Exploring lexical dynamics using diachronic corpora and artificial language experiments

Andres Karjus
CUDAN lab, Tallinn University

In collaboration with: Kenny Smith, Richard A. Blythe, Simon Kirby (University of Edinburgh)

Colloquium for Computational Linguistics and Linguistics in Stuttgart
12.01.2021

- Started postdoc in:

http://cudan.tlu.ee

- PhD from:

All living languages keep changing

- All the time
- Eventually diverge into different languages
- This is weird

- This research: focus on lexical change and competition therein
- What happens when new words are introduced into language?

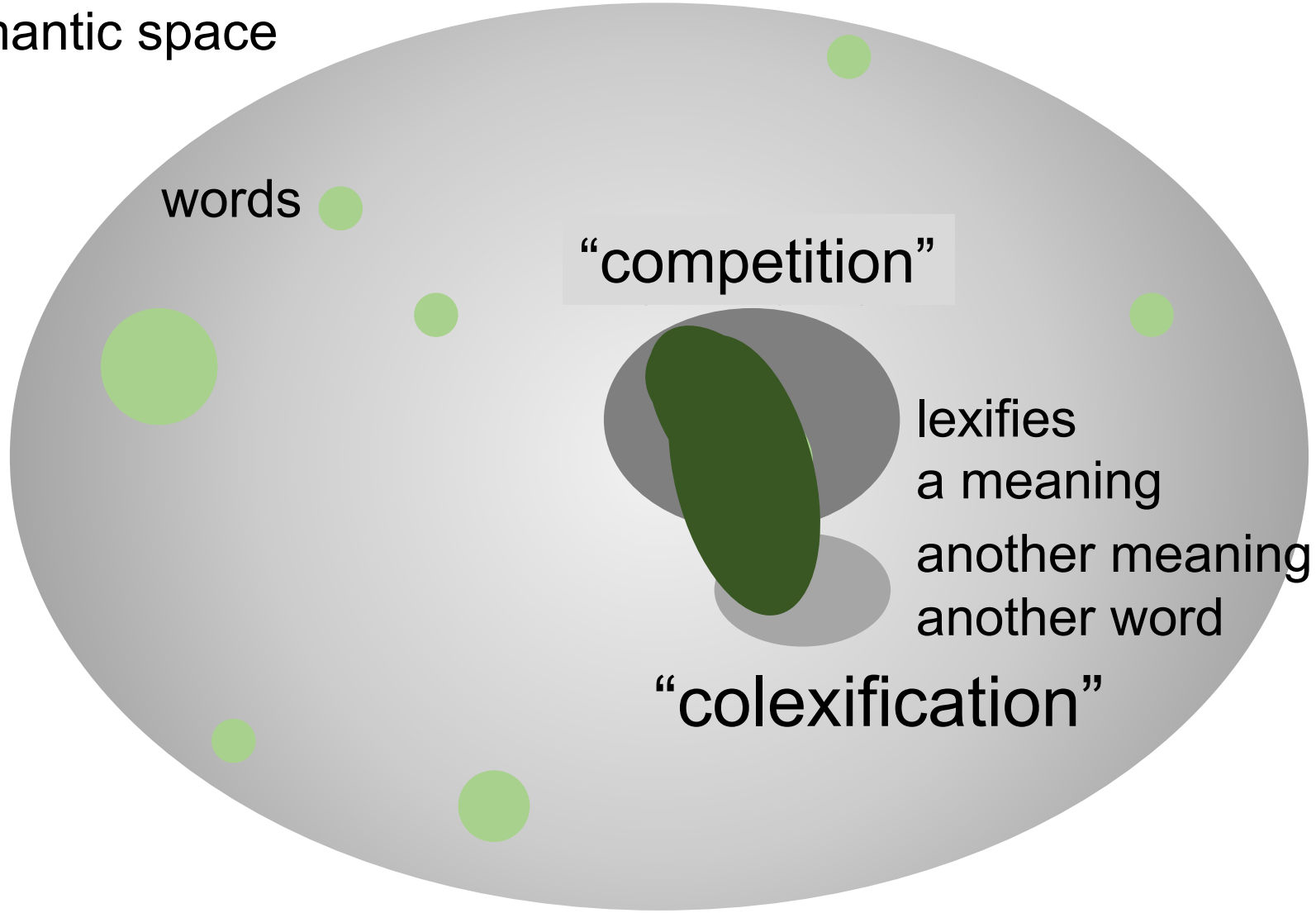
- Massive centuries-spanning corpora open up an unprecedented avenue of possible investigations into language dynamics.
- Variant usage frequencies but also meaning (and change) using distributional semantics methods

In this talk

- Communicative need and lexical competition
  - The topical-cultural advection model
- Semantic similarity and colexification - and communicative need
- Future directions: complexity and informativeness

Some concepts

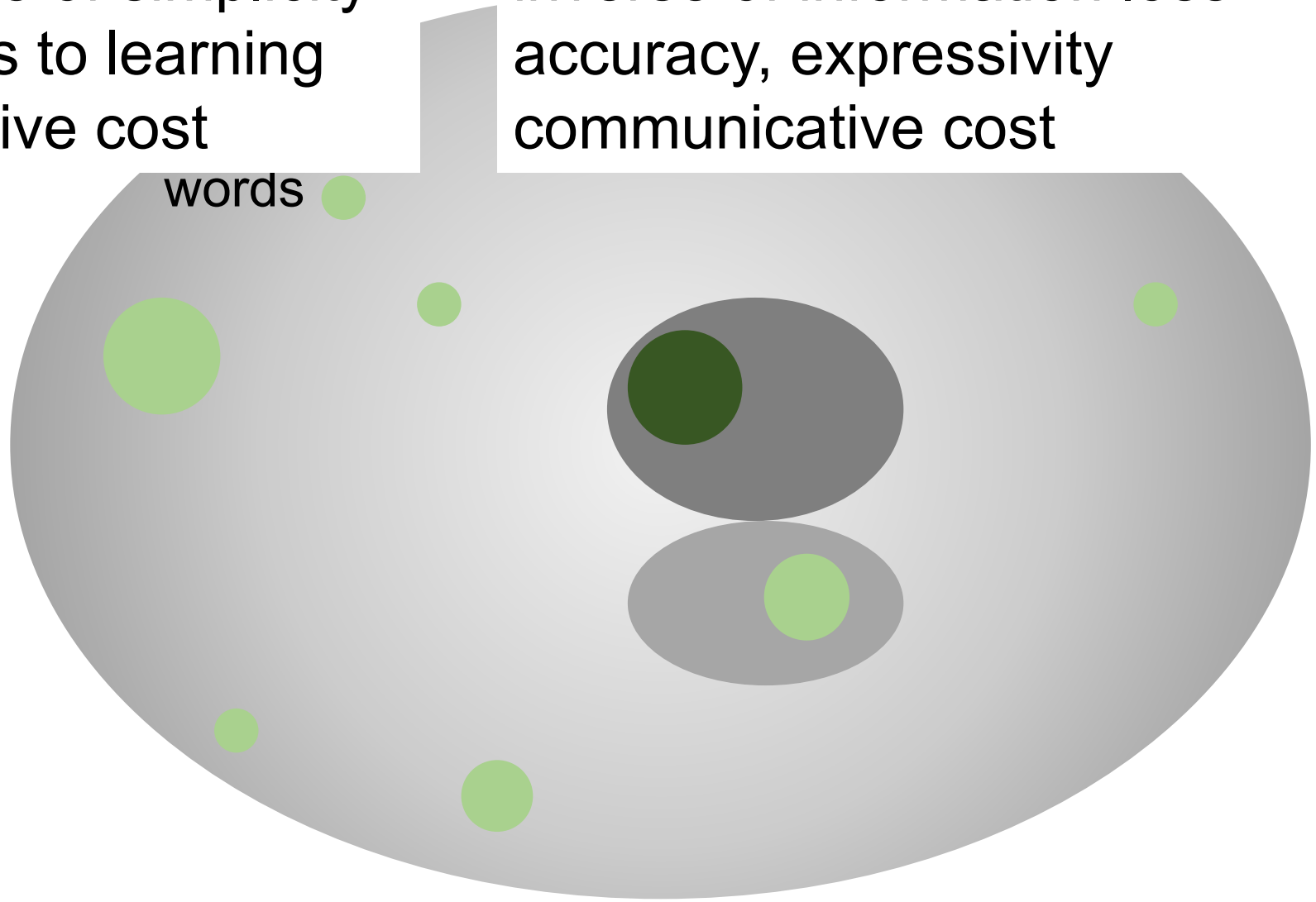
a semantic space



Complexity and informativeness

inverse of simplicity
relates to learning
cognitive cost

inverse of information loss
accuracy, expressivity
communicative cost

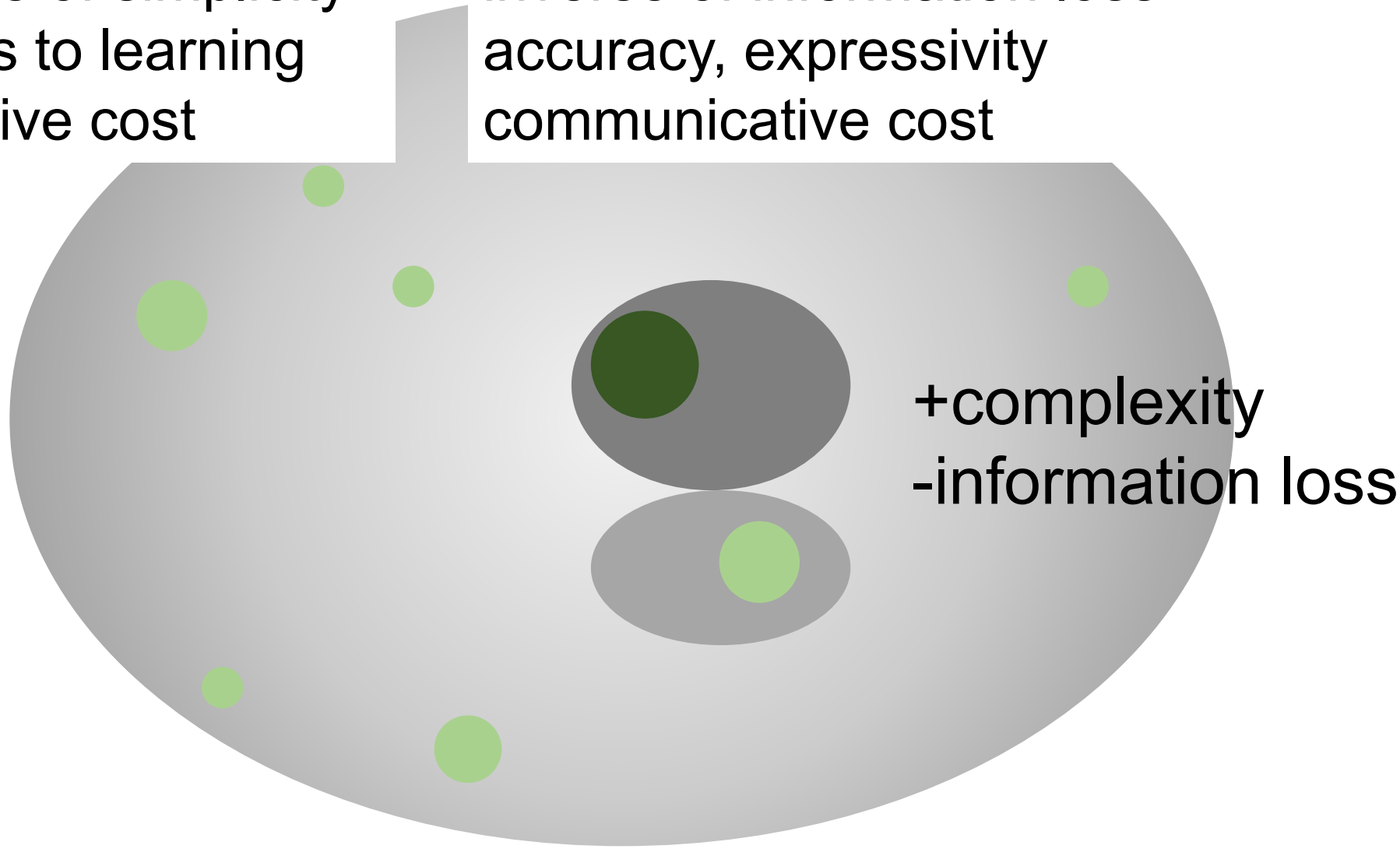


cf. Kirby et al 2015, Kemp et al 2018, Carr et al 2020

Complexity and informativeness

inverse of simplicity
relates to learning
cognitive cost

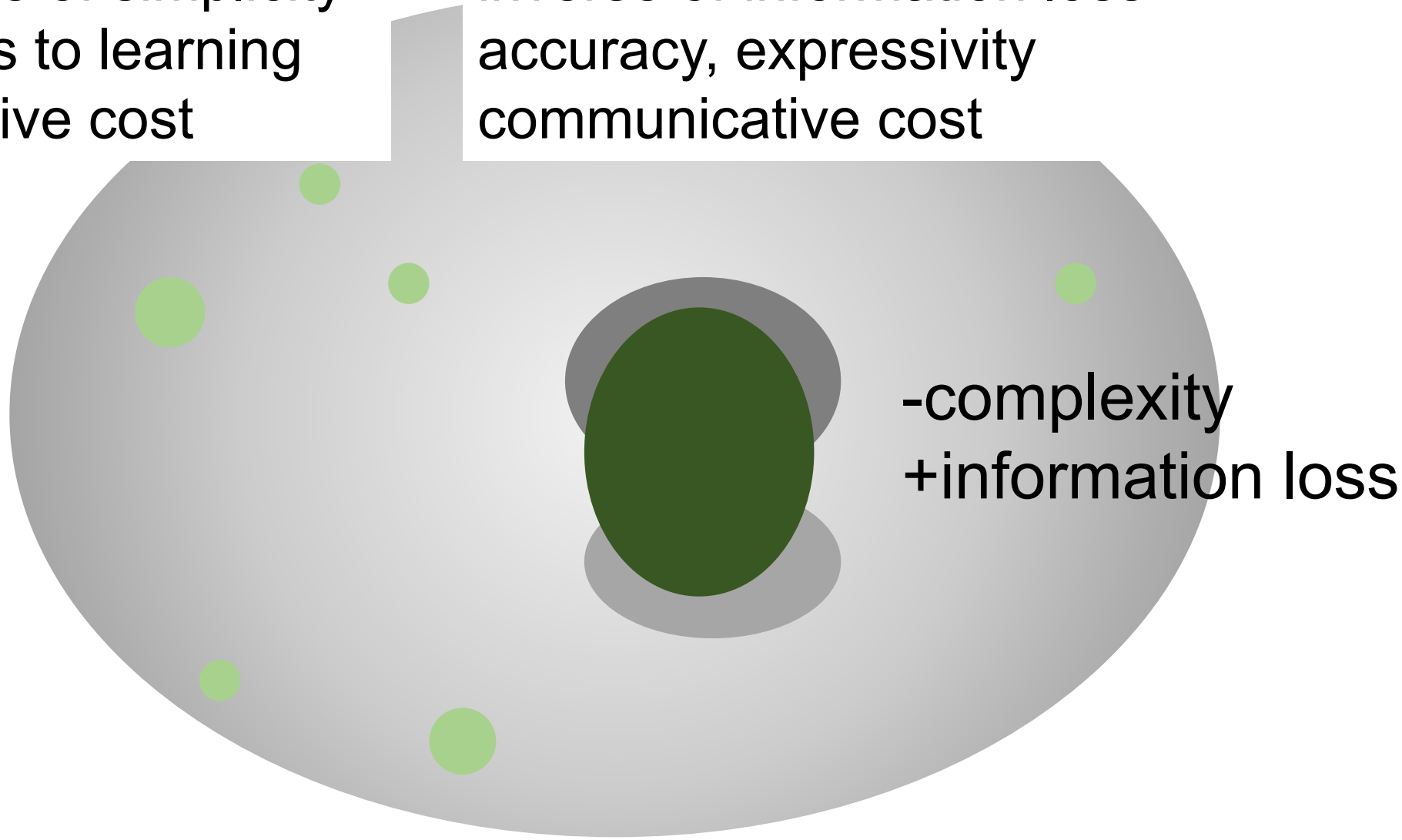
inverse of information loss
accuracy, expressivity
communicative cost



Complexity and informativeness

inverse of simplicity
relates to learning
cognitive cost

inverse of information loss
accuracy, expressivity
communicative cost

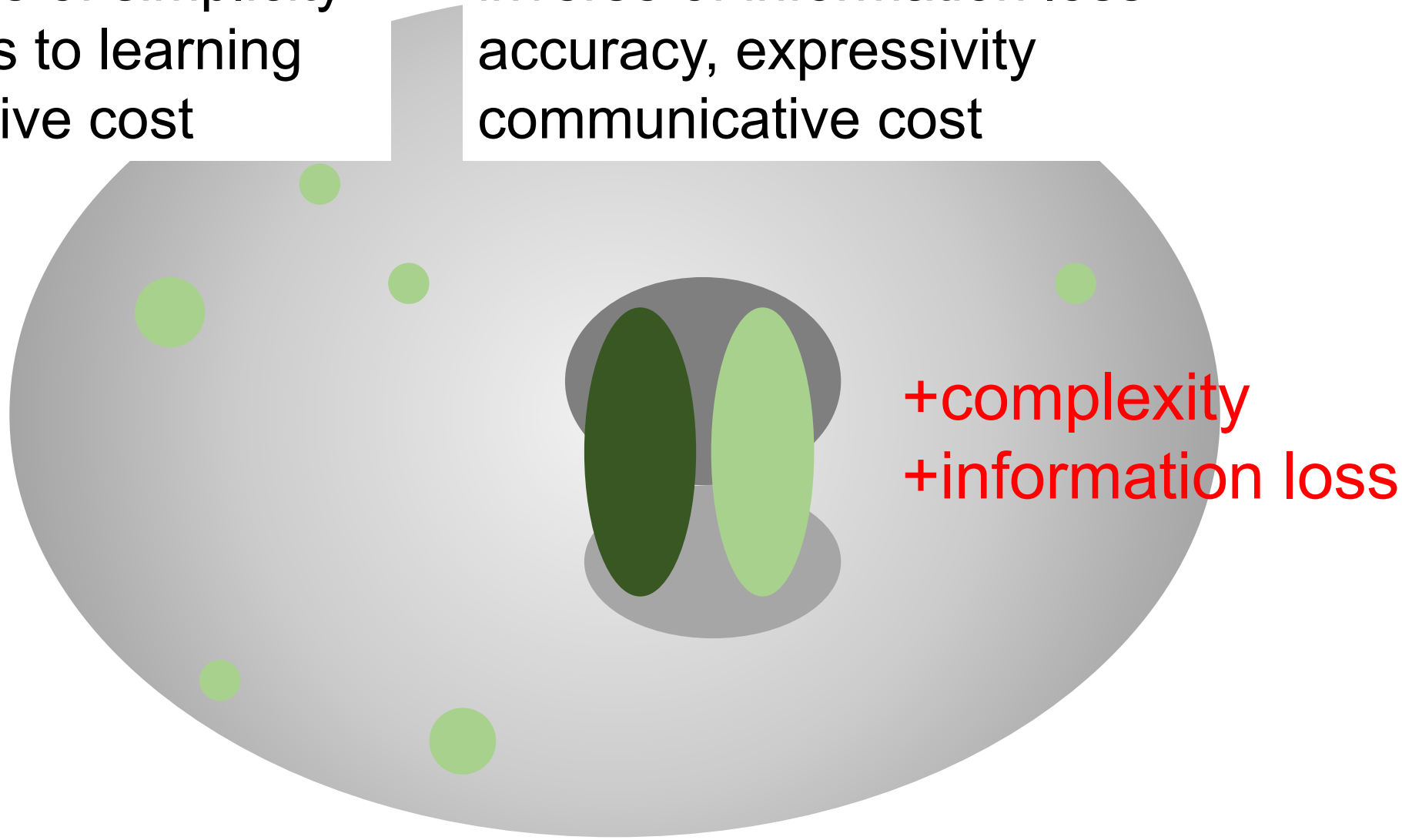


-complexity
+information loss

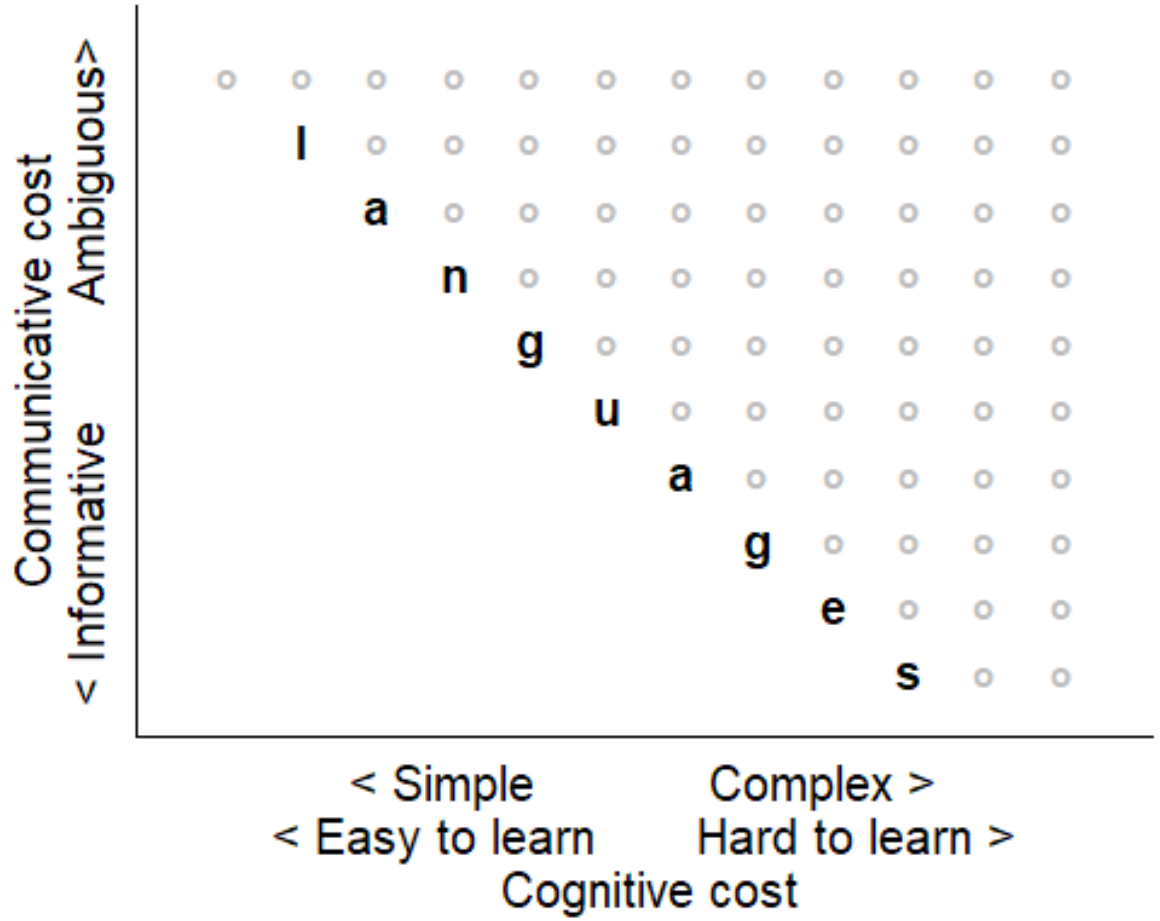
Complexity and informativeness

inverse of simplicity
relates to learning
cognitive cost

inverse of information loss
accuracy, expressivity
communicative cost

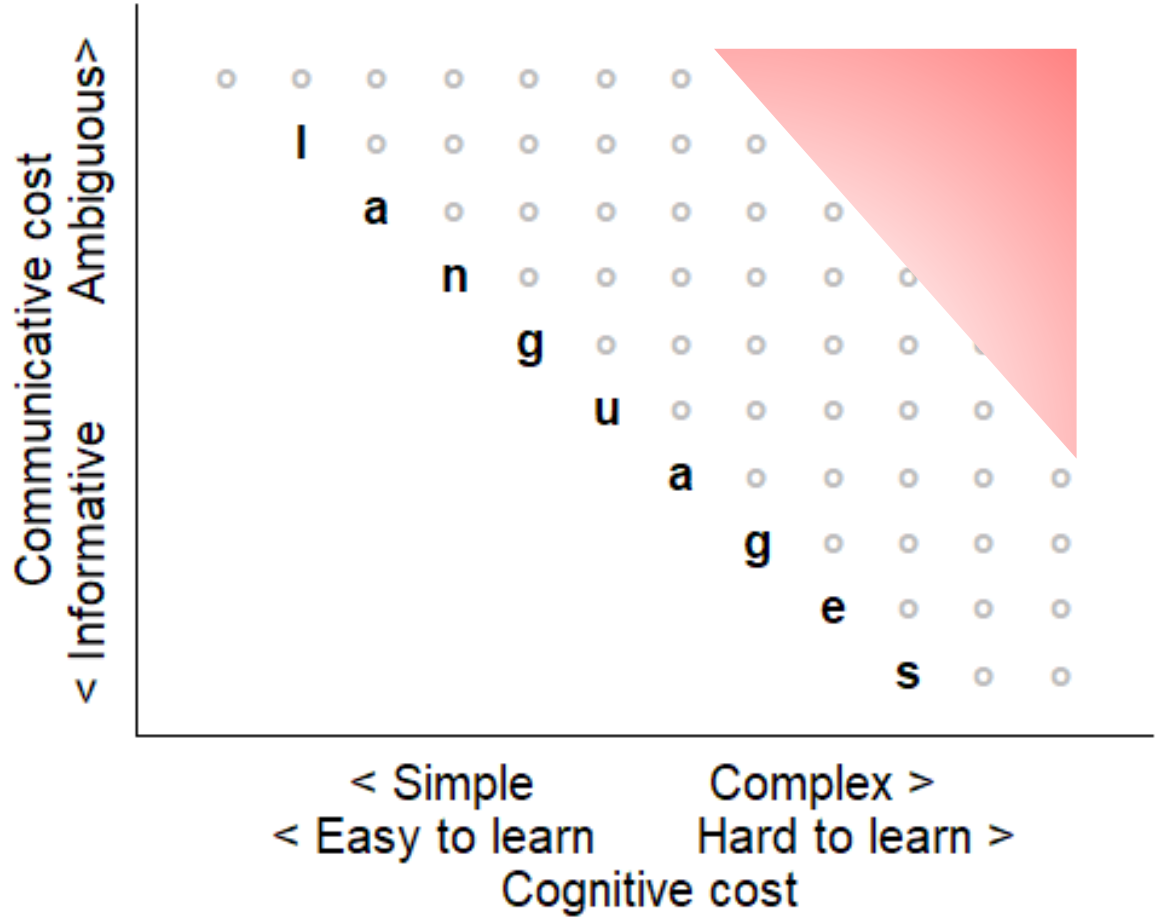


The complexity-informativeness tradeoff and the optimal front



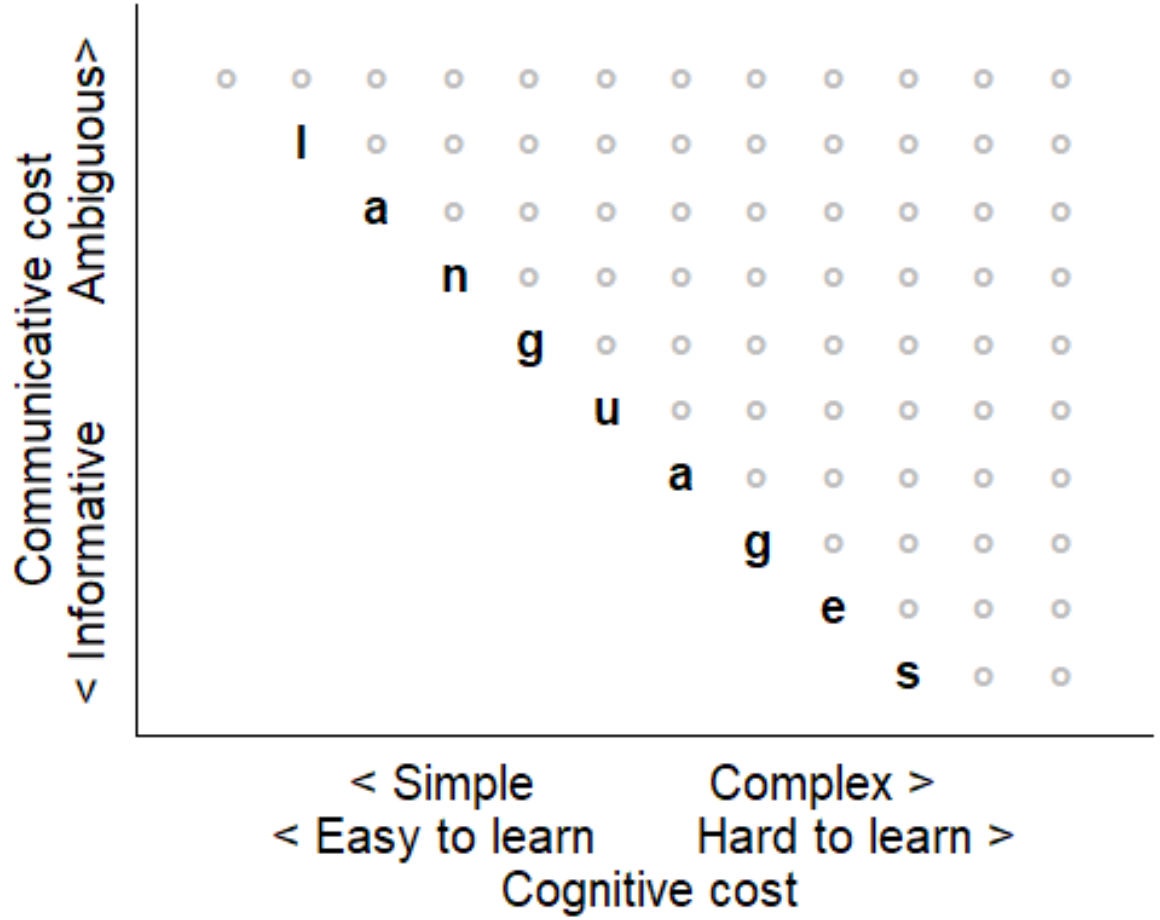
cf. Kemp et al 2012, Kemp et al 2018, Carr et al 2020

The complexity-informativeness tradeoff and the optimal front



cf. Kemp et al 2012, Kemp et al 2018, Carr et al 2020

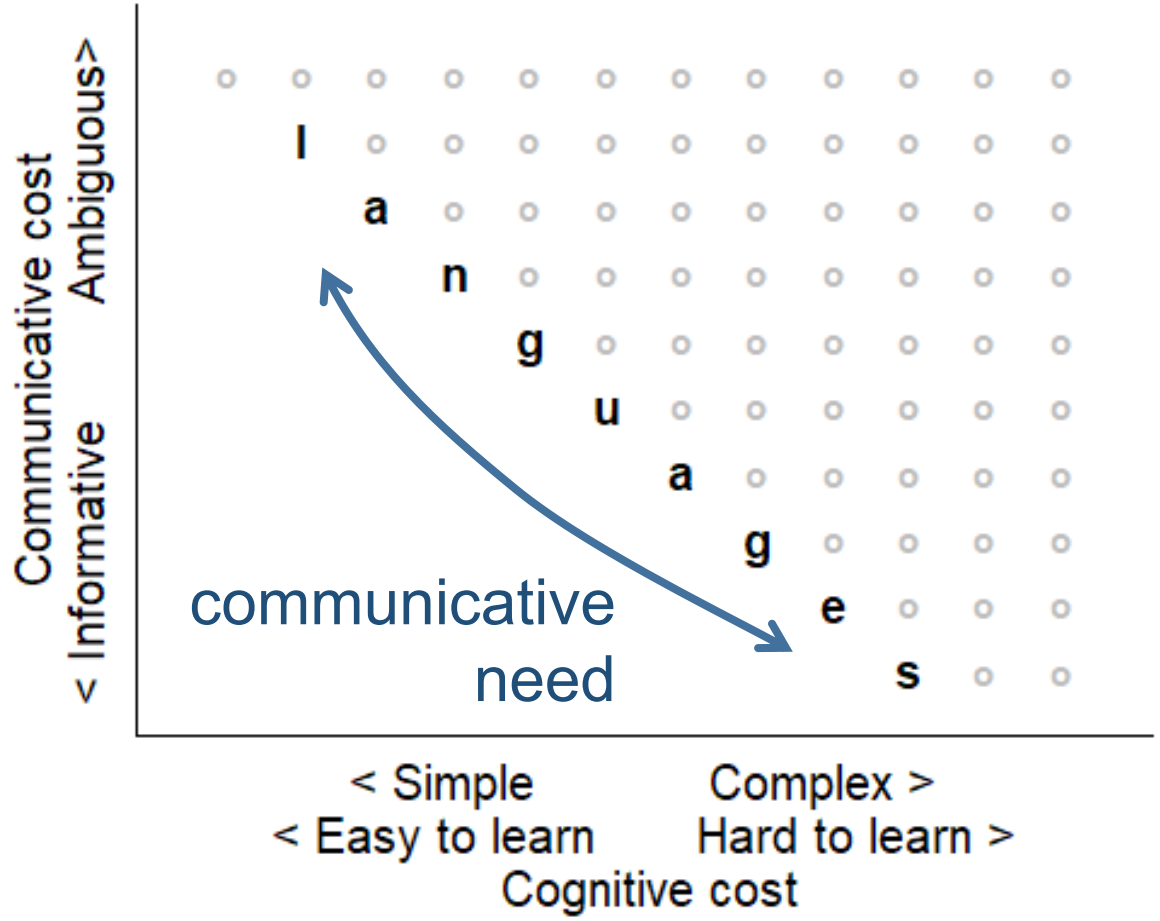
The complexity-informativeness tradeoff and the optimal front



Describes lexicons of kinship terms, colour, numeral systems, negation; similar optimization effects in artificial language experiments

cf. Kemp et al 2012, Kemp et al 2018, Carr et al 2020

The complexity-informativeness tradeoff and the optimal front



cf. Kemp et al 2012, Kemp et al 2018, Carr et al 2020

Communicative need modulates competition in language change

- Preprint: Karjus, Blythe, Kirby, Smith 2020 https://arxiv.org/abs/2006.09277
- As new words, e.g. neologisms & borrowings are selected for, what happens to their older synonyms? Does direct competition always follow local frequency changes?
- Hypothesis:
  - frequency increase in a word will lead to direct competition with (and possibly replacement of) near-synonym(s)
  - unless the lexical subspace experiences high communicative need

Communicative need modulates competition in language change

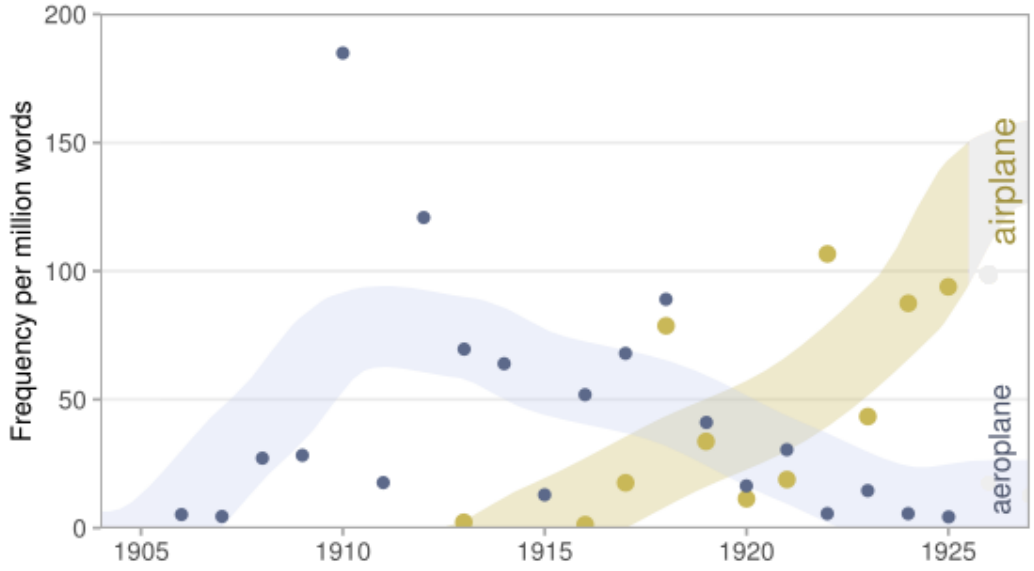
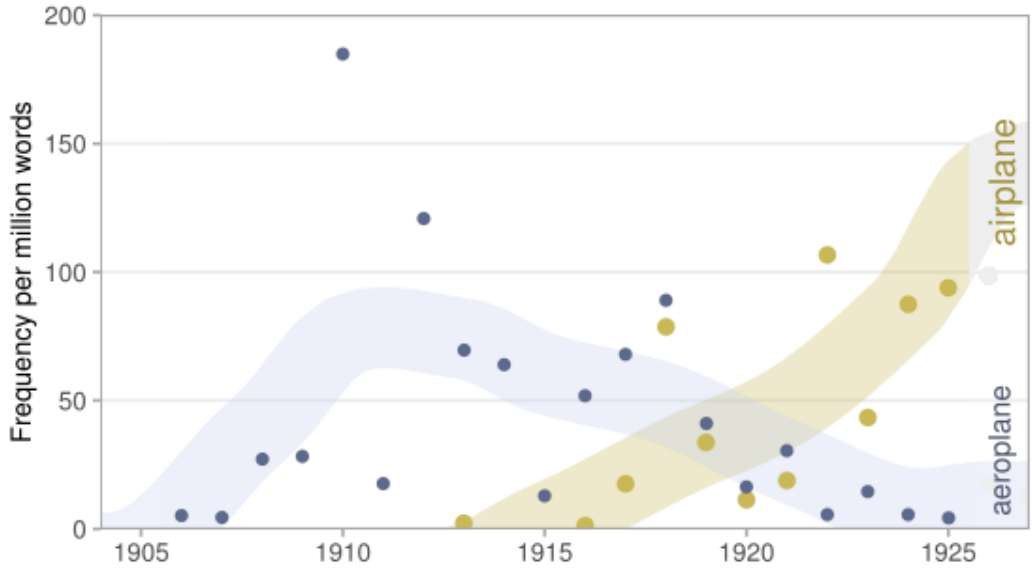
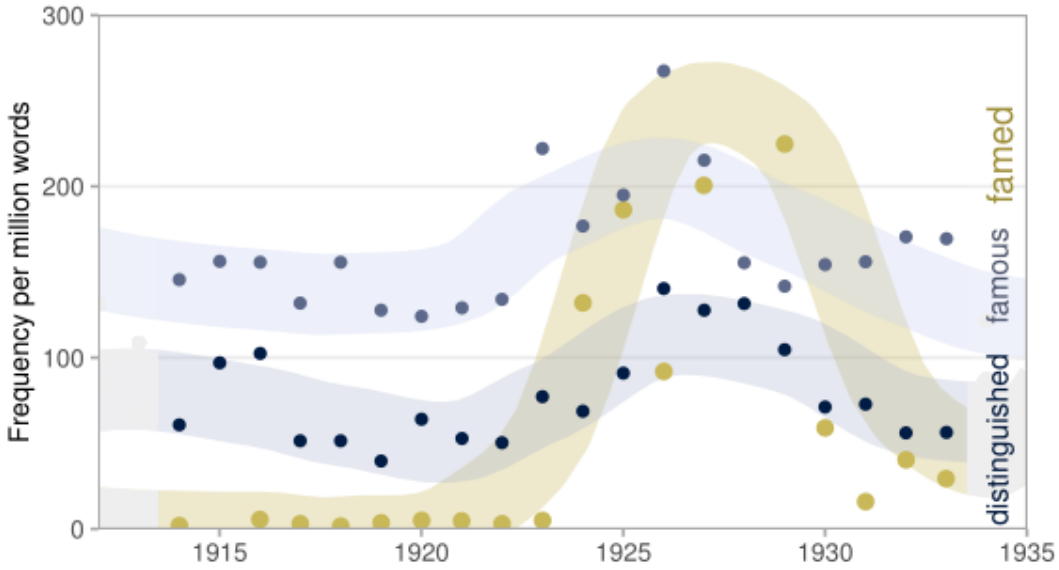


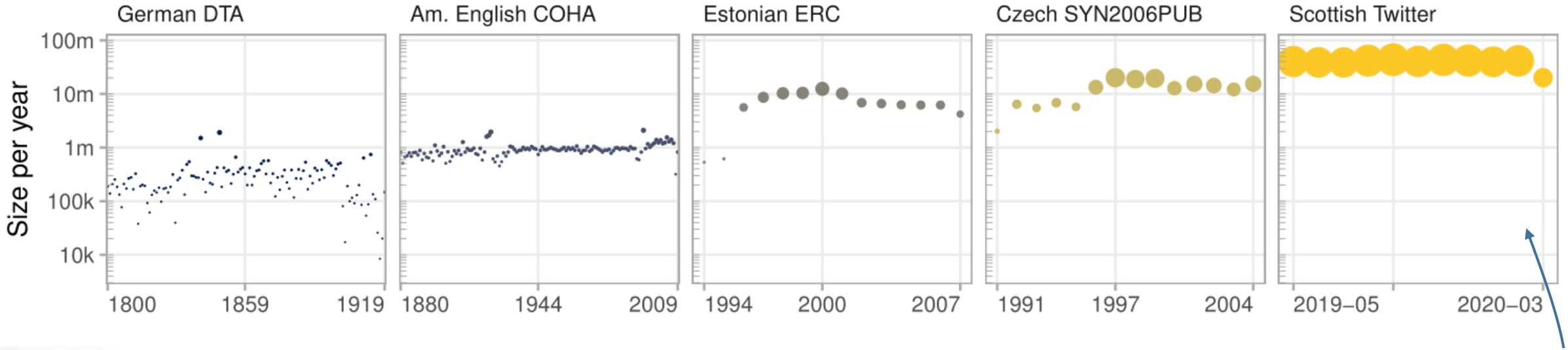
Figure 1: Example time series from the Corpus of Historical American English (COHA). Two

Communicative need modulates competition in language change





The corpora



- COHA&DTA: 10-year bins (5 for ERC, Czech, month for Twitter)
- Targets: min +2 log change, occurs min 100x & in
- A model of communicative need

- Need:
- A model of competition
- A model of communicative need

A model of communicative need

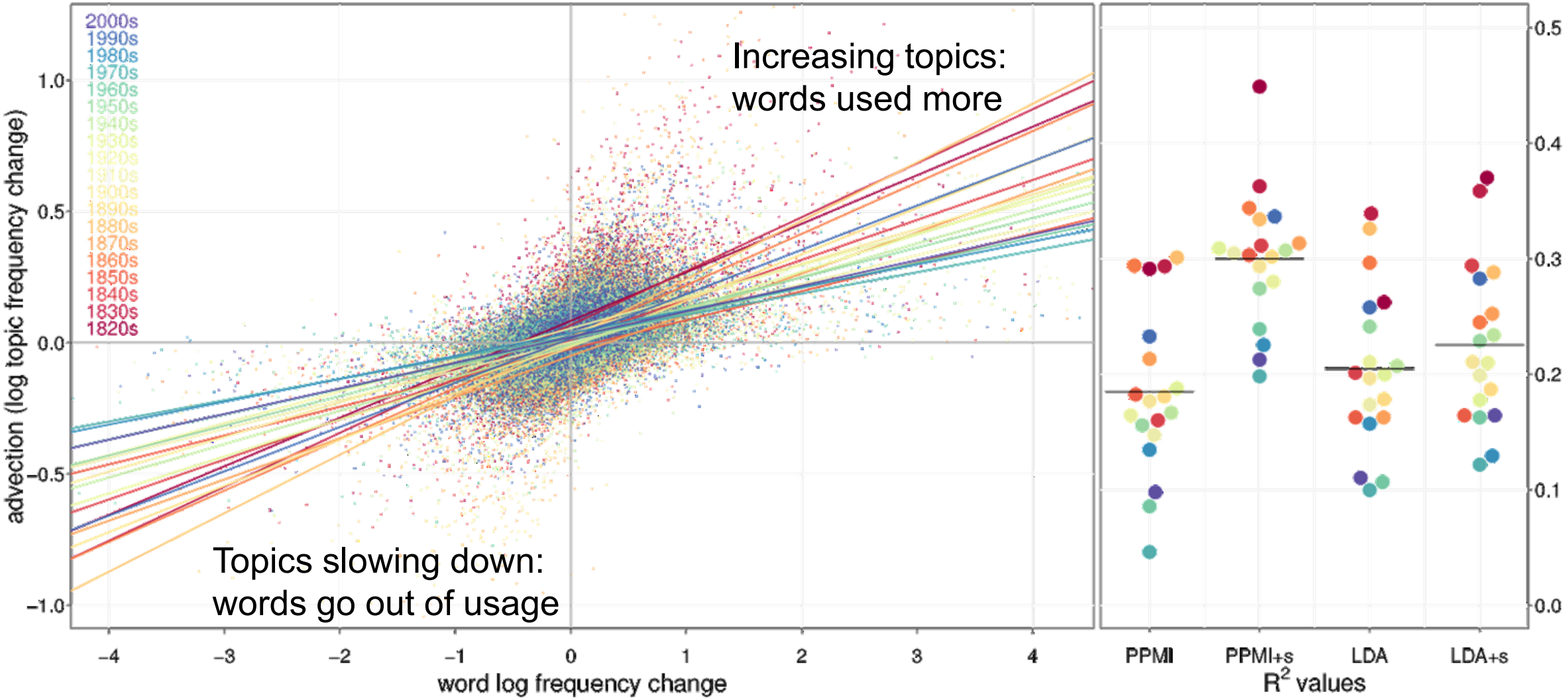
- Karjus, Blythe, Kirby, Smith 2020, Quantifying the dynamics of topical fluctuations in language. Language Dynamics and Change
- Idea: see how much the topic of a target word changes (weighted mean of the log frequency changes of the relevant topic (context) words of the target)
- Discourse topic prevalence ~ how much something needs to be talked about ~ communicative need
- Topics as the latent flow of language, dragging words along

- *advection* - the transfer of matter (or heat) by the flow of a fluid

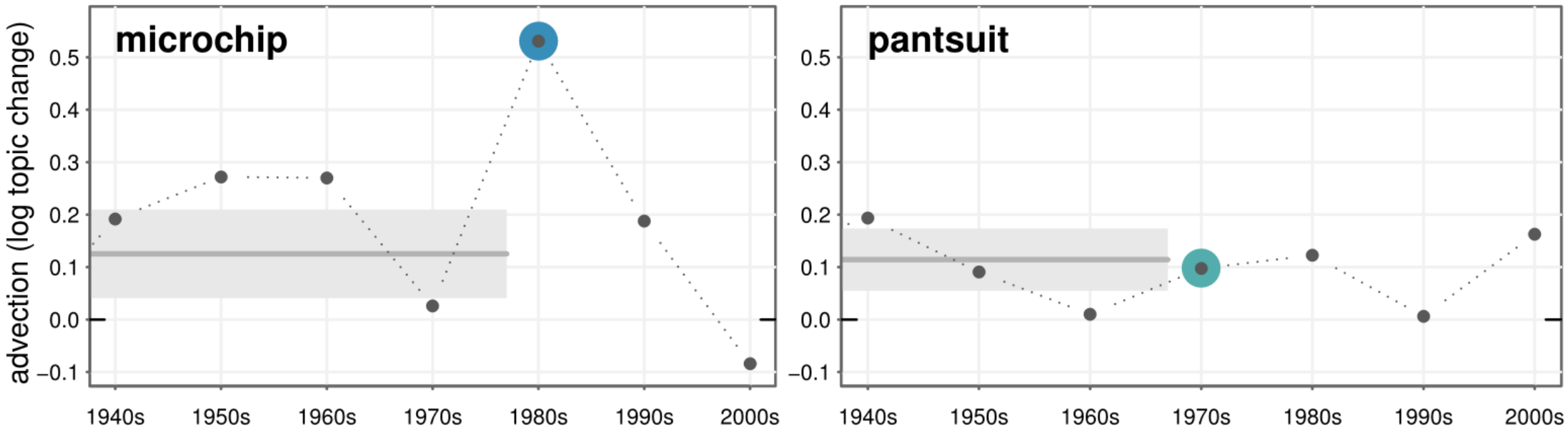
A model of communicative need

- *advection* - the transfer of matter (or heat) by the flow of a fluid

Quantifying the dynamics of topical fluctuations in language



Advection a proxy to communicative need



top topic words

transistor, silicon, implant, fiberoptic, infringement, circuitry, semiconductor, converter, workstation, thickest, computing, smokestack, rile, fabrication, subversion, synthesizer, biotechnology, optics, energize, revolutionize, cerebral, ...

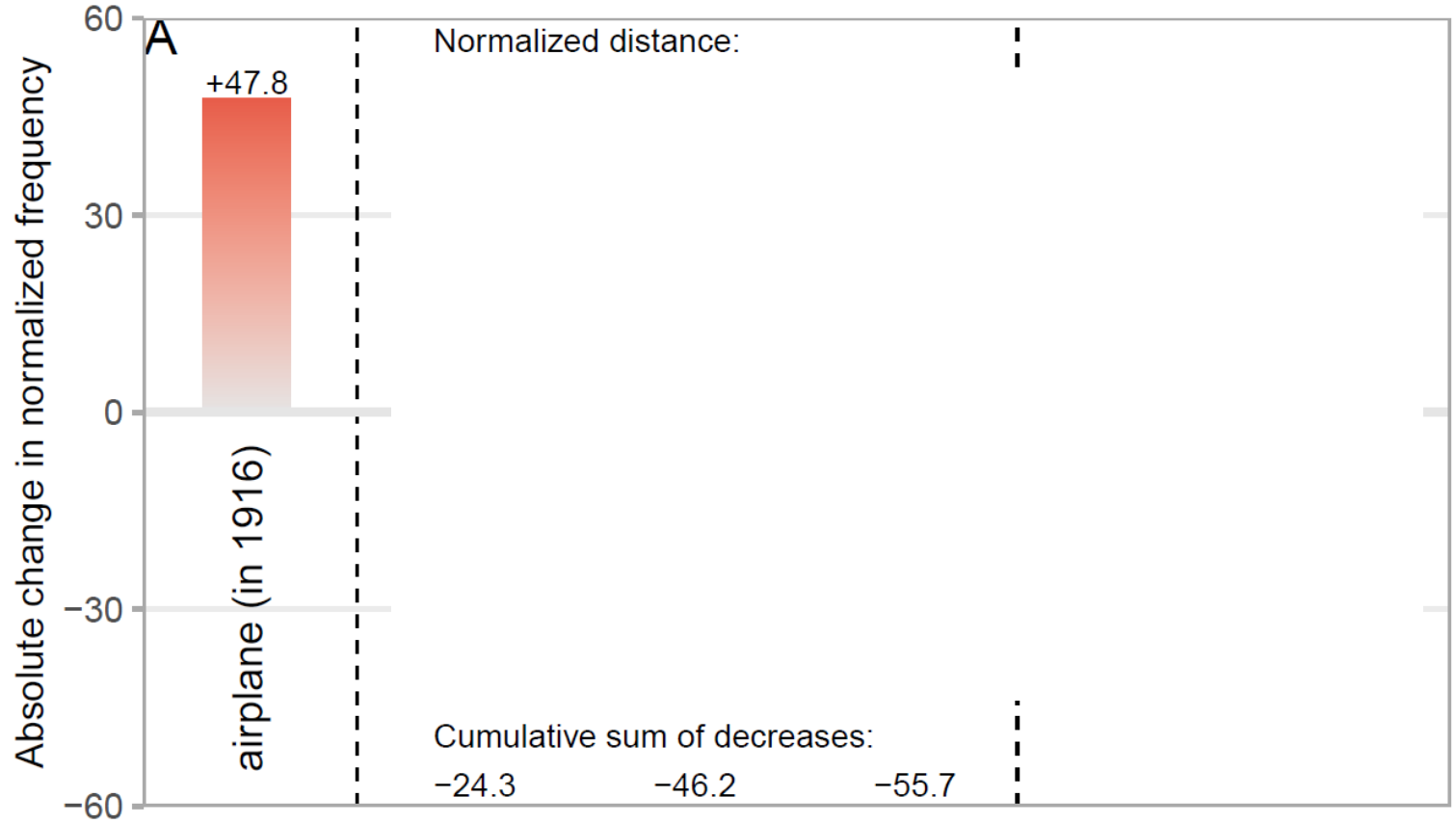
highheeled, polyester, tailored, longsleeved, pajama, creamcolored, deadpan, beige, unimpressed, violet, quip, corduroy, incongruous, sportsman, devoted, safari, maternity, stylish, streamlined, bustle, joker, tuxedo, lustrous, petite, frosted, portly, navy, ...

A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine distance from target where the sum of (normalized) frequency decreases match the increase of the target
- Normalized corpus frequencies sum to 1
- Increase somewhere => decrease somewhere else
- A realistic model of language? Yes: time is finite and learning pressure biases for simpler lexicons. Can’t have infinitely many words.
- Semantics: inferred from LSA, trained for each target word based on (ppmi-weighted) co-occurrence matrix of the preceding time bin, fit target vector into this model – yields neighbours of the *position* where the new word will appear in

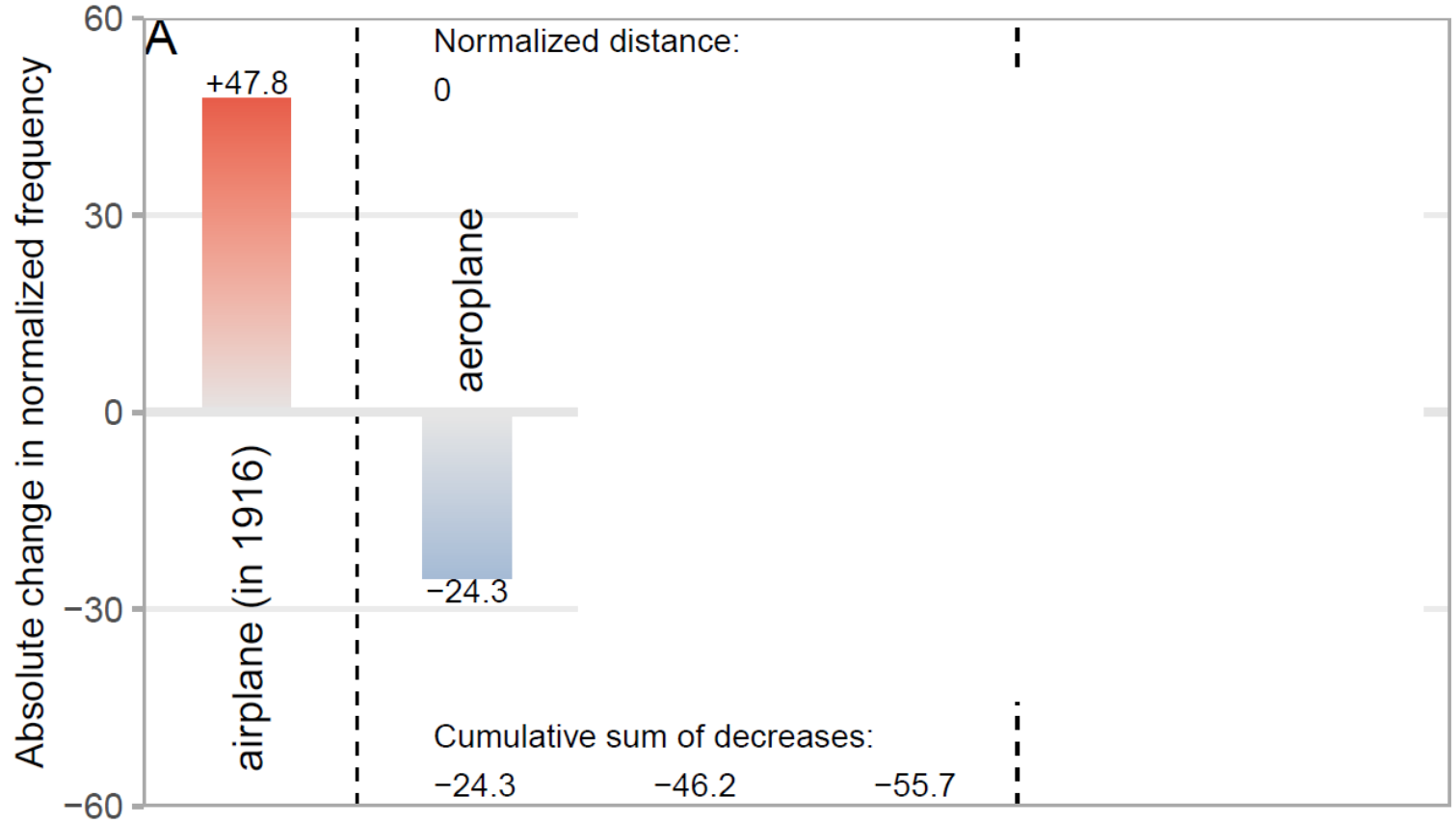
A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine distance from target where the sum of (normalized) frequency decreases match the increase of the target



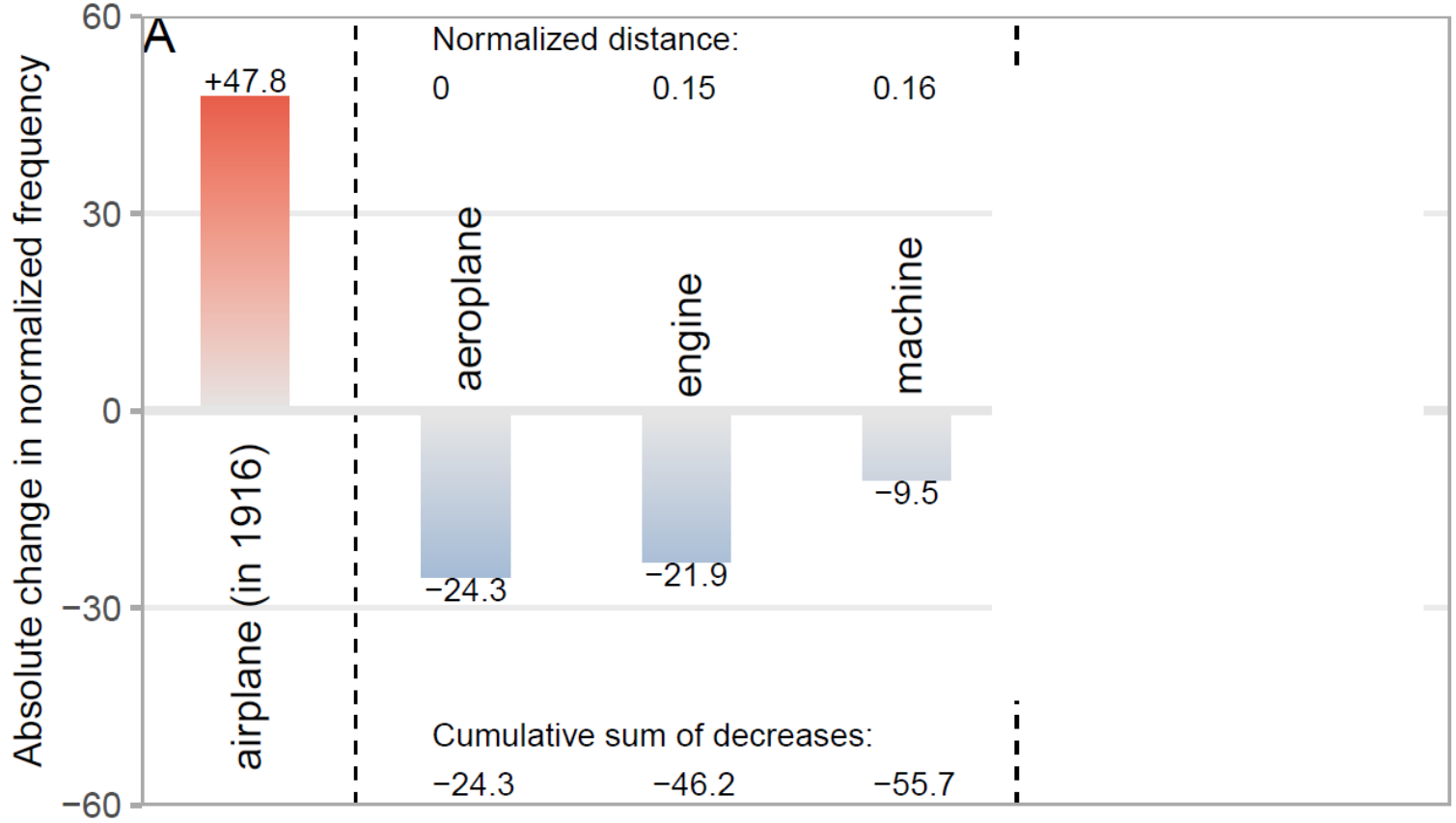
A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine **distance from target** where the sum of (normalized) frequency **decreases** match the increase of the target



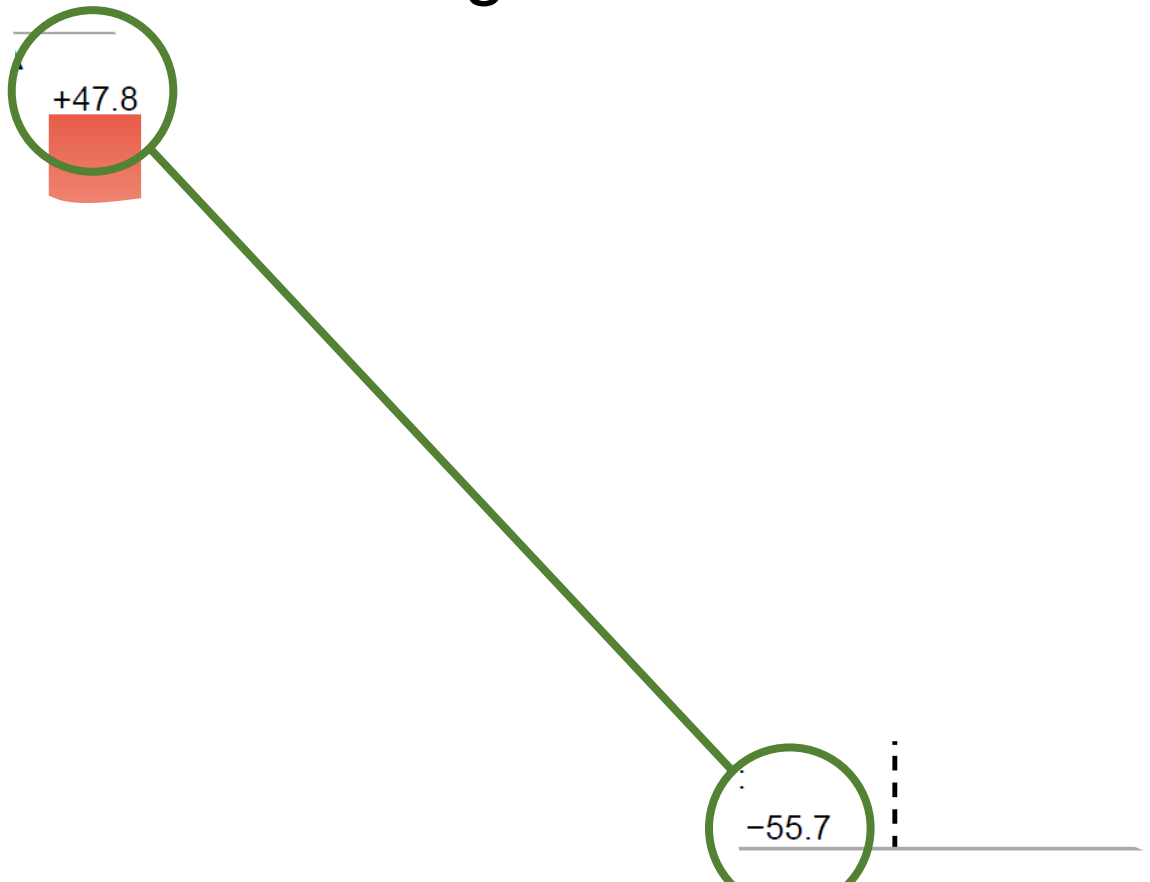
A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine **distance from target** where the sum of (normalized) frequency **decreases match the increase** of the target



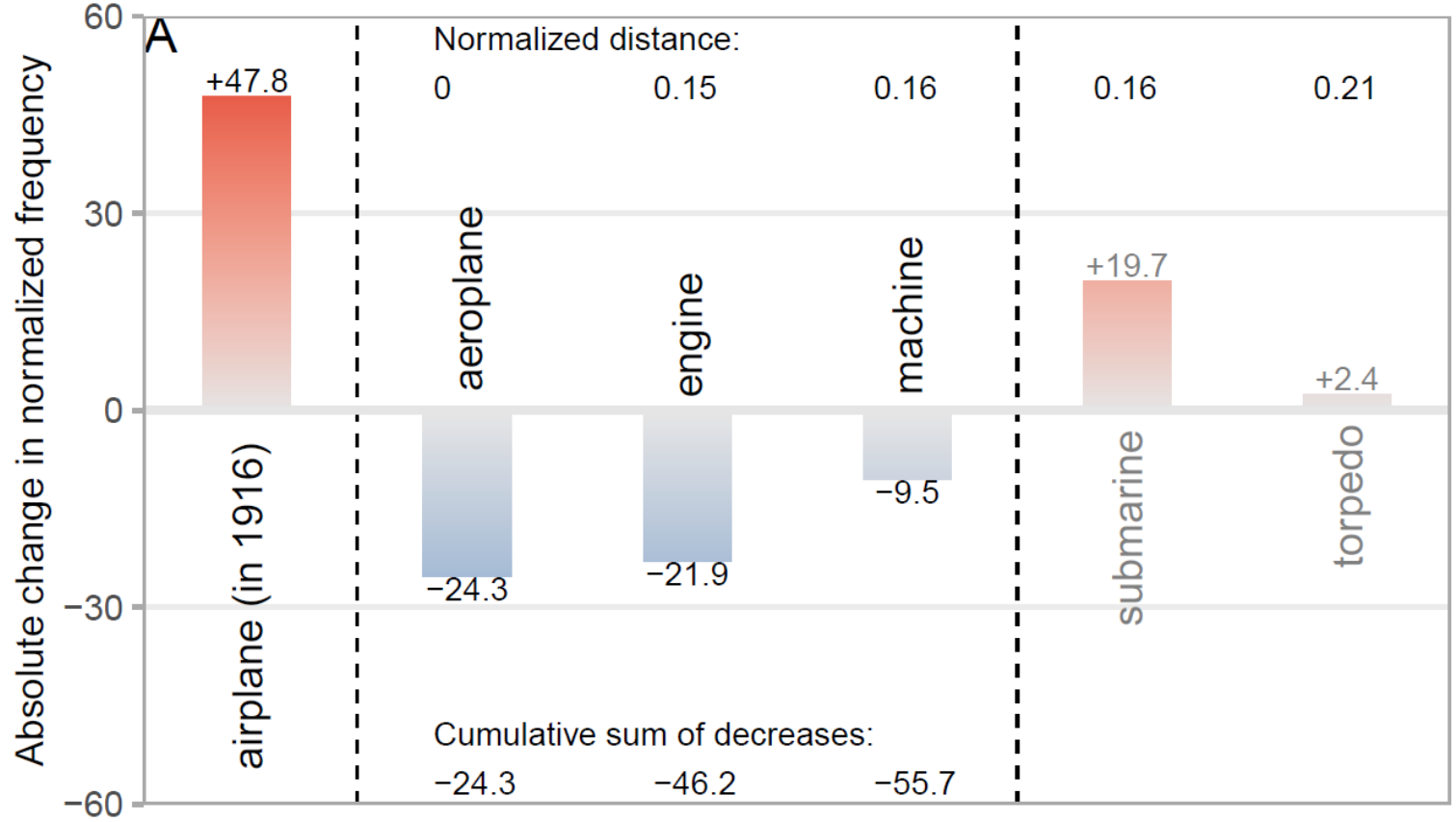
A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine **distance from target** where the sum of (normalized) frequency **decreases match the increase** of the target

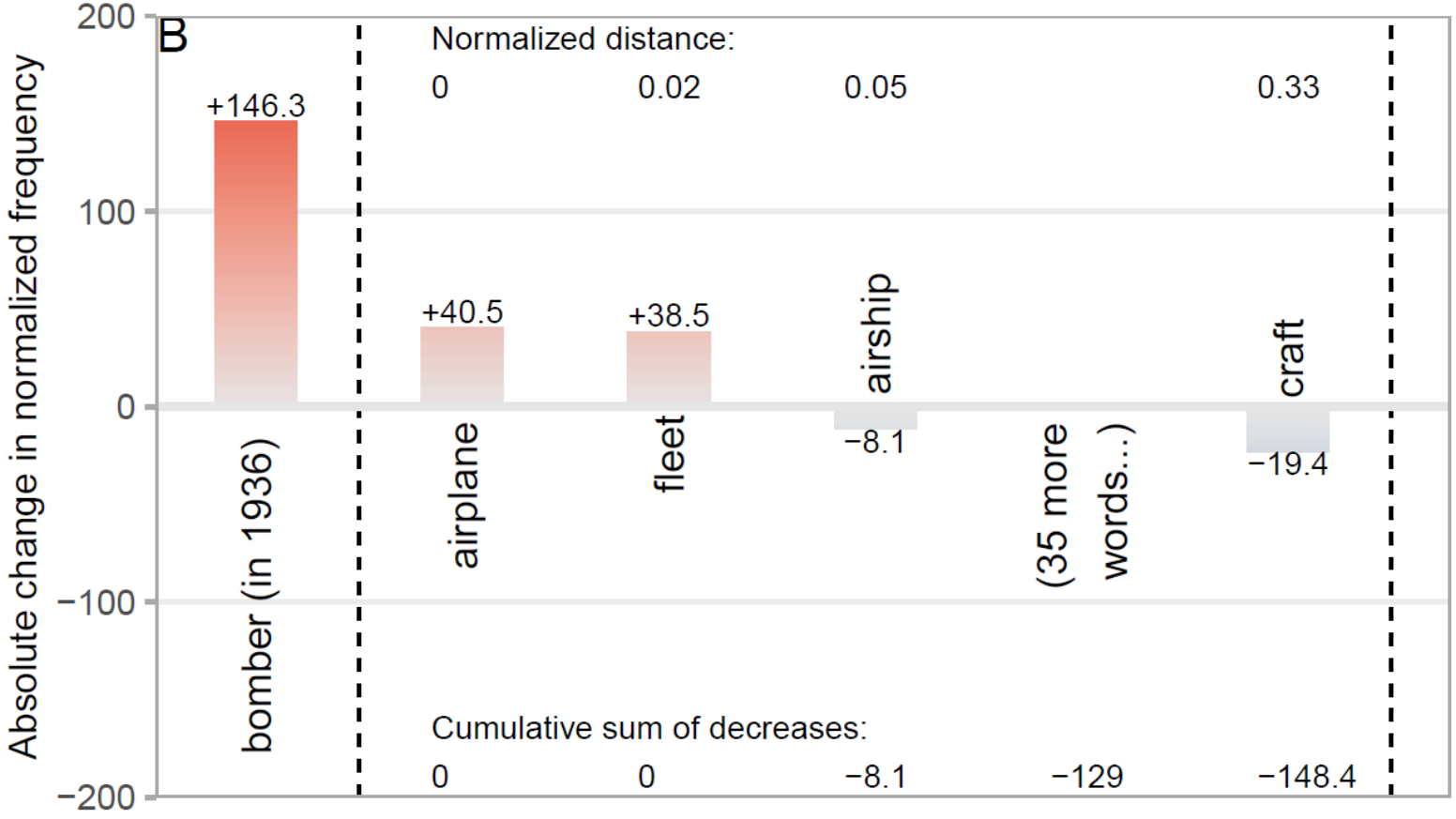


A model of linguistic competition

- Meaning from word embeddings; **equalization range**: norm. cosine **distance from target** where the sum of (normalized) frequency **decreases match the increase** of the target



A model of linguistic competition



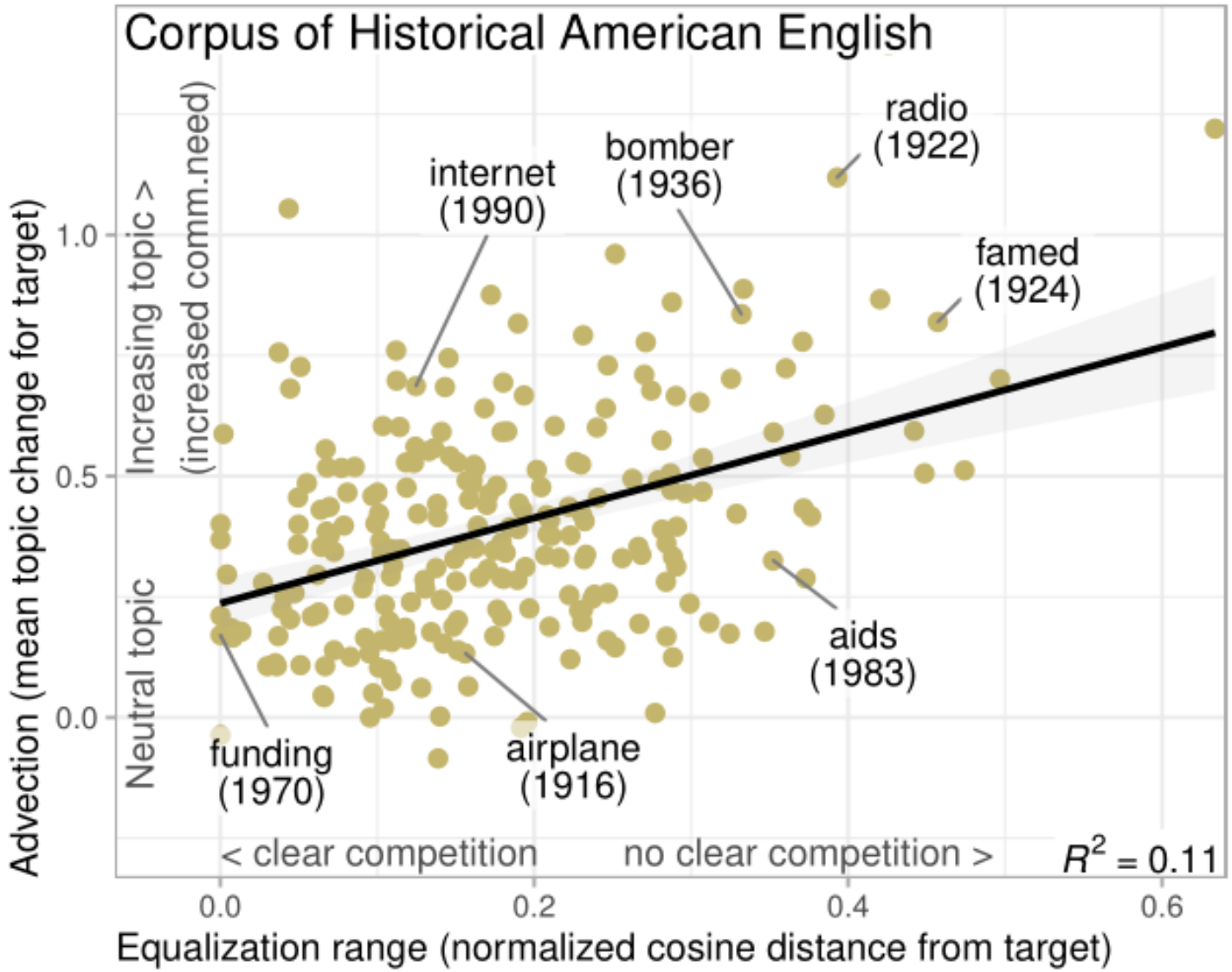
Important

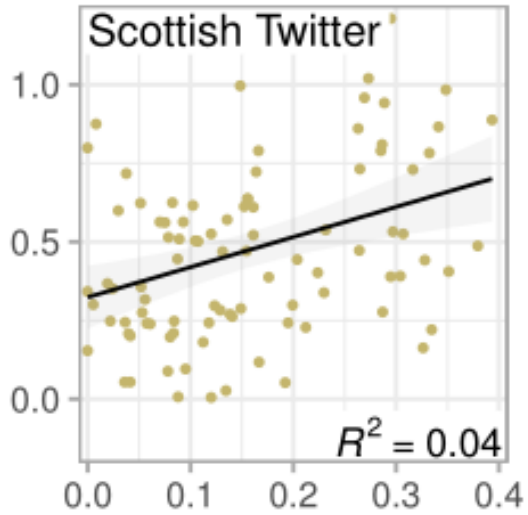
- Both models based on lists of words, but decorrelated:
  - advection: weighted list of associated, co-occurring words (1st order similarity)
  - competition: list of all words, ordered by embedding cosine similarity (2nd order similarity), minus any words in the advection list for a given target

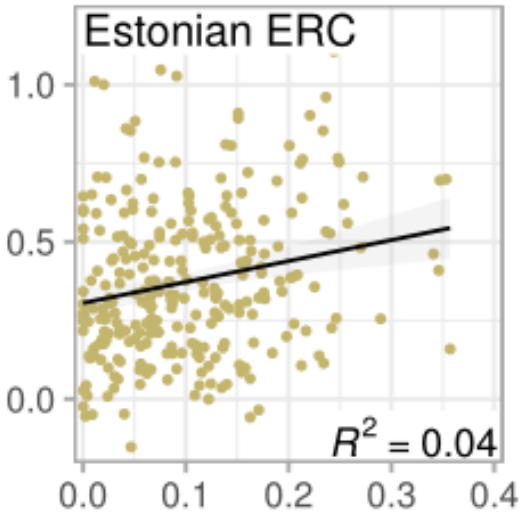
- Necessary, but can weaken the competition model accuracy, if closest neighbours (~synonyms) also co-occur with target:
- *airplane* | *aeroplane* *airship aerial propeller balloon* *engine machine submarine* *biplane wireless* *torpedo*

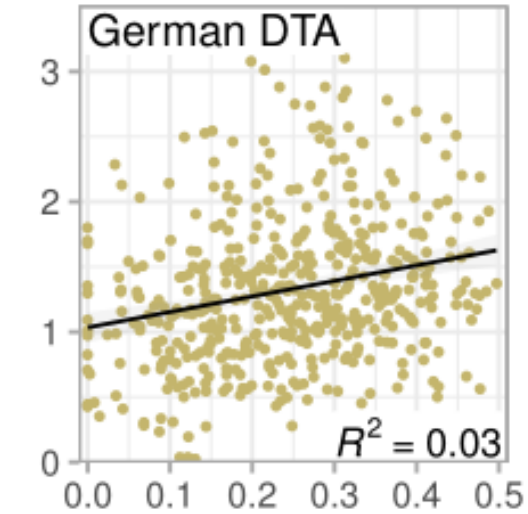
Results

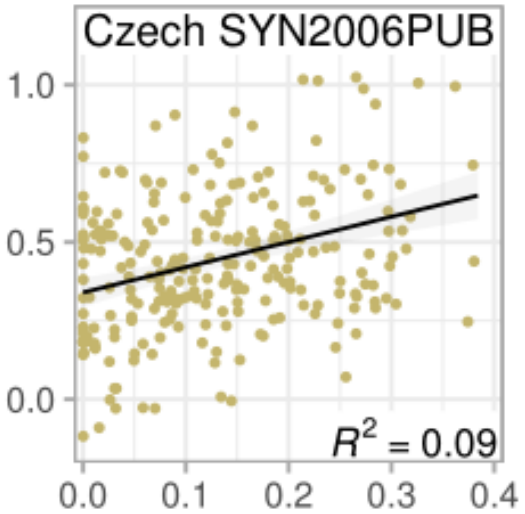
- Topical advection (proxy to communicative need) correlates with
- Equalization range (proxy to extent of competition)

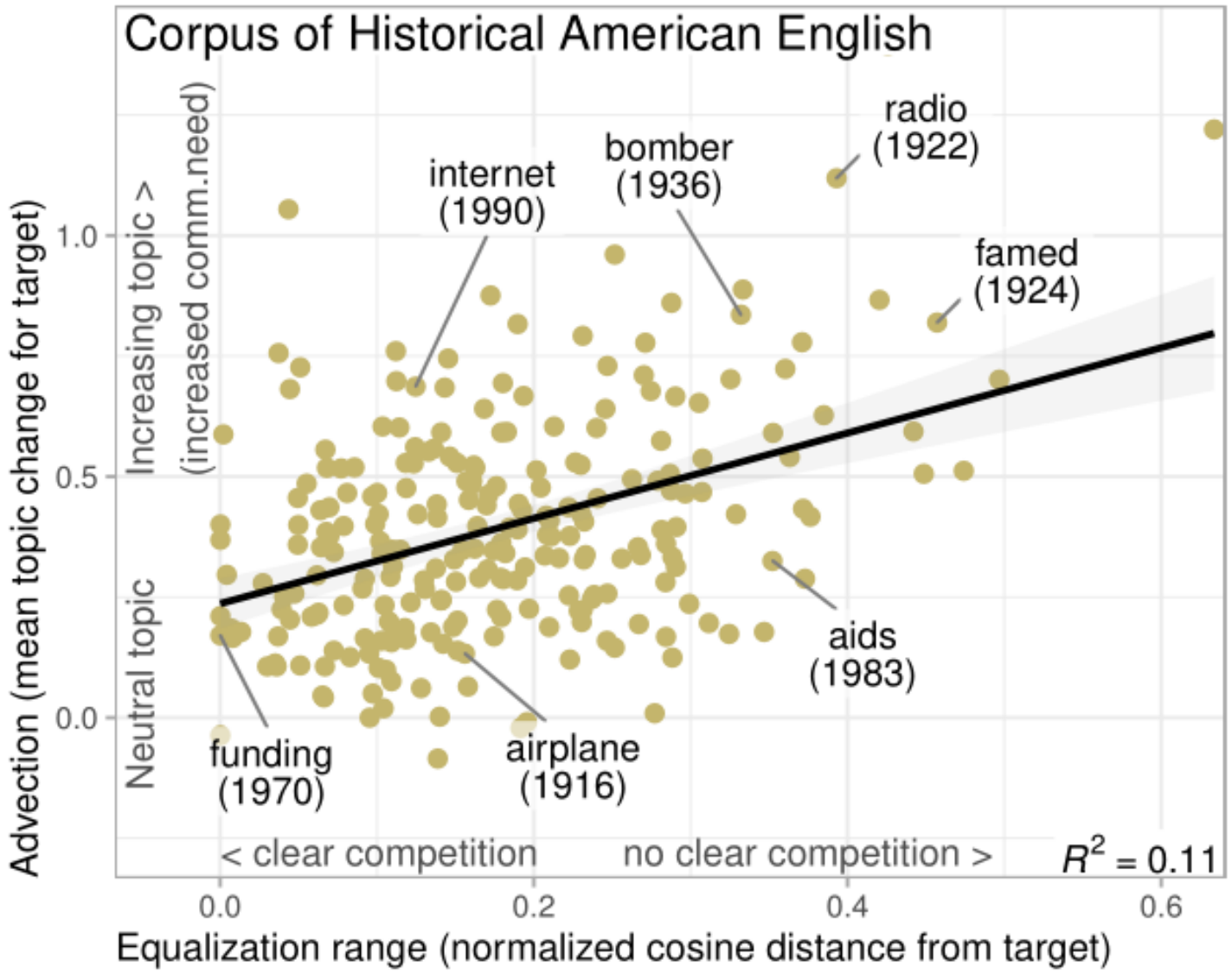

Corpus of Historical American English
Advection (mean topic change for target)
Equalization range (normalized cosine distance from target)
Increasing topic > (increased comm.need)
Neutral topic
< clear competition
no clear competition >
R^2 = 0.11
radio (1922)
bomber (1936)
internet (1990)
famed (1924)
aids (1983)
airplane (1916)
funding (1970)

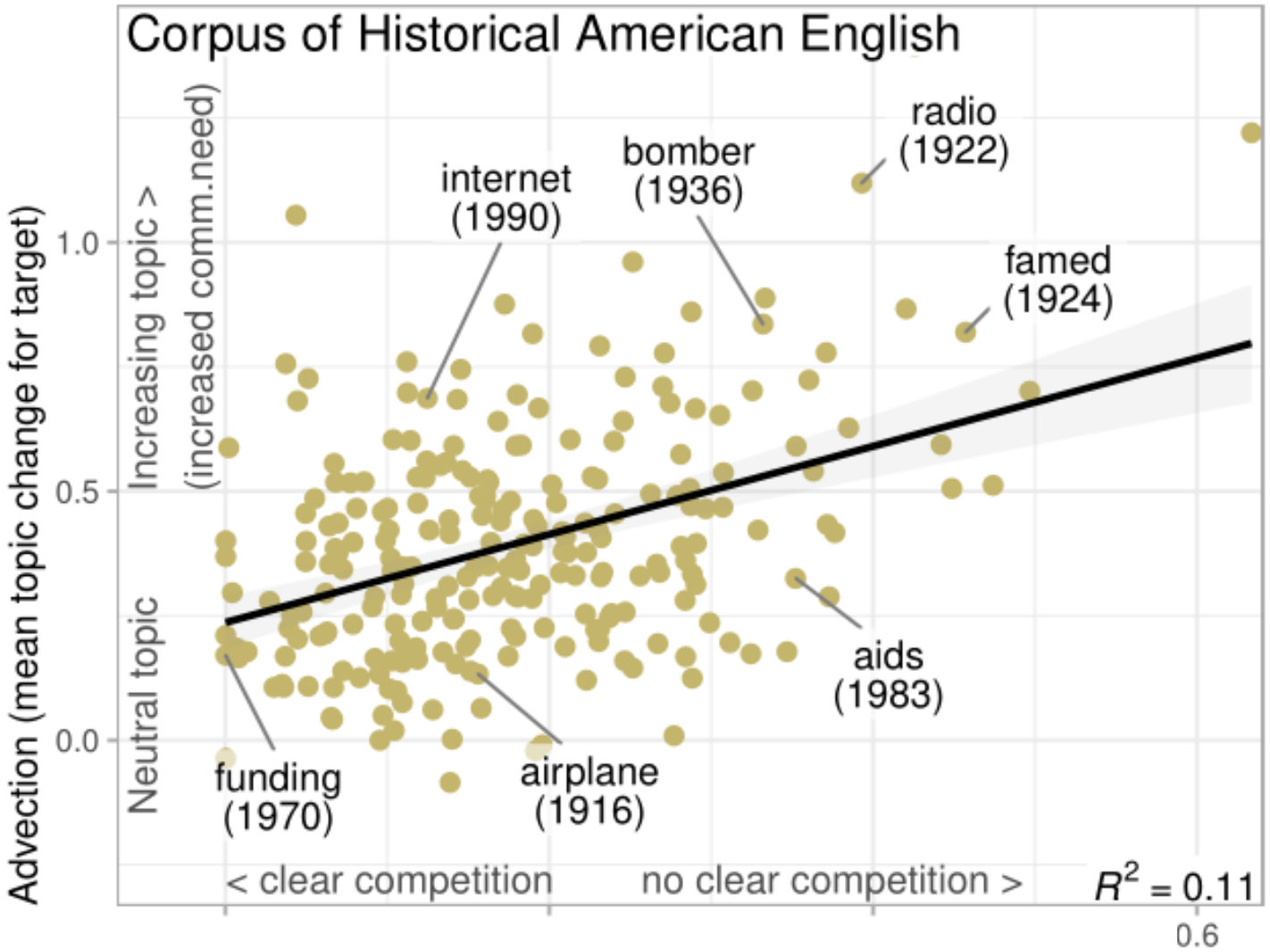

Scottish Twitter
R^2 = 0.04


Estonian ERC
R^2 = 0.04

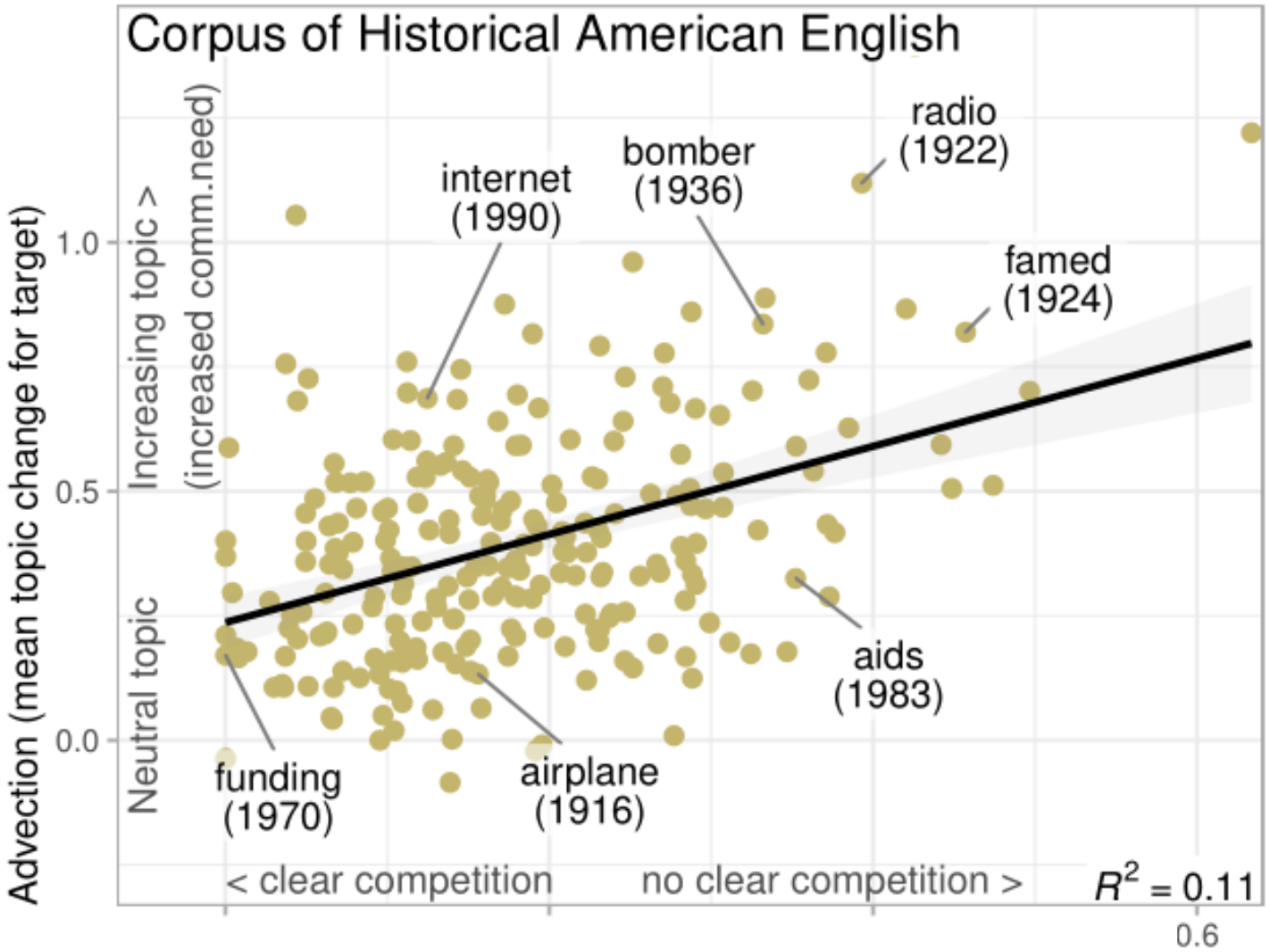

German DTA
R^2 = 0.03


Czech SYN2006PUB
R^2 = 0.09


Corpus of Historical American English
Increasing topic >
(increased comm.need)
Neutral topic
internet
(1990)
bomber
(1936)
radio
(1922)
famed
(1924)
aids
(1983)
funding
(1970)
airplane
(1916)
< clear competition
no clear competition >
R^2 = 0.11
Advection (mean topic change for target)
1.0
0.5
0.0
0.0
0.2
0.4
0.6
Equalization range (normalized cosine distance from target)



- Lower communicative need: competition more likely



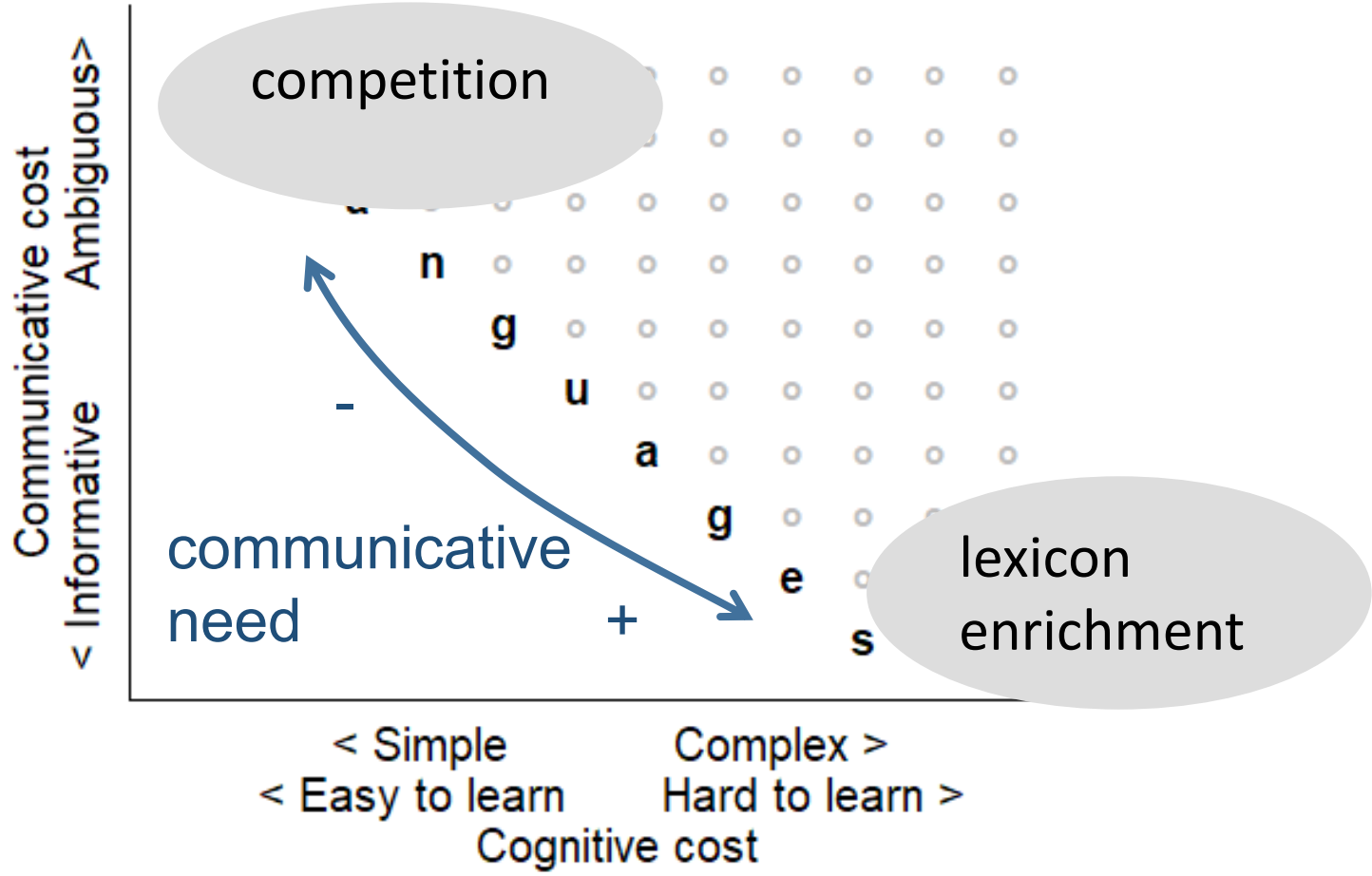
- High communicative need: similar words more likely to coexist

- Lower communicative need: competition more likely

Discussion

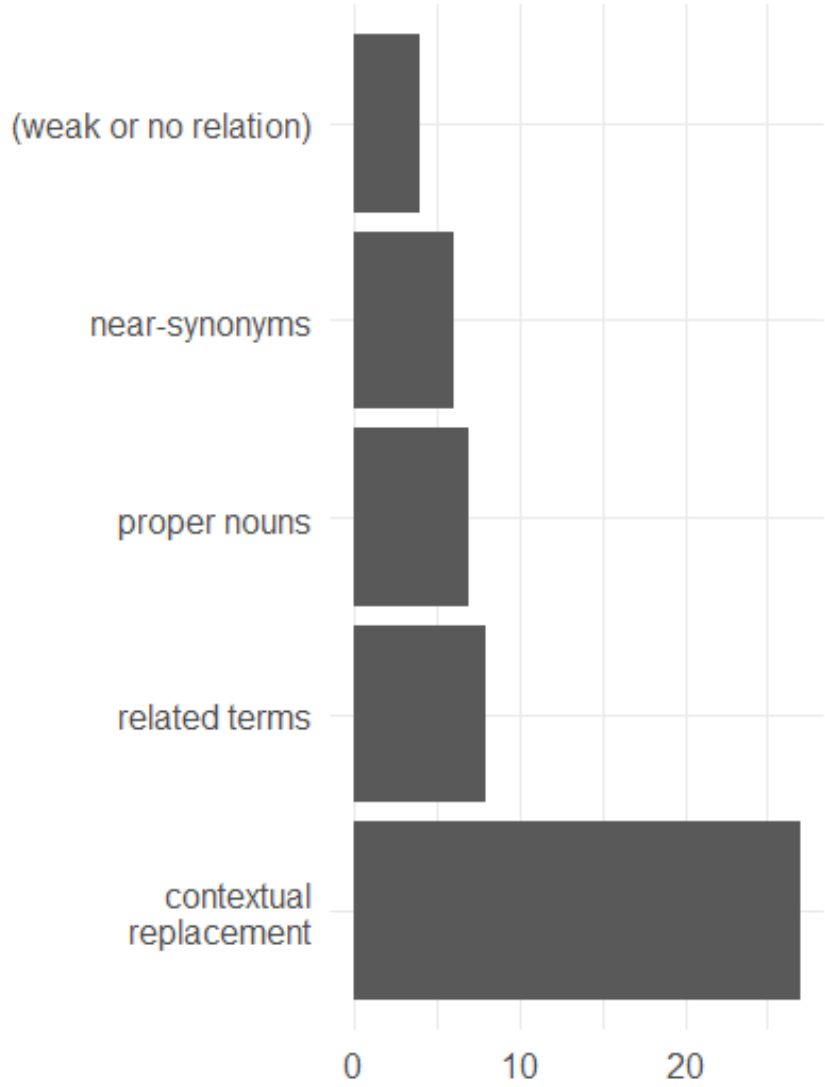
- Communicative need, after controlling for a slew of other lexicostatistical variables, describes a small amount of variance in competitive interactions
- Small effect, but consistent across languages and genres
- Presumably high communicative need facilitates the co-existence of similar words (more complex lexical subspace)

The complexity-informativeness tradeoff and the optimal front



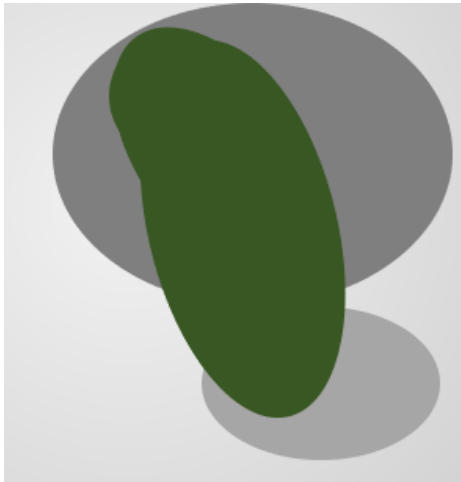
Further evaluation

- But: direct synonym competition is very rare!
- Sample: COHA, equalization range <0.2 & number of losers <4 (n=52)
- near-synonym competition:
  - aeroplane → airplane, close-up → close shot, appropriation → funding
    apartment+inn → motel
- some proper nouns
  - guerrilla → Taliban, Yugoslav → Algerian
- mostly contextual, in-topic replacements:
  - railway → airline, opera+concert → movie
    atomic bomb → ballistic missile
- Still, advection predicts if replacement or not



Conceptual similarity and communicative need shape colexification: an experimental study (Karjus, Blythe, Kirby, Wang, Smith, in prep)

- Xu et al 2020, “Conceptual relations predict colexification across languages”, using 200+ languages
- Similar and associated senses (e.g. FIRE and FLAME) are more frequently **colexified** in world’s languages than unrelated or weakly associated meanings (like FIRE and SALT)



Conceptual similarity and communicative need shape colexification: an experimental study (Karjus, Blythe, Kirby, Wang, Smith, in prep)

- Xu et al 2020, “Conceptual relations predict colexification across languages”, using 200+ languages
- Similar and associated senses (e.g. FIRE and FLAME) are more frequently **colexified** in world’s languages than unrelated or weakly associated meanings (like FIRE and SALT)
- …but culture specific **communicative needs** should affect likelihood of colexification – e.g. if it is necessary for efficient communication to distinguish some similar meanings
- E.g. ICE and SNOW: less likely to be colexified in cold climates (Regier et al 2016)

Conceptual similarity and communicative need shape colexification: an experimental study

- What is the cognitive mechanism though that leads to this cross-linguistic tendency?
- Maybe we can test these two claims experimentally?
- 4 experiments: initial one with student sample, replication on Mechanical Turk, 2 more experiments with different conditions
- Dyadic communication game setup, 2 players, take turn sending and guessing messages (cf. Kirby et al 2008, Winters et al 2015)
- 135 rounds each (data from the first 1/3 of the game excluded)

- 10 meanings total
- 4 distractor meanings
- from Simlex999

- 6 target meanings
- 3 pairs

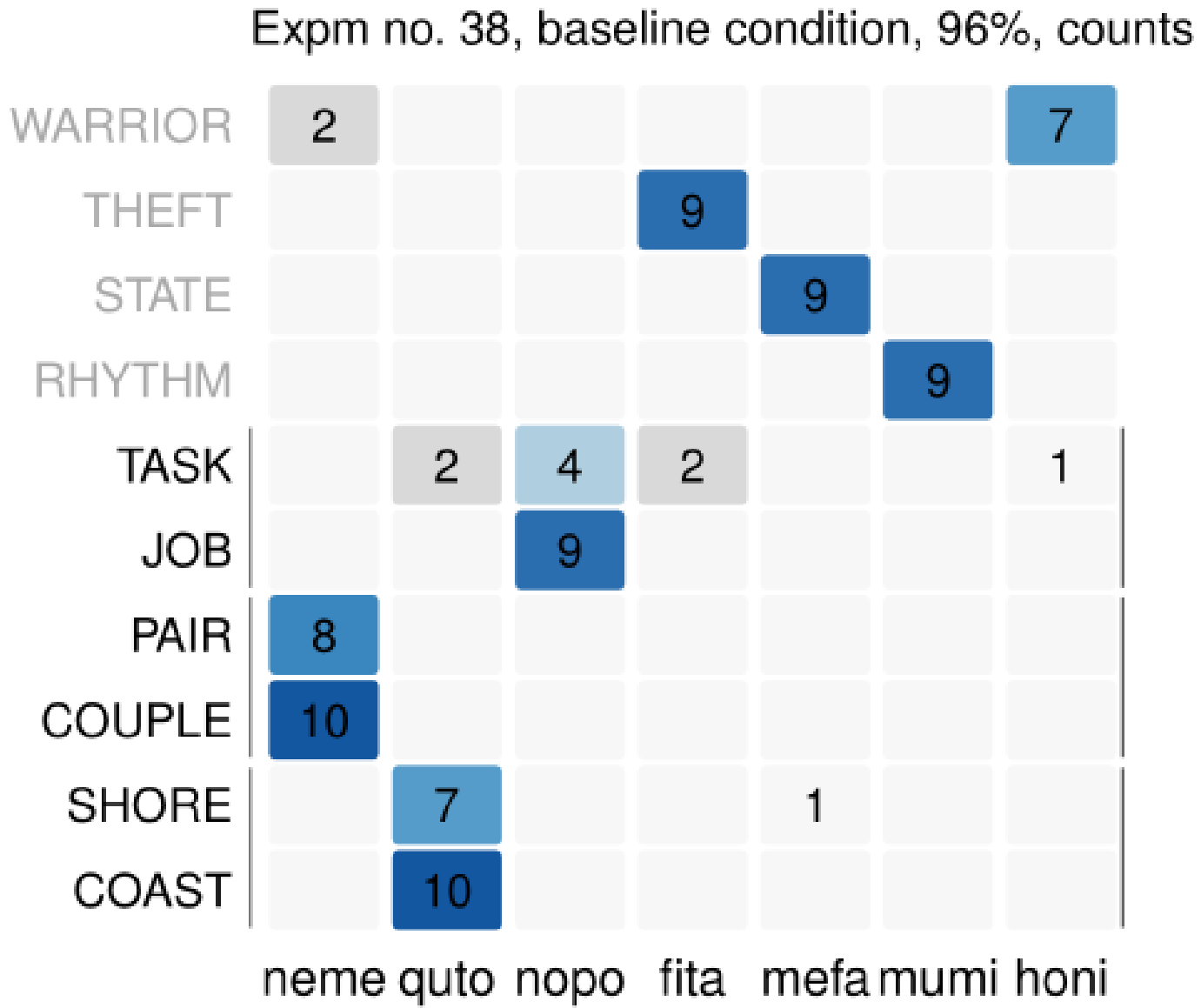
- Baseline: pairs co-occur uniformly
- Target condition: **similar ones occur together more often!**

The game


Player 1
Players connected: 2. Score: 0/2
area fashion
Communicate area using...
piti
wuli
liha
naru
mano
himu
qata

The game

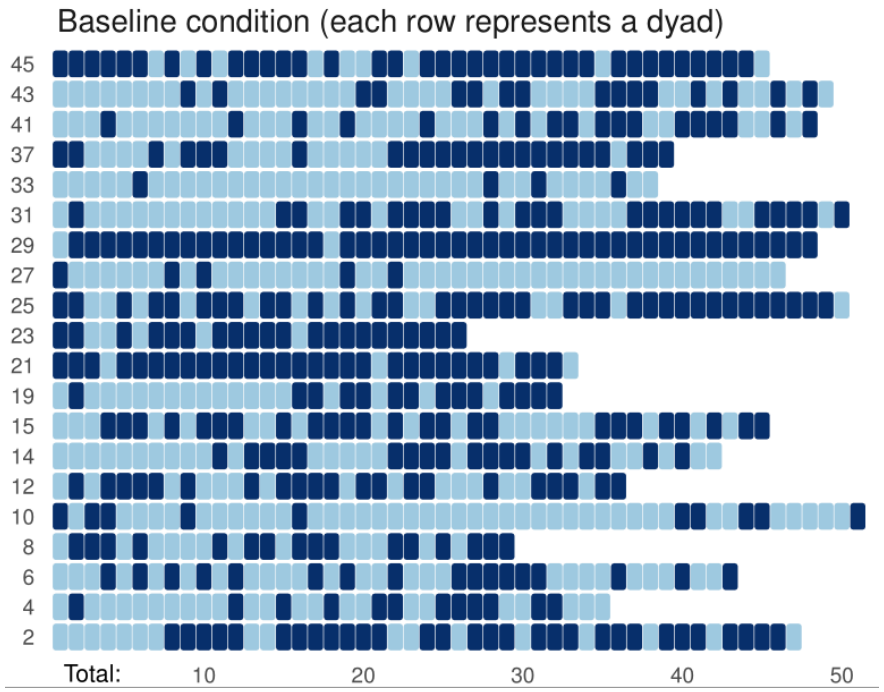

Player 1
Players connected: 2. Score: 0/2
area fashion
Communicate area using...
piti
wuli
liha
naru
mano
himu
qata
Player 2
Players conn
area fashion
Waiting for message...
Player 1
Players connected: 2. Score: 0/2
Sent area using piti
Stand by...
Player 2
Players connected: 2. Score: 0/2
area fashion
Message: piti
This means:
area
fashion

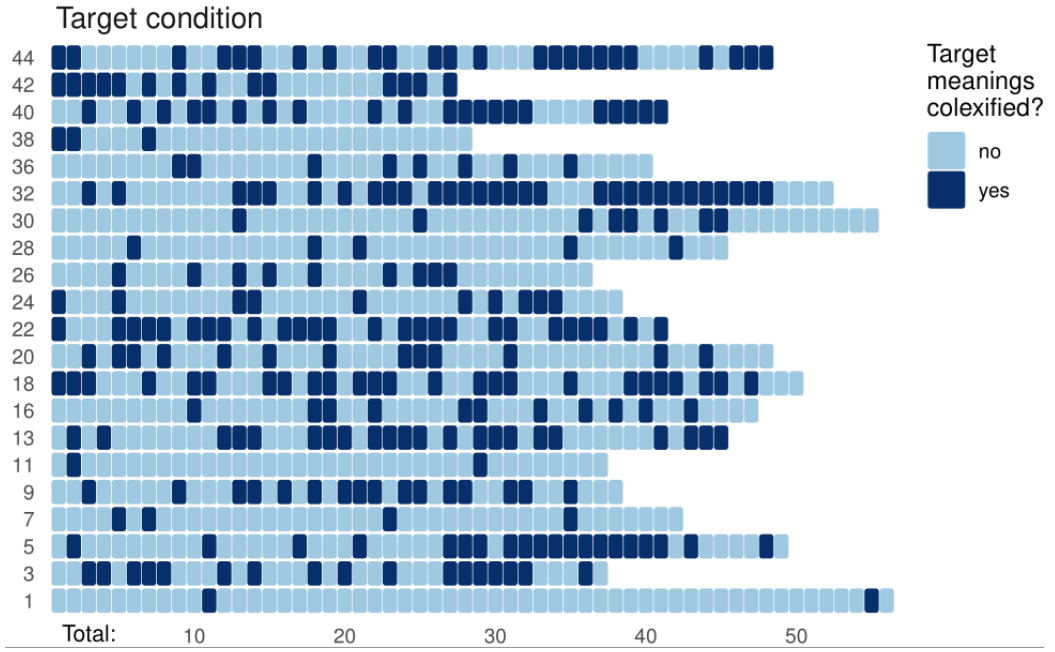

Expm no. 38, baseline condition, 96%, counts
WARRIOR
THEFT
STATE
RHYTHM
TASK
JOB
PAIR
COUPLE
SHORE
COAST
2
7
9
9
9
2
4
2
1
9
8
10
7
1
10
neme
quto
nopo
fita
mefa
mumi
honi

7 signals

Analysis

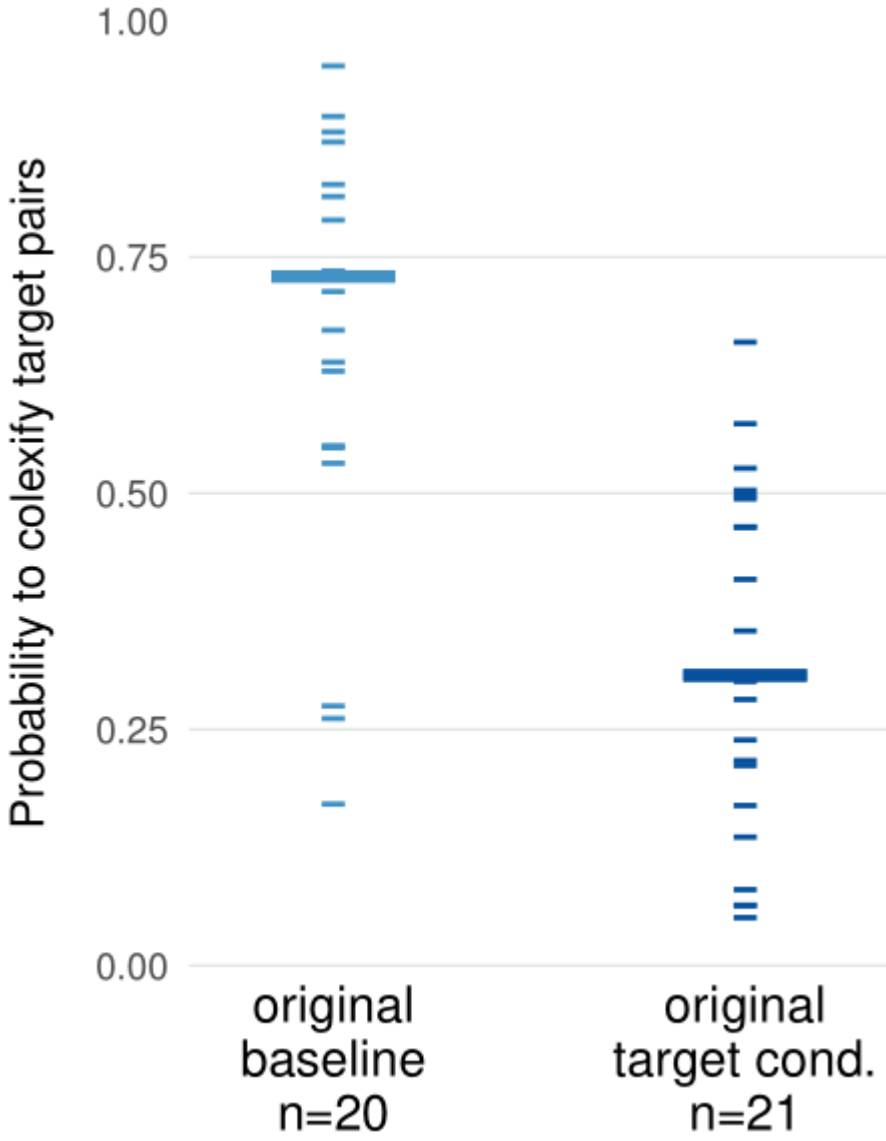
- Exclude low-accuracy dyads (41 left)
- Iterate through each experiment, record each instance of colexification (same signal, different meaning) involving a target meaning; n=1218.
- Logistic mixed effects regression; control for dyads, meaning pairs. Are similar meanings less likely to be colexified in the target condition?





Results

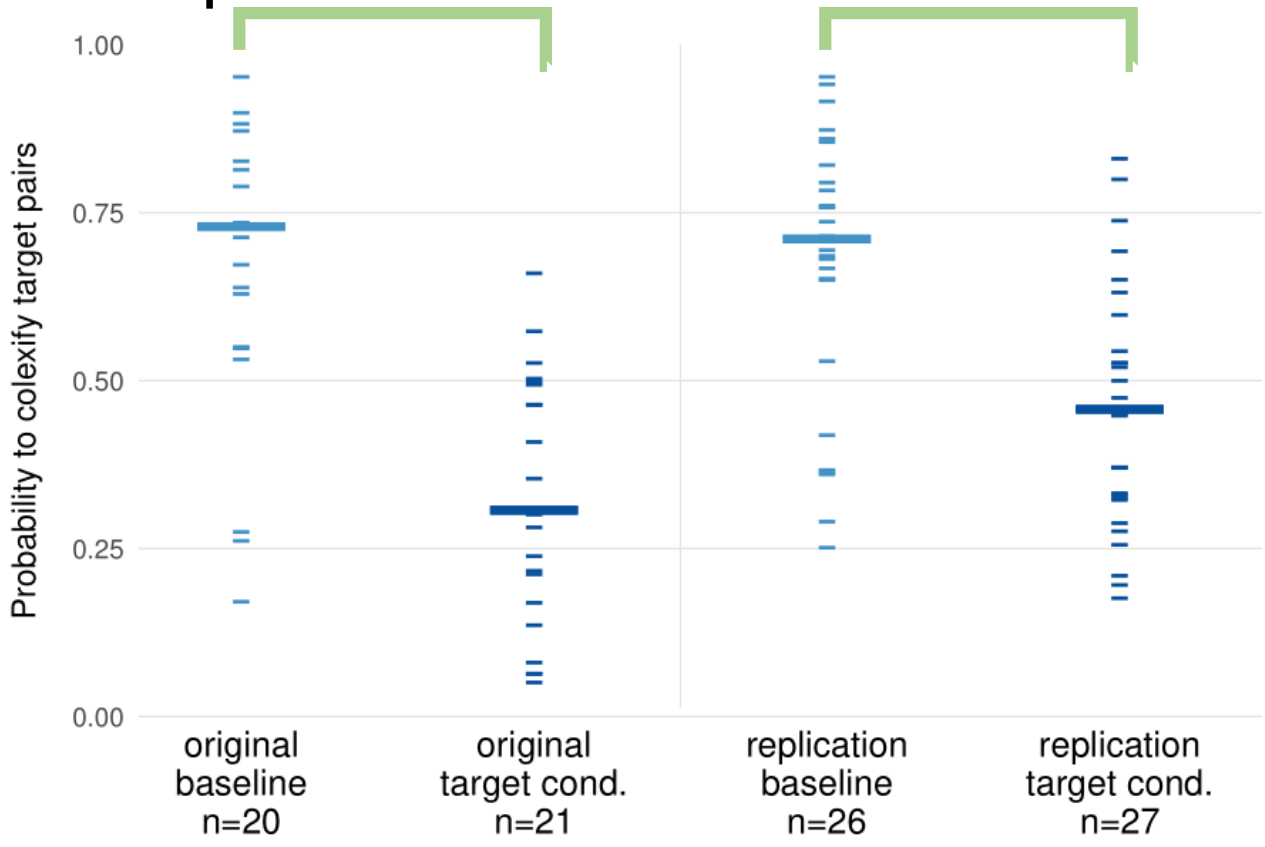
- Yes (p=0.001). This includes interaction with round – some dyads change preferences over the course of the game.
- When no pressure to distinguish particular meanings (baseline condition), speakers prefer to colexify similar meanings (confirms Xu et al 2020)
- When need arises to distinguish similar meanings (target condition), speakers less likely to colexify them (confirms hypothesis that communicative needs may block colexification of related concepts)



Follow-up experiments

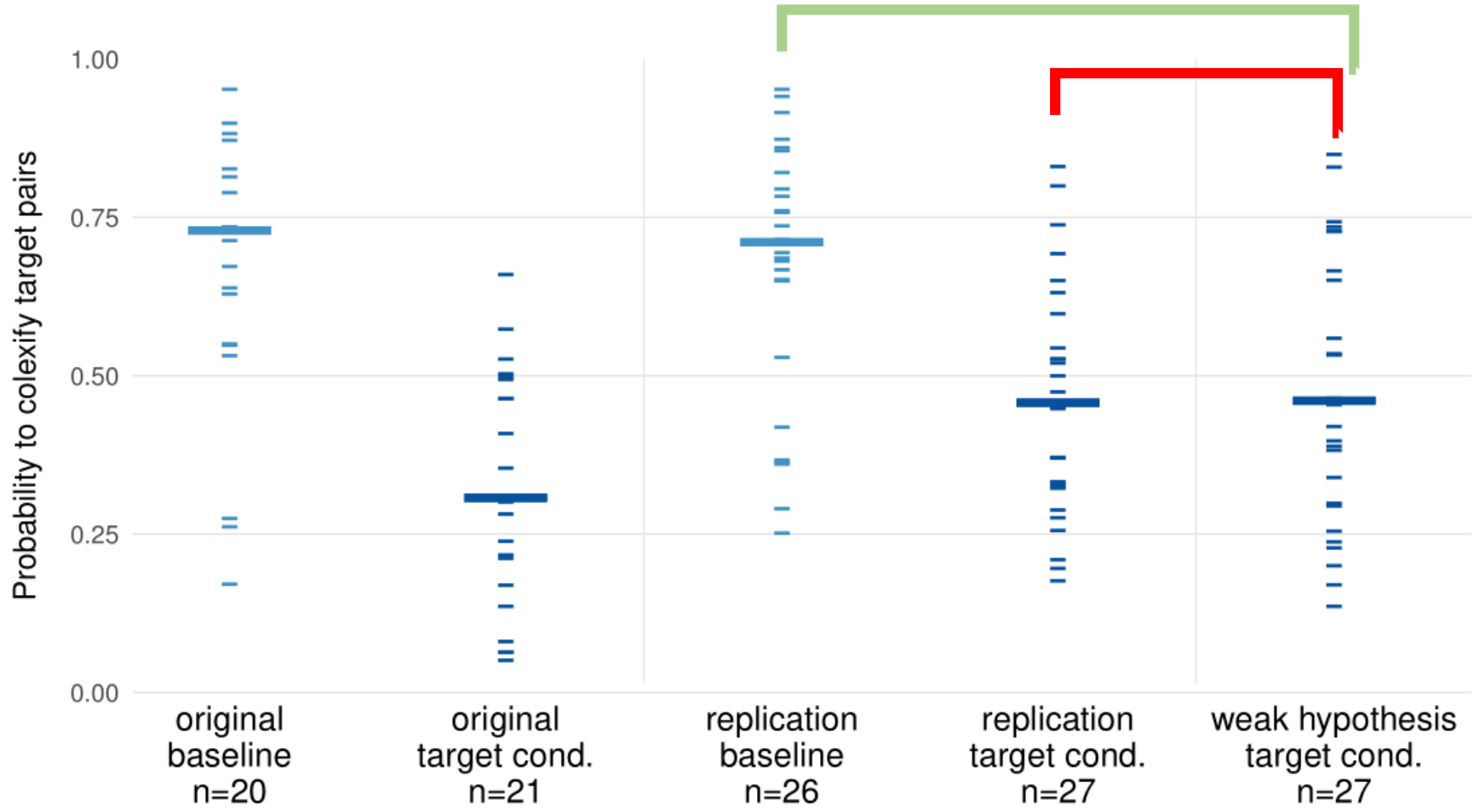
- Switch to Mechanical Turk: initial experiment was planned to be run the lab in spring 2020, but the apocalypse happened
- Experiment 2, replication on MTurk: exact same setup with 2 conditions; results of experiment 1 replicated.
- Lower accuracy: 79 dyads,

  could use data only from 53.



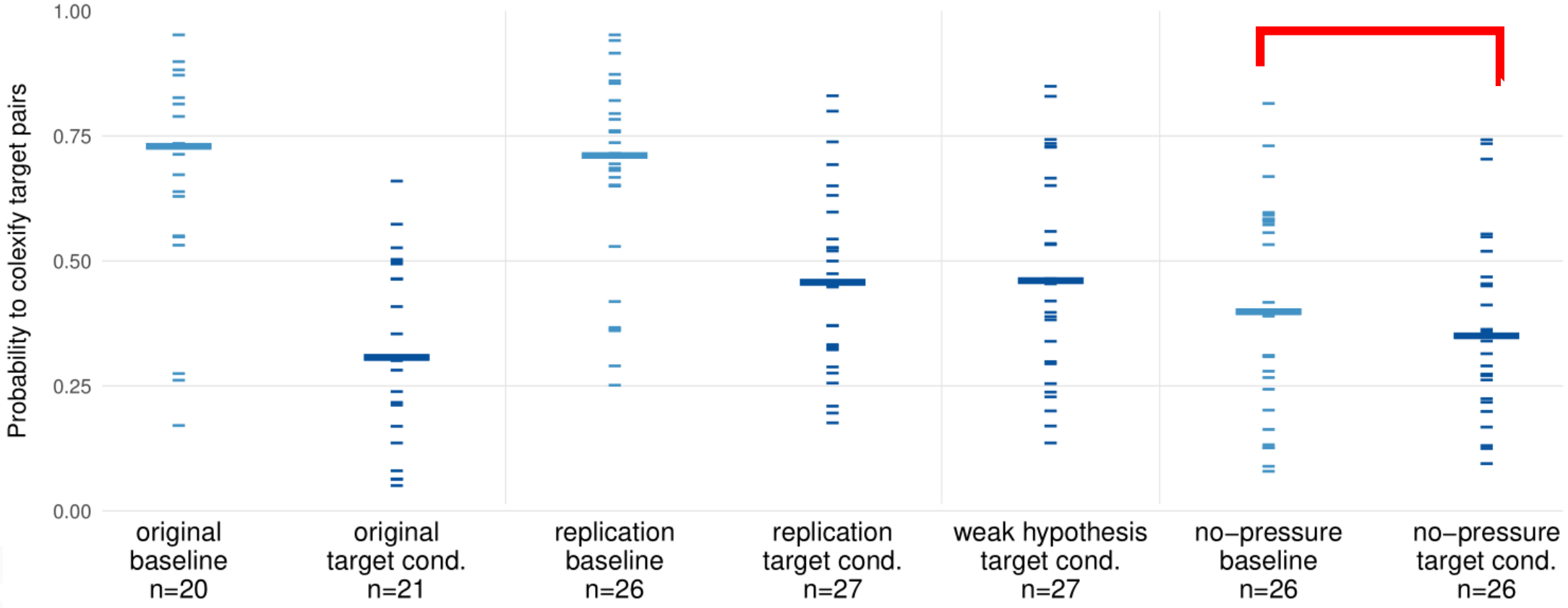
Follow-up experiments

- Experiment 3, target condition only: introduce similar-meaning pairs into the distractor set to make colexifying them more natural.

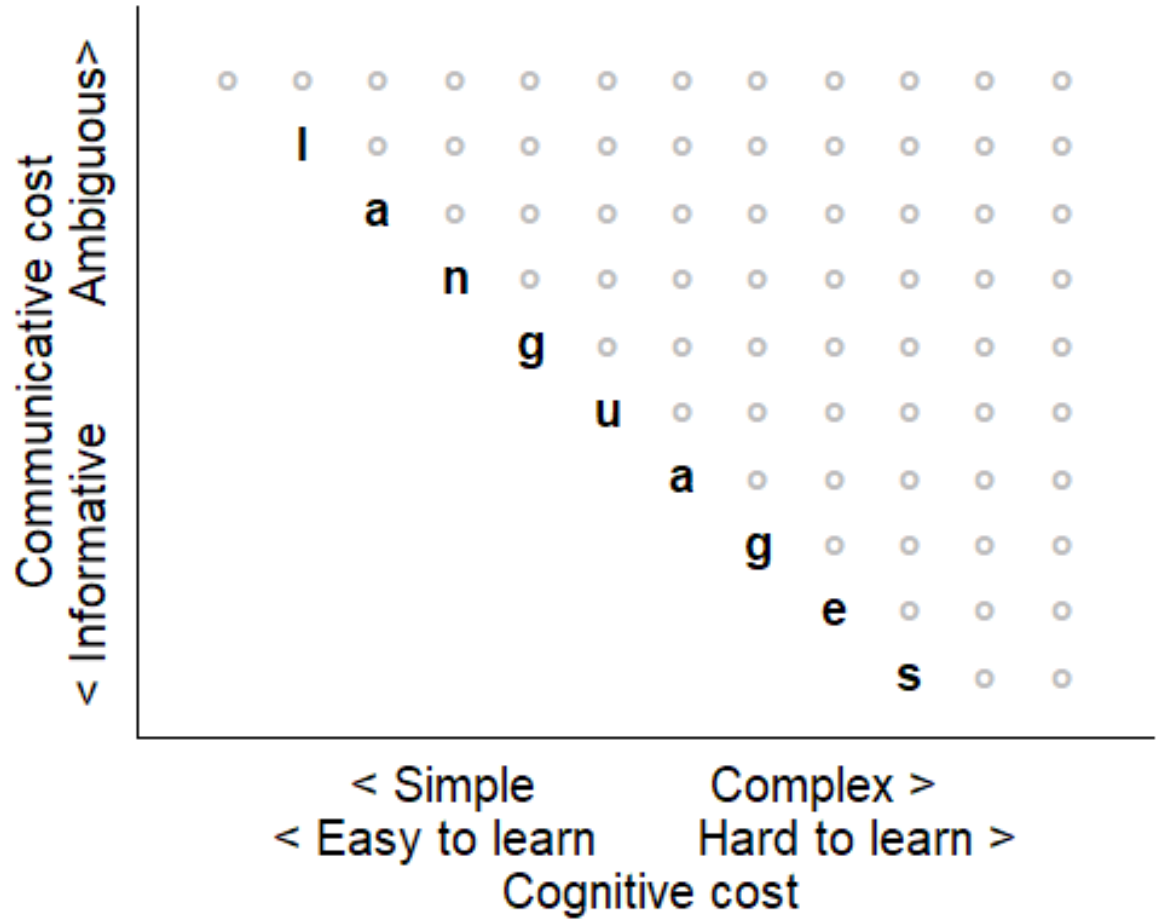


Follow-up experiments

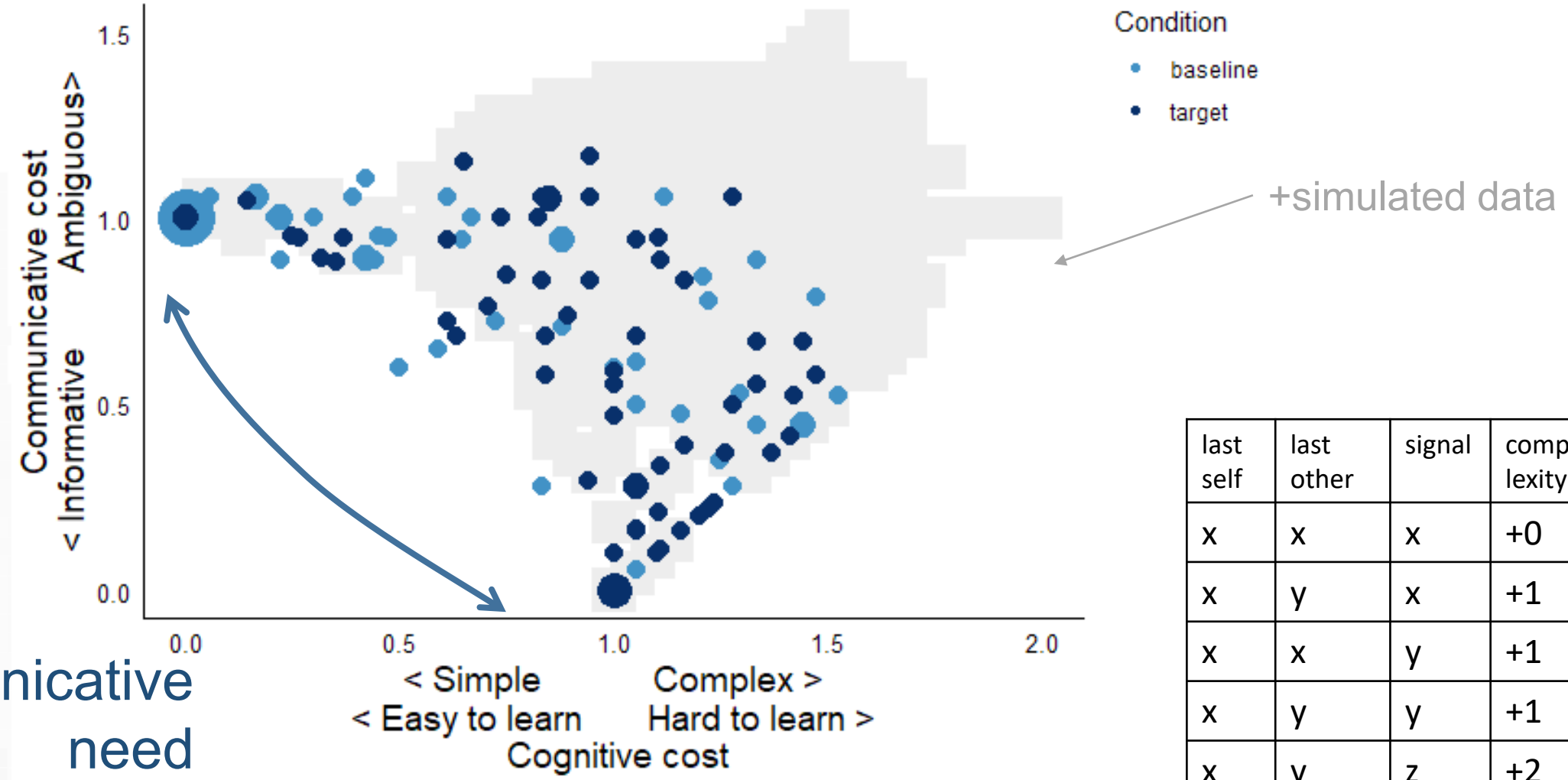
- Experiment 4: no pressure to colexify (10 signals for 10 meanings). No effect, and participants make significantly more use of the bigger signal space. But: natural language does have pressure to simplify (can't have infinite lexicons).



The complexity-informativeness tradeoff and the optimal front



The complexity-informativeness tradeoff and the optimal front

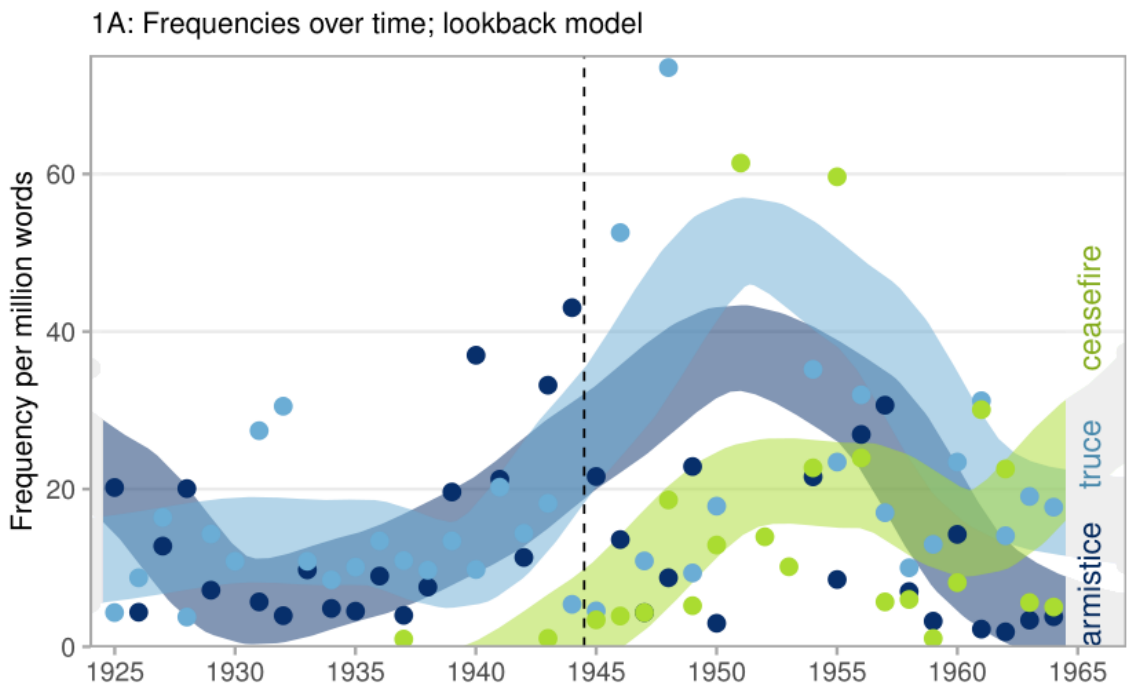


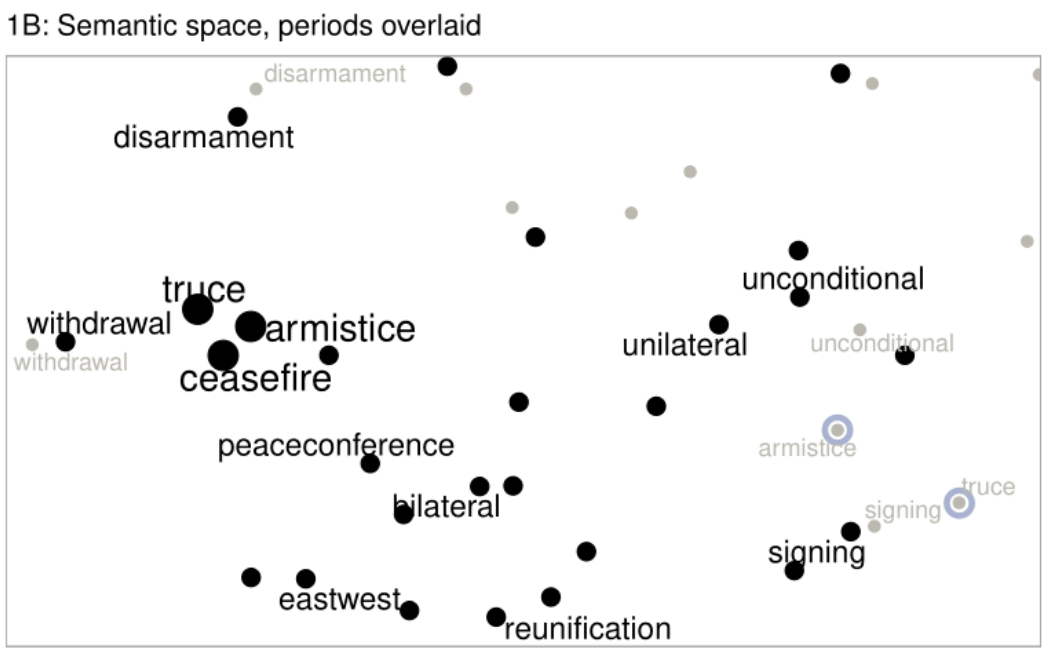
communicative need

last self	last other	signal	comp-lexity	ambi-guity
x	x	x	+0	+1
x	y	x	+1	+0
x	x	y	+1	+1
x	y	y	+1	+2
x	y	z	+2	+1

Discussion

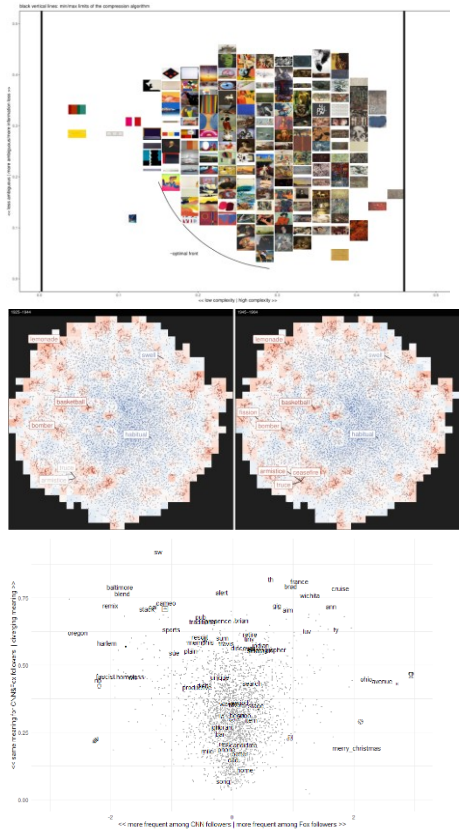
- Experimental results describe an individual-level lexical choice mechanism which produces results in line with typological colexification tendencies (Xu et al 2020) as well as the communicative need hypothesis
- Work in process: a model of lexical density (~extent of colexification) applied to embeddings trained on diachronic corpora



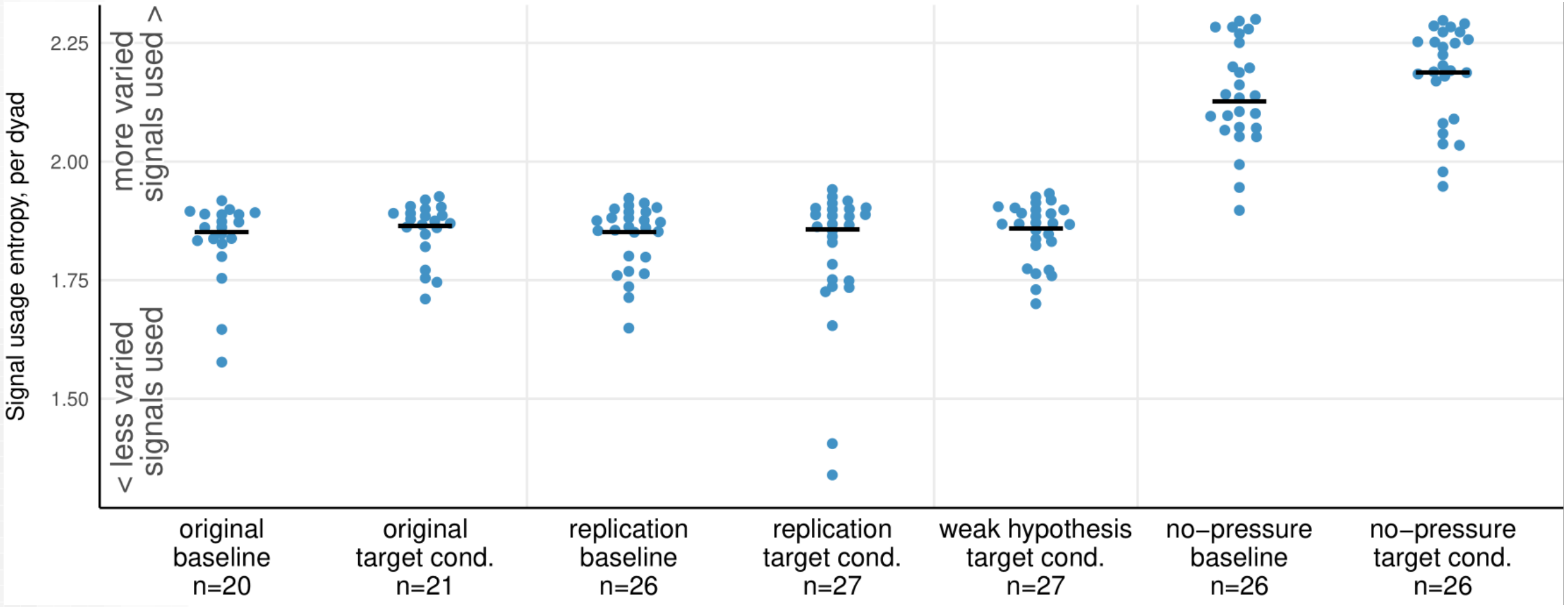


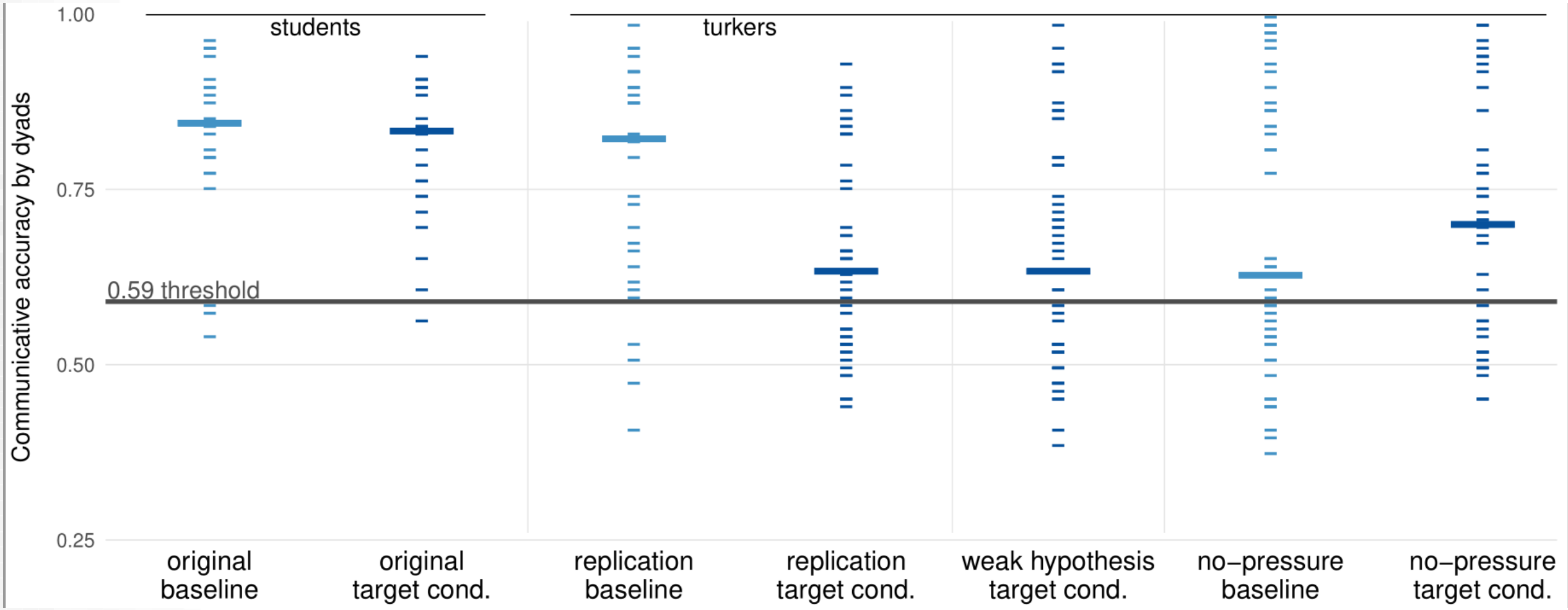
Conclusions

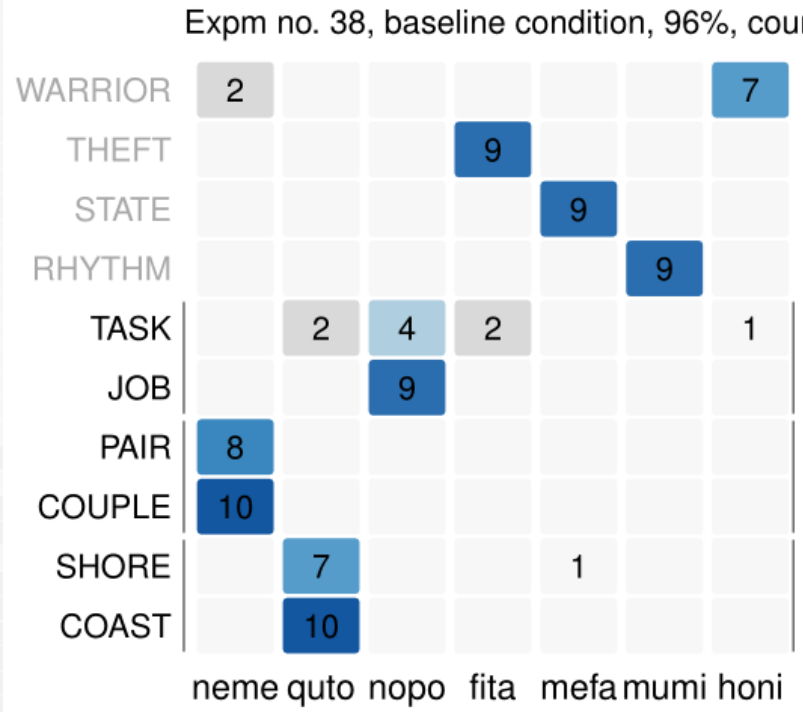
- Converging typological, experimental and corpus evidence supports the argument for the role of communicative need from earlier cross-linguistic research
- There are many reasons why languages change; one of them is adaption to the changing needs of their speakers
- Future: apply the complexity-informativeness approach to products of cumulative cultural evolution other than language
- Iron out the competition model, apply to data other than language
- Other stuff: research into semantics-driven misunderstanding and semantic divergence on social media

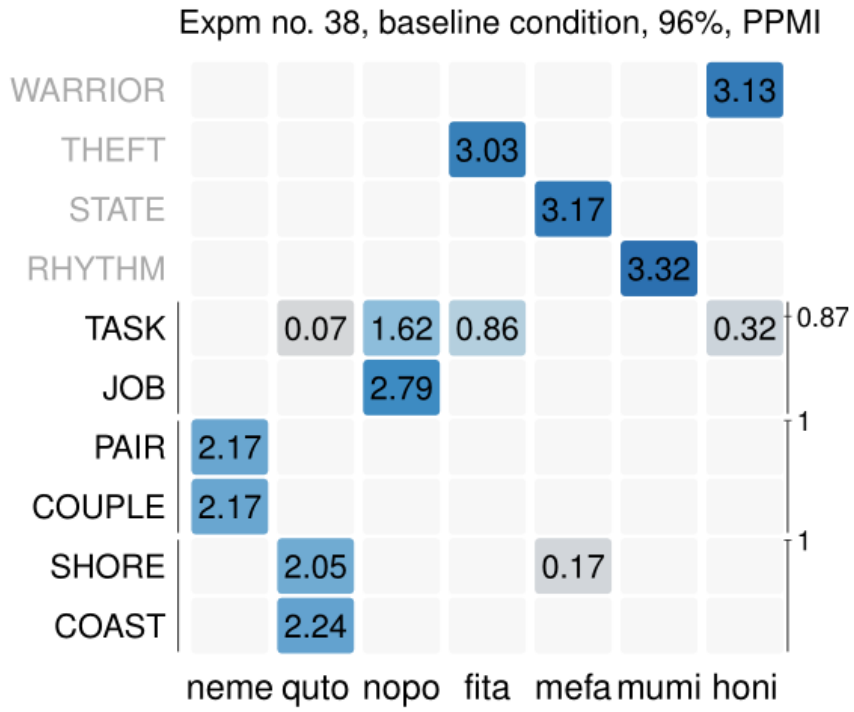


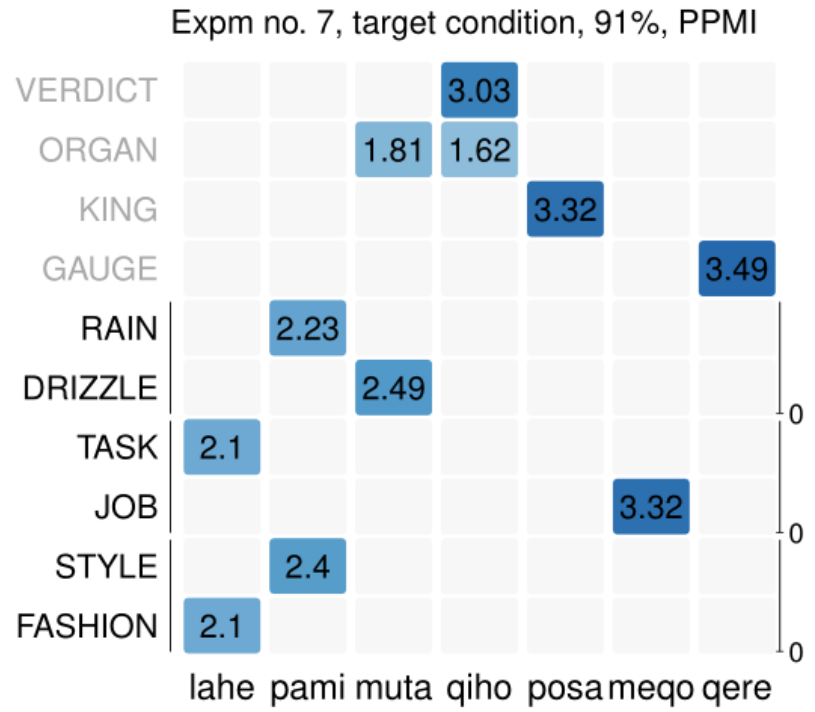
Appendix

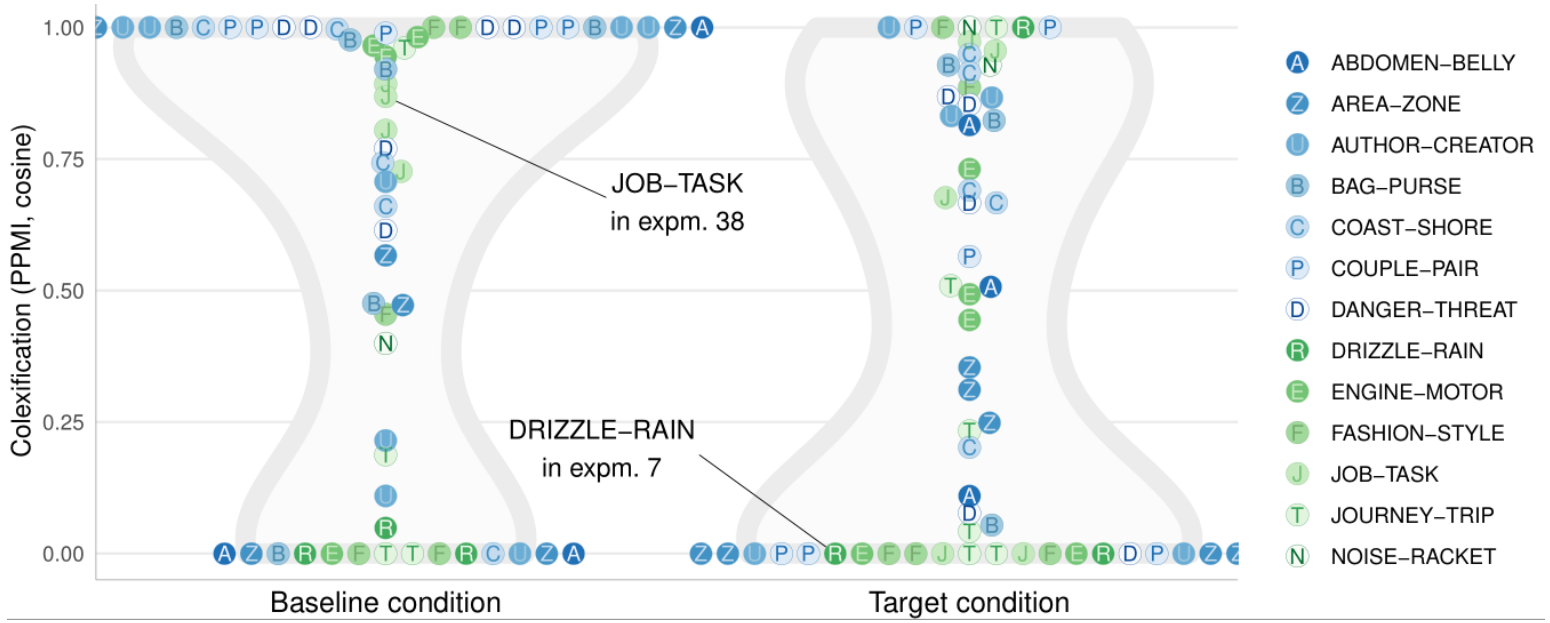

Signal usage entropy, per dyad
2.25
2.00
1.75
1.50
more varied signals used >
< less varied signals used
original baseline n=20
original target cond. n=21
replication baseline n=26
replication target cond. n=27
weak hypothesis target cond. n=27
no−pressure baseline n=26
no−pressure target cond. n=26

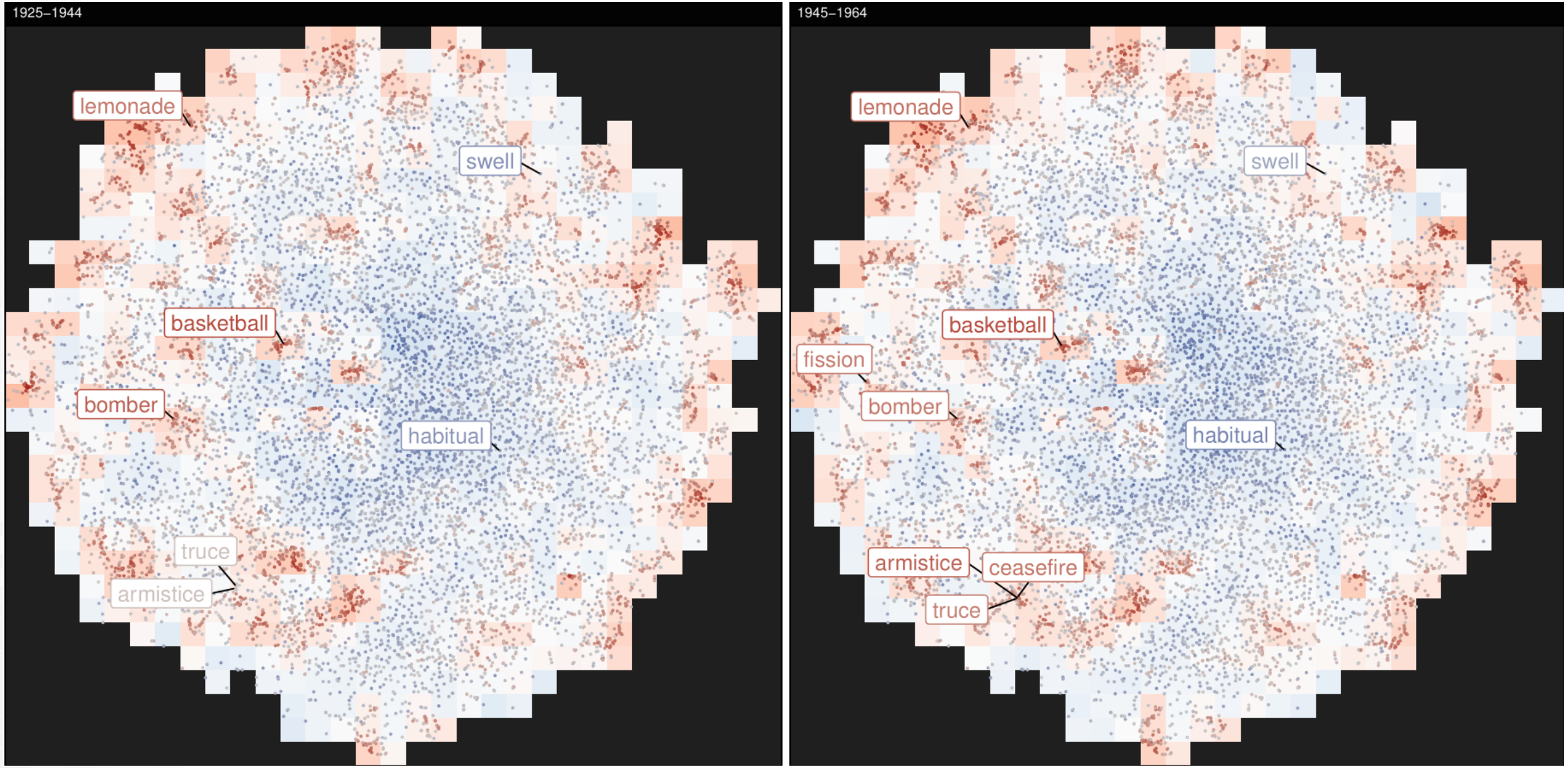

1.00
0.75
0.50
0.25
Communicative accuracy by dyads
students
turkers
0.59 threshold
original baseline
original target cond.
replication baseline
replication target cond.
weak hypothesis target cond.
no–pressure baseline
no–pressure target cond.

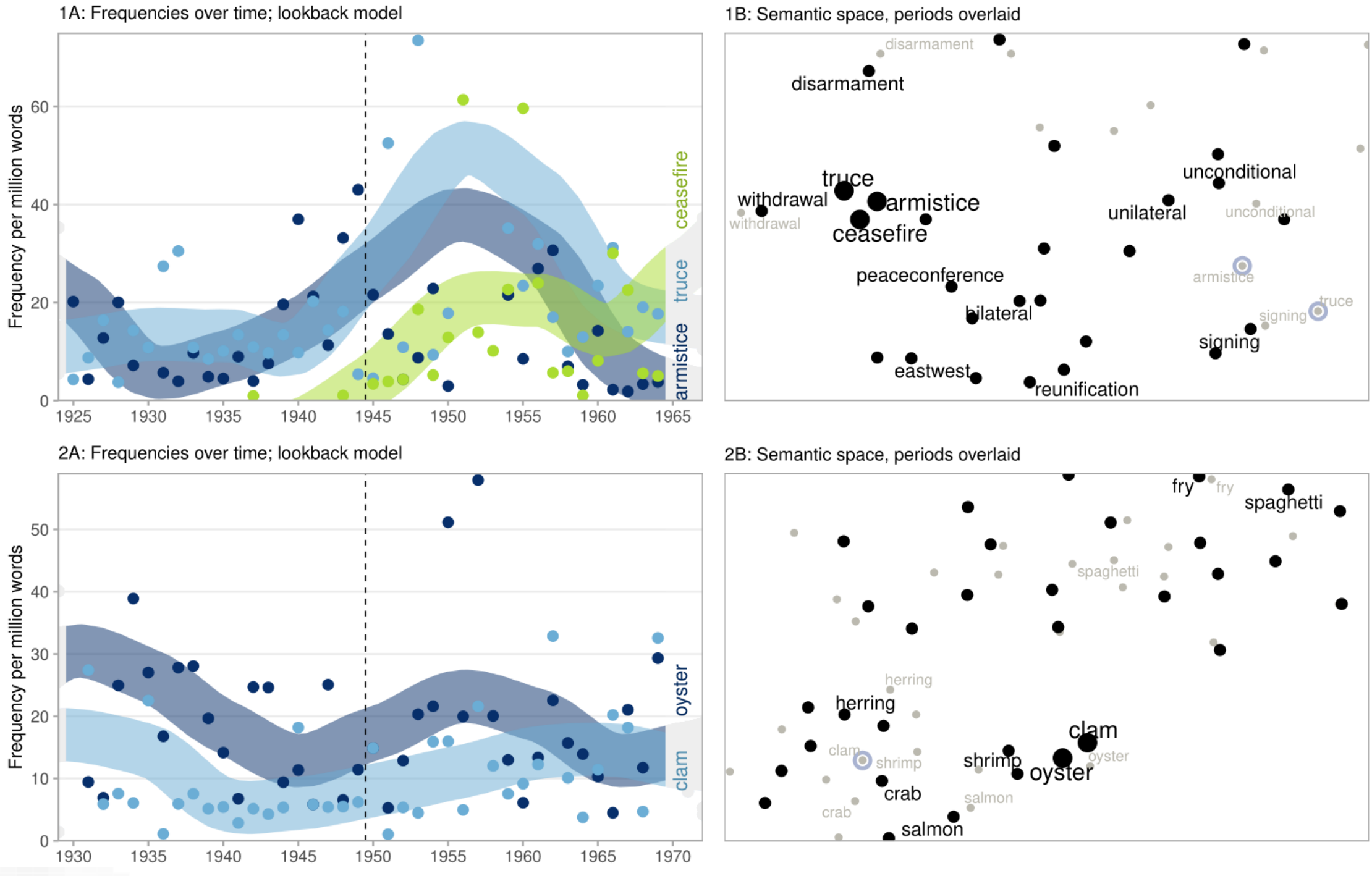

Expm no. 38, baseline condition, 96%, counts
WARRIOR 2 7
THEFT 9
STATE 9
RHYTHM 9
TASK 2 4 2 1
JOB 9
PAIR 8
COUPLE 10
SHORE 7 1
COAST 10
neme quto nopo fita mefa mumi honi


Expm no. 38, baseline condition, 96%, PPMI
WARRIOR 3.13
THEFT 3.03
STATE 3.17
RHYTHM 3.32
TASK 0.07 1.62 0.86 0.32
JOB 2.79
PAIR 2.17
COUPLE 2.17
SHORE 2.05 0.17
COAST 2.24
0.87
1
1
neme quto nopo fita mefa mumi honi


Expm no. 7, target condition, 91%, PPMI
VERDICT 3.03
ORGAN 1.81 1.62
KING 3.32
GAUGE 3.49
RAIN 2.23
DRIZZLE 2.49
TASK 2.1
JOB 3.32
STYLE 2.4
FASHION 2.1
0
0
0
lahe pami muta qiho posa meqo qere


Colexification (PPMI, cosine)
1.00
0.75
0.50
0.25
0.00
JOB–TASK
in expm. 38
DRIZZLE–RAIN
in expm. 7
Baseline condition
Target condition
ABDOMEN–BELLY
AREA–ZONE
AUTHOR–CREATOR
BAG–PURSE
COAST–SHORE
COUPLE–PAIR
DANGER–THREAT
DRIZZLE–RAIN
ENGINE–MOTOR
FASHION–STYLE
JOB–TASK
JOURNEY–TRIP
NOISE–RACKET


1925–1944
lemonade
swell
basketball
bomber
habitual
truce
armistice
1945–1964
lemonade
swell
basketball
fission
bomber
habitual
armistice
ceasefire
truce


1A: Frequencies over time; lookback model
Frequency per million words
0
20
40
60
1925
1930
1935
1940
1945
1950
1955
1960
1965
ceasefire
truce
armistice
1B: Semantic space, periods overlaid
disarmament
disarmament
truce
withdrawal
armistice
withdrawal
ceasefire
unconditional
unilateral
unconditional
peaceconference
armistice
bilateral
truce
signing
signing
eastwest
reunification
2A: Frequencies over time; lookback model
Frequency per million words
0
10
20
30
40
50
1930
1935
1940
1945
1950
1955
1960
1965
1970
oyster
clam
2B: Semantic space, periods overlaid
fry
fry
spaghetti
spaghetti
herring
herring
clam
clam
shrimp
shrimp
oyster
oyster
crab
salmon
crab
salmon