct exp. support for the explanation of ZLA based on PLE.

Generalizing Principle of Least Effort

But, word length is only one possible source of effort.

Principle of Least Effort Generalized: languages optimizes the trade-off between communicative accuracy and cognitive effort arising from dimensions of *linguistic complexity in general*, not just from word length.

We pursue this general interpretation of PLE by testing its prediction about competition between lexical and compositional forms.

Outline

Principle of Least Effort and Lexical/Compositional distinction

Principle of Least Effort and Lexical/Compositional distinction

Our hypothesis

Our hypothesis

If a language contains a compositional form F_c and a lexical form F_l as candidate forms for a frequent meaning M_{freq} and an infrequent meaning M_{inf}, the combined effect of accuracy and effort results in the mapping:

$$F_l \mapsto M_{freq}$$
$$F_c \mapsto M_{inf}$$

Or: Under the combined effect of accuracy and effort, *the more frequent a meaning is, the more it tends to be expressed as a lexical form*, as opposed to a compositional form.

PLE derives the hypothesis

- Using a lexical form incurs less cognitive effort than using a compositional form.
- Using compositional forms allows for more accurate messages.
- How to use the available lexical forms optimally?
  Solution: allocate more **frequent** meanings to lexical forms.

Cognitive effort and compositional/lexical

Using a compositional form is more cognitively effortful than using a lexical form because:

- It involves retrieving multiple stored lexical items;
- It requires accurate syntactic and semantic composition of lexical items;
- (It requires recognizing covert scopal configurations a given composition allows)

Principle of Least Effort and Lexical/Compositional distinction

Empirical manifestation of the hypothesis

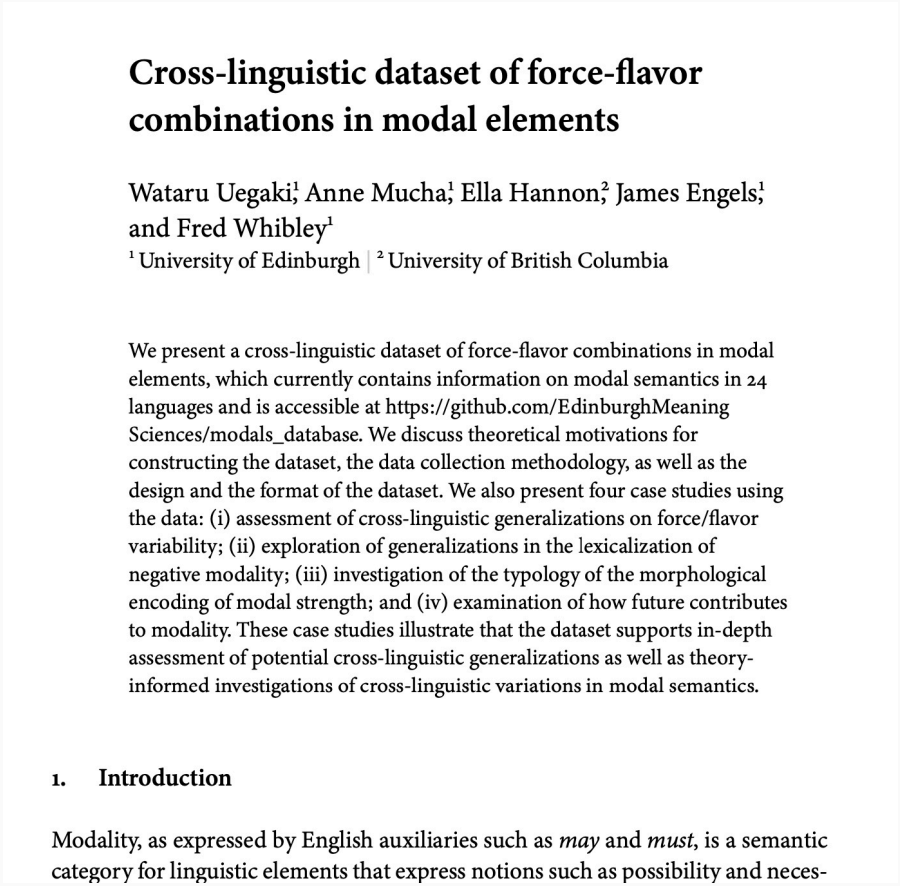
Crosslinguistic study on modals

24 languages (10 language families, 23 genera, 4 isolates)

We adapted **Vander Klok's (2021) revised modal questionnaire** for crosslinguistic **fieldwork**

- comparable data, suitable for detecting crosslinguistic patterns
- systematic elicitation of force/flavor combinations

... and added **contexts for eliciting negative modality**: *non-necessity* ($\neg\Box$ p) and *impossibility* ($\neg\Diamond$ p) with **epistemic, deontic, teleological, pure circumstantial flavor**



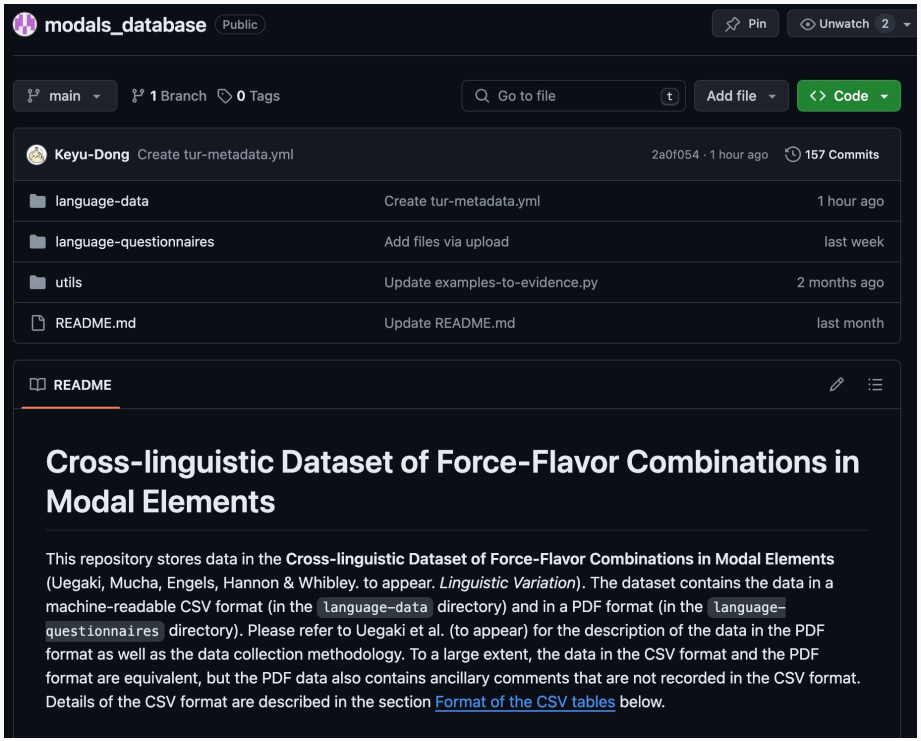
Cross-linguistic dataset of force-flavor combinations in modal elements

Wataru Uegaki[1], Anne Mucha[1], Ella Hannon[2], James Engels[1], and Fred Whibley[1]
[1] University of Edinburgh | [2] University of British Columbia

We present a cross-linguistic dataset of force-flavor combinations in modal elements, which currently contains information on modal semantics in 24 languages and is accessible at https://github.com/EdinburghMeaning Sciences/modals_database. We discuss theoretical motivations for constructing the dataset, the data collection methodology, as well as the design and the format of the dataset. We also present four case studies using the data: (i) assessment of cross-linguistic generalizations on force/flavor variability; (ii) exploration of generalizations in the lexicalization of negative modality; (iii) investigation of the typology of the morphological encoding of modal strength; and (iv) examination of how future contributes to modality. These case studies illustrate that the dataset supports in-depth assessment of potential cross-linguistic generalizations as well as theory-informed investigations of cross-linguistic variations in modal semantics.

1. Introduction

Modality, as expressed by English auxiliaries such as *may* and *must*, is a semantic category for linguistic elements that express notions such as possibility and neces-



- Uegaki, Mucha, et al. (2024) in *Linguistic Variation*
- `https://github.com/EdinburghMeaningSciences/modals_database`

Modal flavors

(1) **Epistemic**
(In view of the available evidence,) John *may* / *must* be the murderer.

(2) **Deontic**
(In view of his parents' orders,) John *may* watch TV, but he *must* go to bed at 8 pm.

(3) **Circumstantial / ability**
(In view of his physical abilities,) John *can* lift 200 lbs.

(4) **Teleological**
(In view of his goal to get a PhD,) John *must* write a dissertation.

Deontic Priority Generalization

Generalization: Lexicalized impossibility modals tend to express non-epistemic (specifically deontic modality), more likely than epistemic modality.

	Impossibility	
	epistemic	deontic
Pattern A	✓	✓
Pattern B	×	✓
*Pattern C	✓	×

Table 1: ✓ means the meaning is lexicalized, × means it is not

- Pattern A: Basque, Turkish
- Pattern B: Korean, Hausa, Thai, Hungarian, Hebrew, Russian

Concrete examples (a: Hausa; b: Thai)

(5) Context (*deontic*): Visiting hours in a hospital.

a. **Kada** maziyarta su wuce ƙarfe 6 na yamma
¬♢ visitors 3pl.sbjv stay hour 6 pm

b. Yardpubuay **harm** yuu lang hok morng
patient.visitor ¬♢ exist after six o'clock
"Visitors can't stay after 6pm."

(6) Context (*epistemic*): Ben swims every day at this time.

a. **Ba** zai **yiwu** Ben ya kasance a gida **ba**
NEG 3sg.fut MODAL(♢) Ben 3sg be at house NEG

b. Ben yuu tii baan **mai dai**
Ben exist LOC house NEG MODAL(♢)
"Ben can't be at home."

Frequency of epistemic vs. deontic modals

The generalization may be an empirical manifestation of our hypothesis—the more frequent a meaning is, it tends to be expressed as a lexical form.

- Corpus studies suggest that epistemic uses of modals are less frequent than non-epistemic uses (e.g. Van Dooren et al. 2022; Veselinovic 2019)
- deontic impossibility in particular is expressed more frequently than epistemic impossibility (Bell et al., 2024).

Frequency data in child-directed speech

Bell et al. (2023): Force-flavour annotation of English child-directed speech for two 3/4yo children in CHILDES.

Flavour	**Possibility**	**Necessity**	**Impossibility**
Deontic	110	86	39
Epistemic	74	117	25

A dyadic experiment

A dyadic experiment

Design

Dyadic experiment — idea

Goal: Test if lexicalization of a frequent meaning occurs as a result of the combined effect of interaction and cognitive cost.
$\rightsquigarrow$ Design a language game where...

- there are two synonymous forms: lexical and compositional
- both forms are, by design, ambiguous between two meanings: a more frequent and less frequent

Expectation: When participants interact in this language, they will converge on a language where

- $[\![\text{lex}]\!]$ = the frequent meaning
- $[\![\text{comp}]\!]$ = the infrequent meaning

Dyadic experiment — basic design

(7) a. Nothing in this picture is blue. (lexical)
b. Everything in this picture is not blue. (compositional)

These two messages are in principle compatible with two kinds of domains: circles and triangles, where one shape is more frequent than the other. Communication succeeds when both force and domain are conveyed accurately.

- Using *nothing* for both circles and triangles is less cognitively costly, but will result in domain-ambiguity
- Using *every...not* along with *nothing* allows disambiguation between shapes.
- **Solution**: allocate the more frequent shape to *nothing*.

Conditions

4 conditions that vary in 2 variables:

[± Partner] Whether there is an interaction with a partner.

[± Distractor] Whether there is a "distractor" in the task that incurs an additional cognitive effort for using the compositional form.

Main prediction: In the [+Partner, +Distractor] condition, participants tend to map *nothing* to the more frequent shape and *everything...not* to the more infrequent object. This tendency is reduced if either of the manipulations is removed.

Preregistration.

Stimuli sentences and pictures

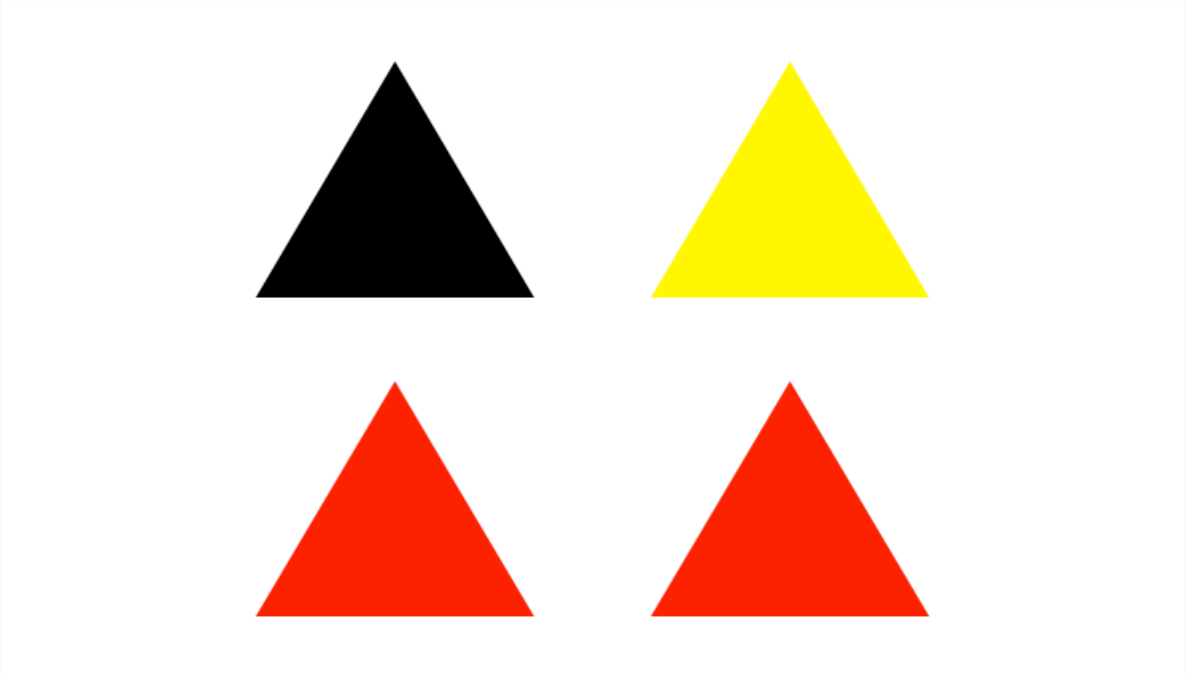
Sentences:

1. Lexicalized negative existential (nex)
   E.g., ***Nothing*** *in this picture is red.*
2. Compositional negative existential (nex)
   E.g., ***Everything*** *in this picture is* ***not*** *red.*
3. **In [+dist]:** distractor weak
   E.g., ***Not everything*** *in this picture is red.*
   In [–dist]: non-distractor weak:
   There are multiple objects in this picture.

Pictures:

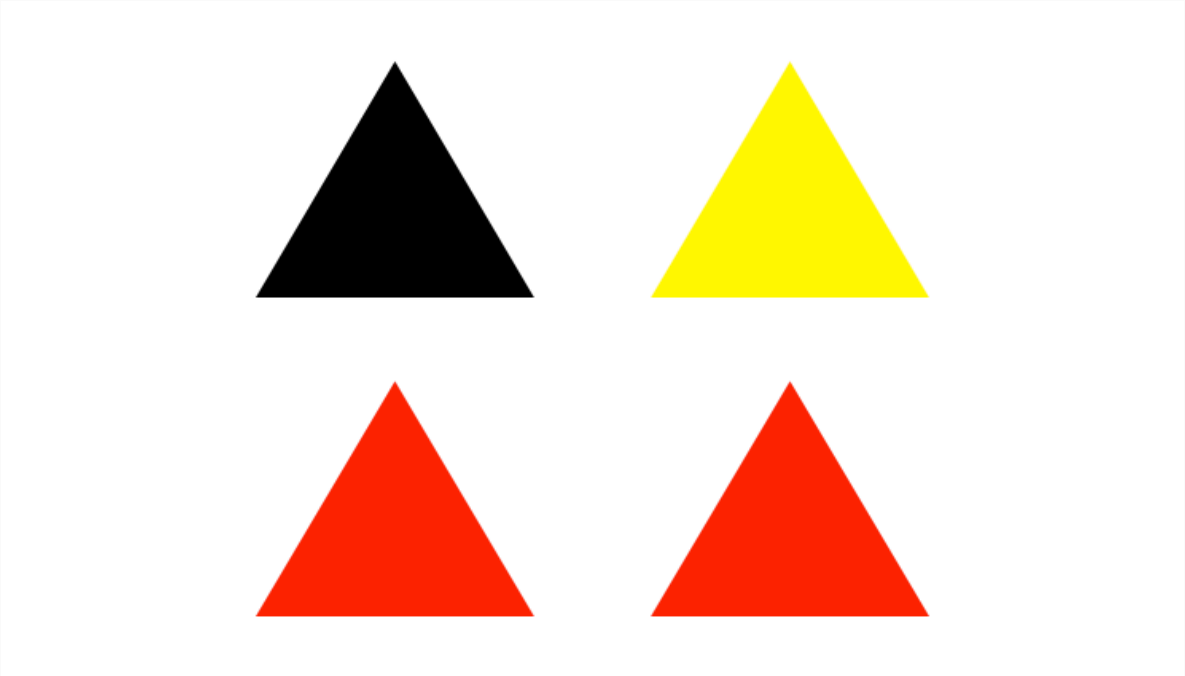
1. *circles* where nex is true (henceforth denoted as $\neg\exists_{\bigcirc}$);
2. *triangles* where nex is true ($\neg\exists_{\triangle}$);
3. *circles* where only the weak sentence is true ($\neg\forall_{\bigcirc}$);
4. *triangles* where only the weak sentence is true ($\neg\forall_{\triangle}$).

Stimulus (non-existence; triangle; lexical)



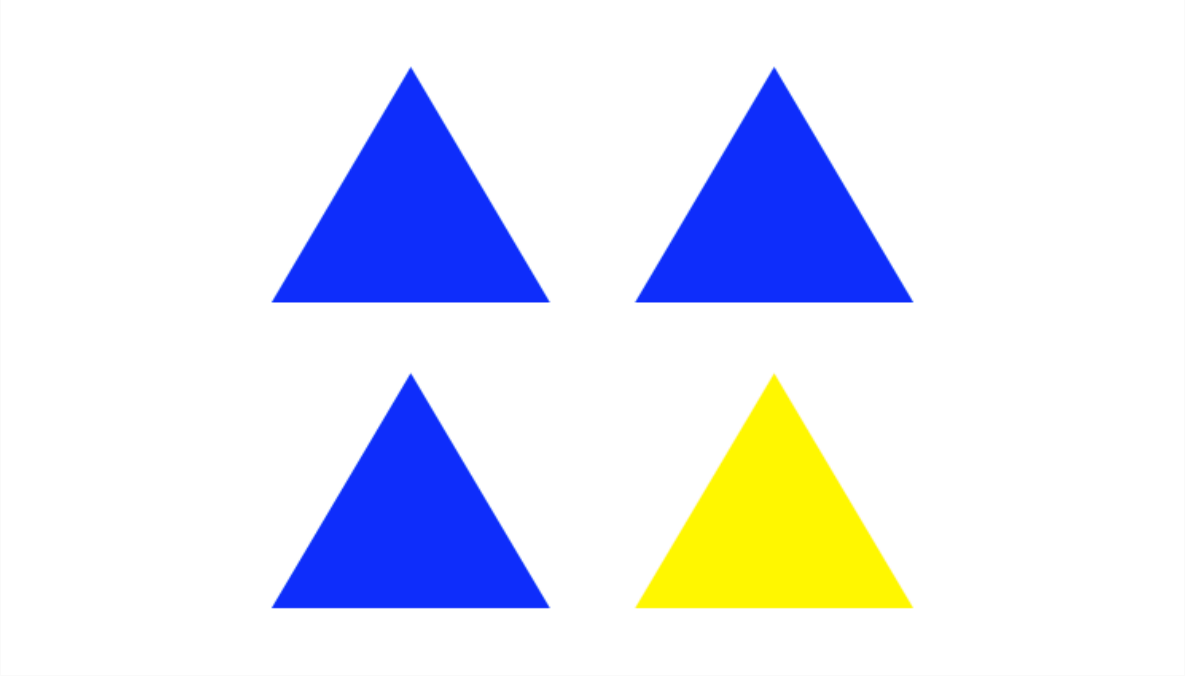
Nothing in this picture is blue.

Stimulus (non-existence; triangle; comp.)



Everything in this picture is not blue.

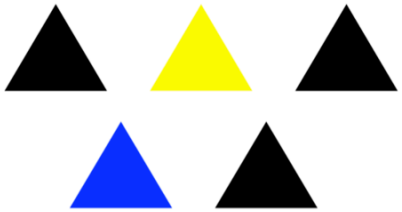
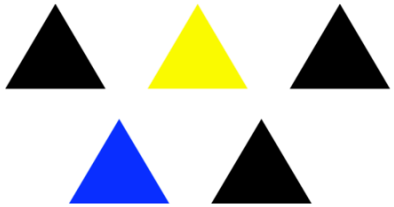
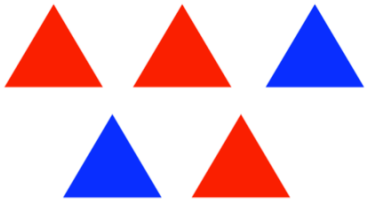
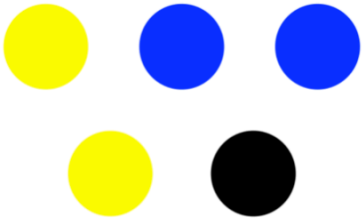
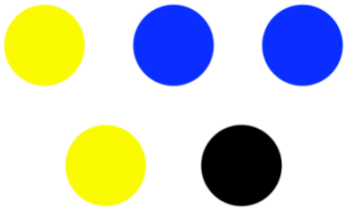
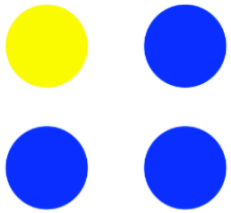
Stimulus (non-universal; triangle)



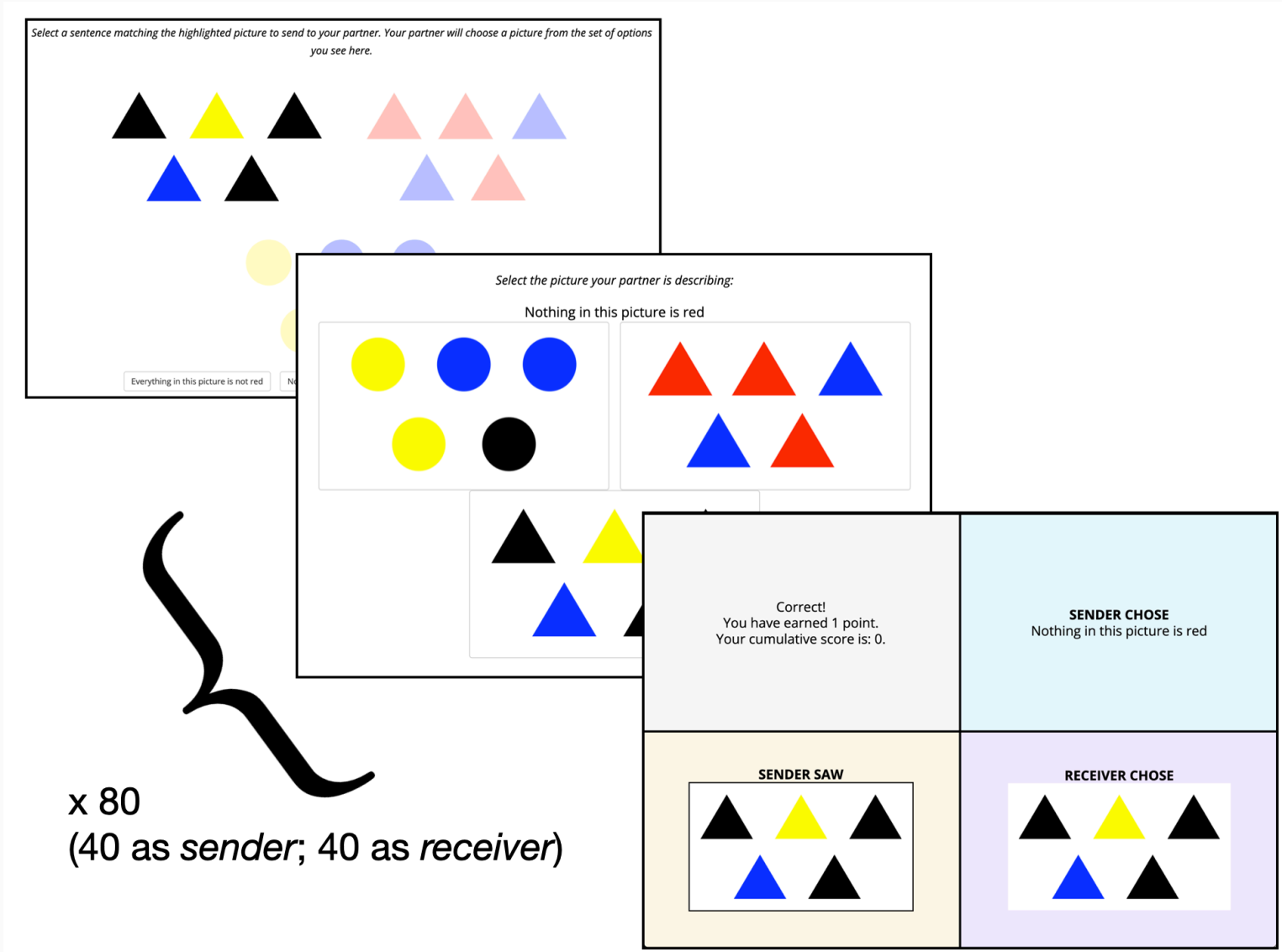
Not everything in this picture is blue (distractor)

There are multiple objects in this picture (non-distractor)

Exposure format and frequency

¬∃ pic with lexical nex form	¬∃ pic with compositional nex form	¬∀ pic with weak form
 *Nothing in this picture is red* x 12	 *Everything in this picture is not red* x 12	 *Not everything in this picture is red* x 6
 *Nothing in this picture is red* x 4	 *Everything in this picture is not red* x 4	 *Not everything in this picture is blue* x 2

Testing format in [+Partner] conditions



DEMO

Conditions

4 conditions that vary in 2 variables:

[± Partner] Whether there is an interaction with a partner.

[± Distractor] Whether there is a "distractor" in the task that that incurs an additional cognitive effort for using the compositional form.

Predictions: In the [+Partner, +Distractor] condition, participants tend to map *nothing* to the more frequent shape and *everything...not* to the more infrequent object. This tendency is reduced if either of the manipulations is removed.

Distractor and cognitive effort

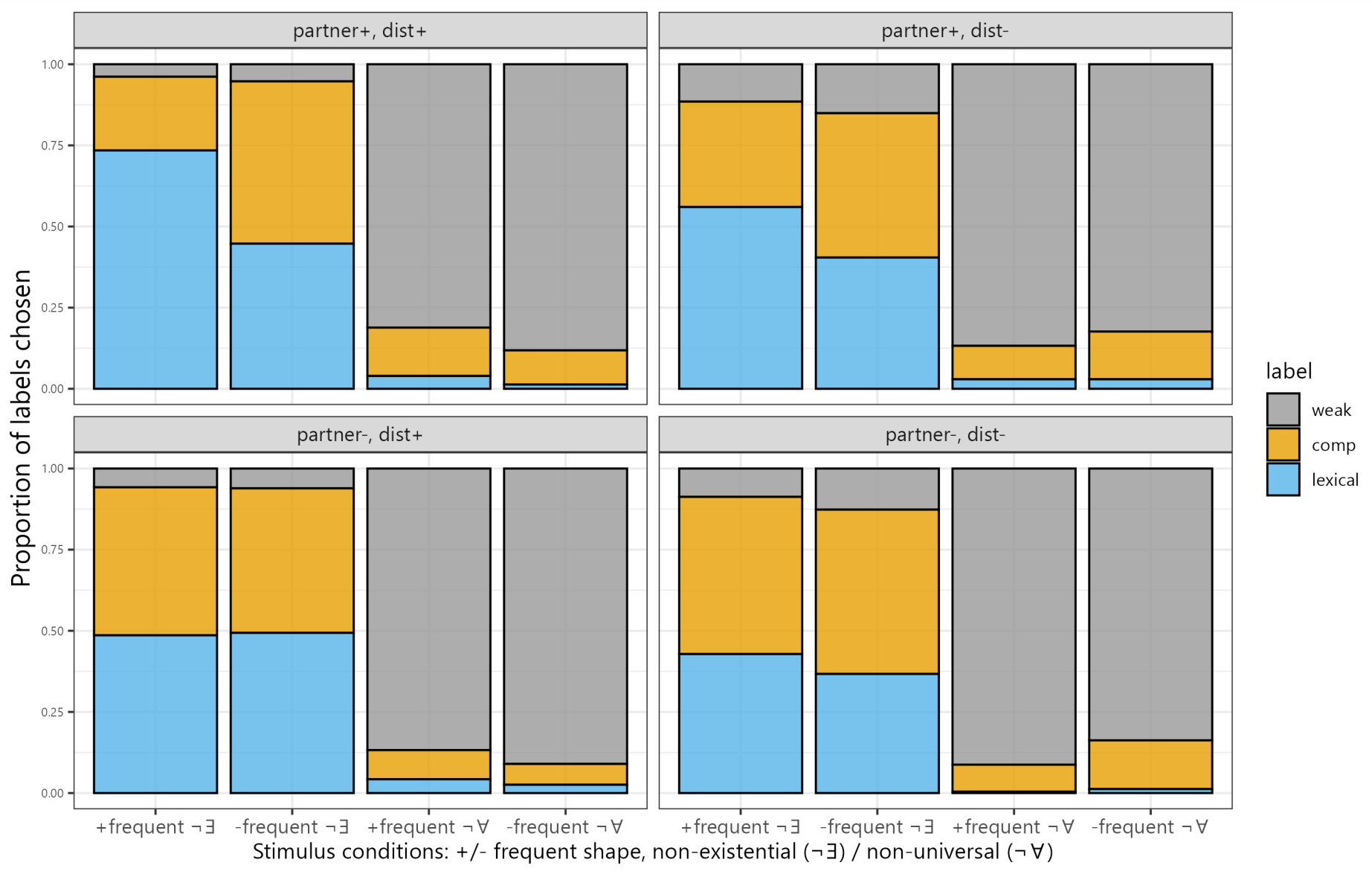
The presence of the distractor weak form in [+dist] conditions is designed to increases the cognitive effort for choosing the *compositional* nex form, relative to the *lexical* nex form.

- The distractor (*Not everything...*) is formally similar to the compositional nex form (*Everything...not...*)
- Mapping the forms to correct meanings requires awareness of their compositional structures, not merely recognizing them as a collection of words

In contrast, the non-distractor weak sentence is not formally similar to compositional nex.

A dyadic experiment

Results and discussion



N = 151 (38 [+part,+dist]; 34 [+part,–dist]; 39 [–part,+dist]; 40 [–part;–dist])

Main results

Three significant main contrasts:

- The preference for lexicalizing the more frequent object is reduced if we remove the distractor ($\beta = -0.877, p < .001$).
- The preference for lexicalizing the more frequent object is reduced if we remove the partner ($\beta = -1.697, p < .001$).
- The effect of reducing the distractor is stronger if the partner is present than when the partner is absent ($\beta = 1.19, p < .001$).

Exploratory analysis: the preference for lexicalization for the frequent object is greater in [+part, –dist] than in [–part, –dist] ($\beta = 0.502, p = 0.036$).

Discussion

- In the [+part; +dist] condition, the participants prefer lexicalizing the more frequent meaning.
- This preference is significantly reduced if we remove either of the pressures.
- Unlike previous dyadic experiments on PLE which uses time pressure, the current experiment shows that cognitive effort can be manipulated in terms of the formal structures of the linguistic stimuli.
- Even when the distractor is absent, having a partner significantly increases the preference for lexicalizing the more frequent object. $\rightsquigarrow$ The cognitive effort for the compositional form exists in the absence of the distractor.

Conclusions

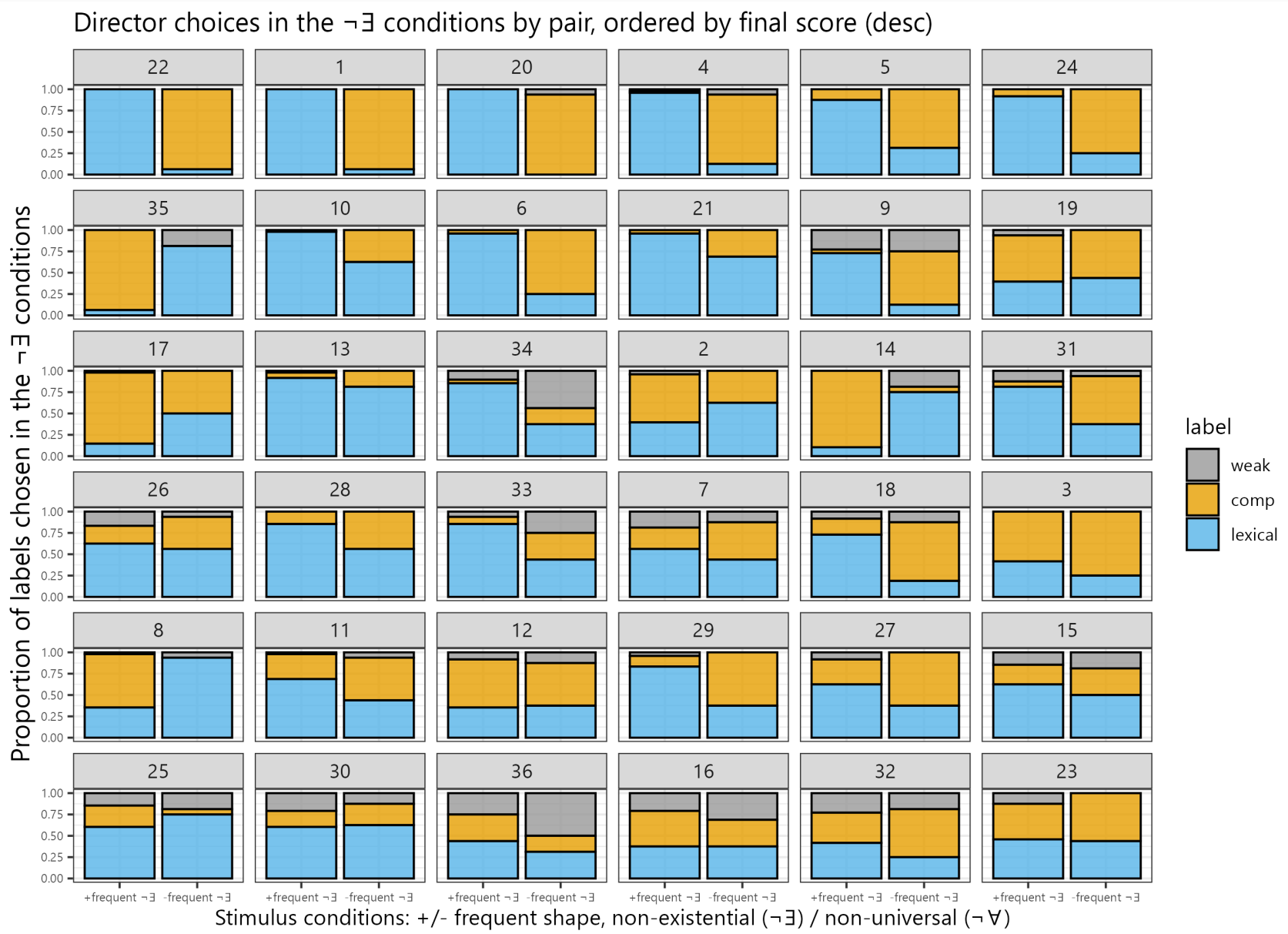
Conclusions

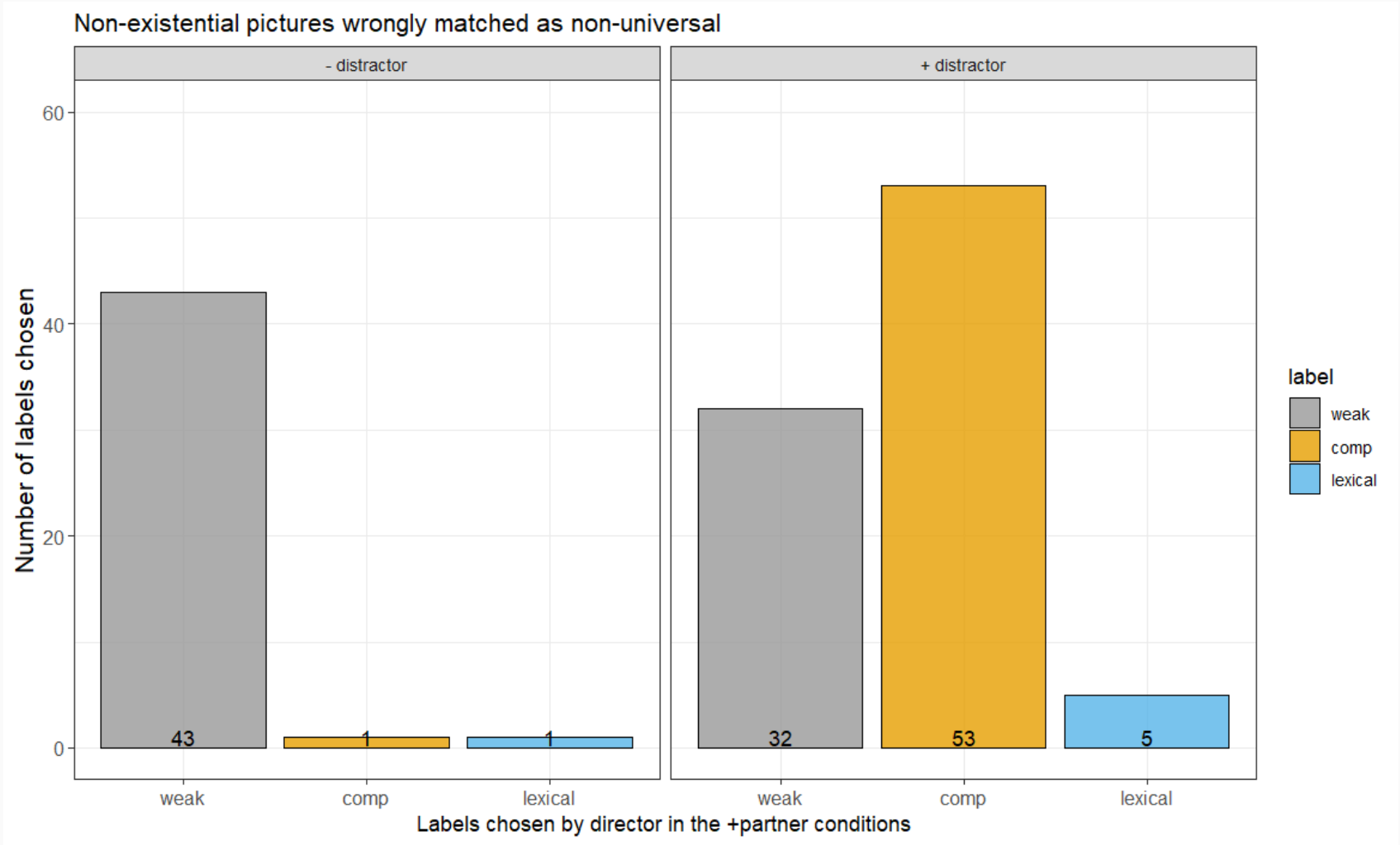
- Overarching question: To what extent can universals be accounted for by the idea that languages support efficient communication?
- Our dyadic communication experiment provides direct experimental evidence that efficiency considerations in terms of communicative accuracy and cognitive effort jointly shape lexicalization, biasing speakers toward using lexical forms for more frequent meanings over compositional alternatives.
- New experimental support for the effect of efficiency in lexicalization in natural language, aligned with existing analysis of cross-linguistic data (e.g. Khishigsuren et al., 2025; Xu et al., 2024)

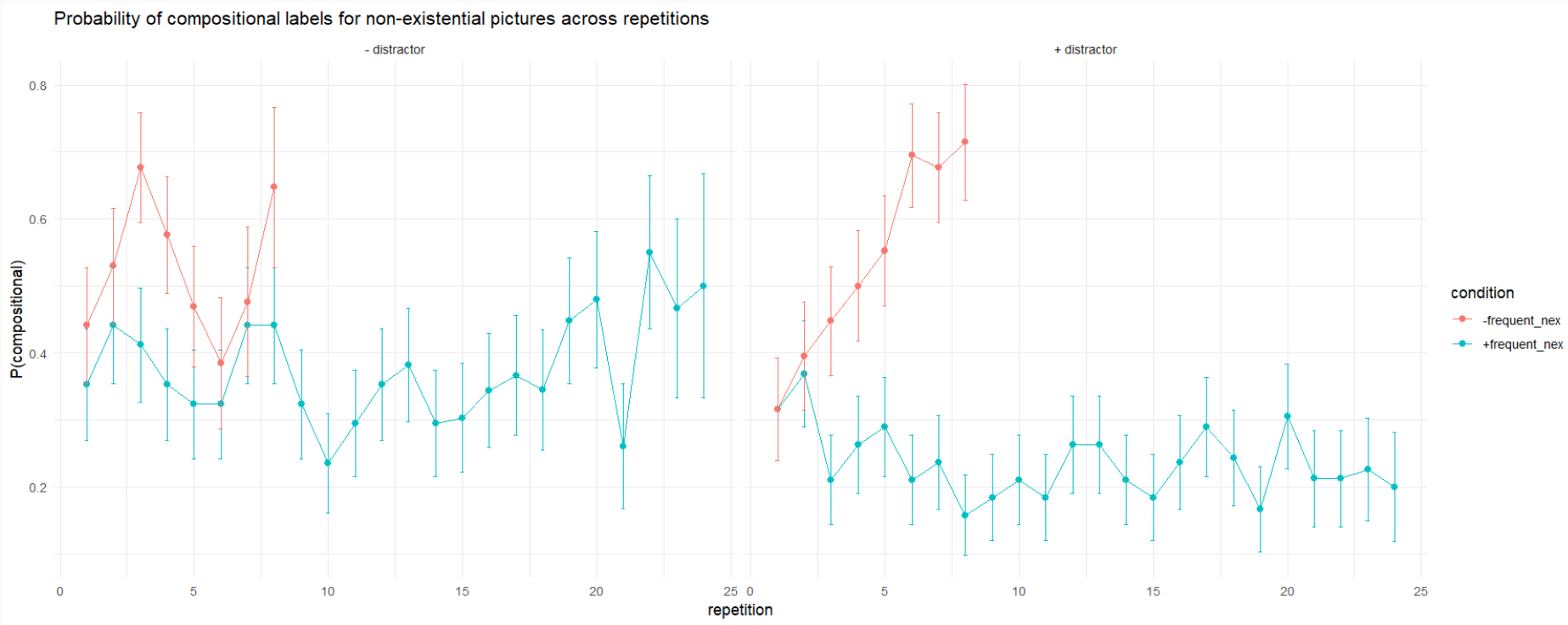
The end

Thank you!

Supported by UKRI Future Leaders Fellowship (Ref: MR/V023438/1).


Director choices in the ¬∃ conditions by pair, ordered by final score (desc)
Proportion of labels chosen in the ¬∃ conditions
22
1
20
4
5
24
35
10
6
21
9
19
17
13
34
2
14
31
26
28
33
7
18
3
8
11
12
29
27
15
25
30
36
16
32
23
1.00
0.75
0.50
0.25
0.00
+frequent ¬∃
-frequent ¬∃
Stimulus conditions: +/- frequent shape, non-existential (¬∃) / non-universal (¬∀)
label
weak
comp
lexical


Non-existential pictures wrongly matched as non-universal
- distractor
+ distractor
Number of labels chosen
60
40
20
0
43
1
1
32
53
5
weak
comp
lexical
weak
comp
lexical
Labels chosen by director in the +partner conditions
label
weak
comp
lexical


Probability of compositional labels for non-existential pictures across repetitions
- distractor
+ distractor
P(compositional)
0.8
0.6
0.4
0.2
0
5
10
15
20
25
repetition
condition
-frequent_nex
+frequent_nex