# Not just form, Not just meaning

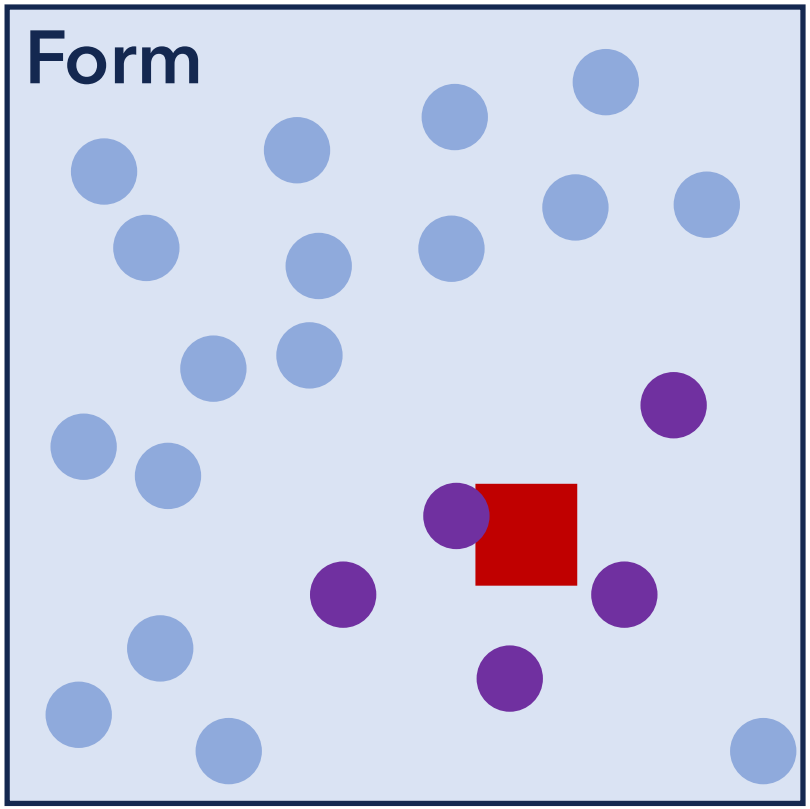
## Consistency in form-meaning mappings predicts age of acquisition beyond semantic and form neighborhood density

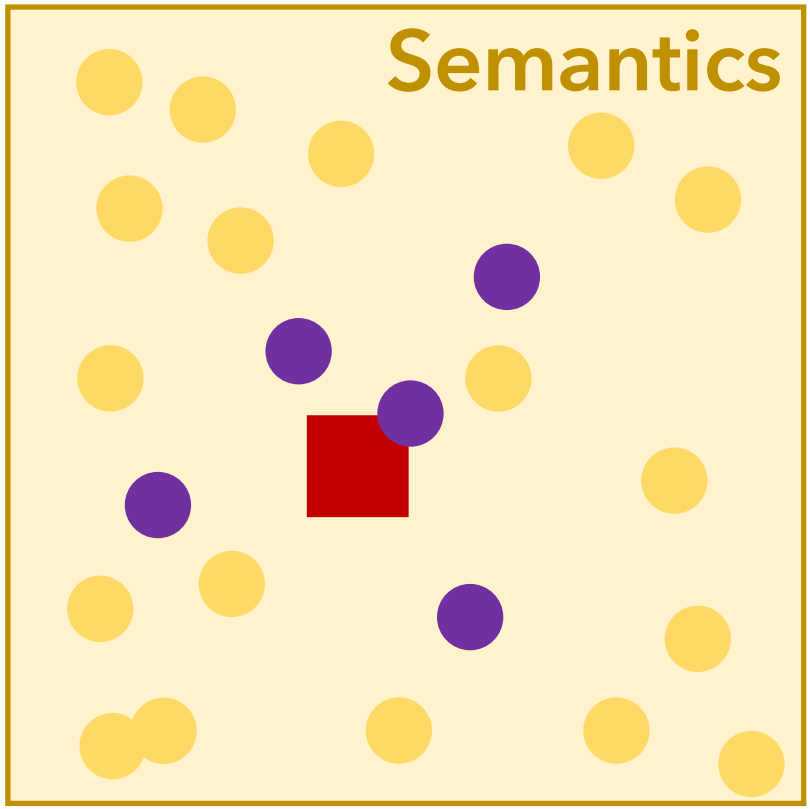
Giovanni Cassani
Niklas Limacher

AMLaP 2021

Tilburg University

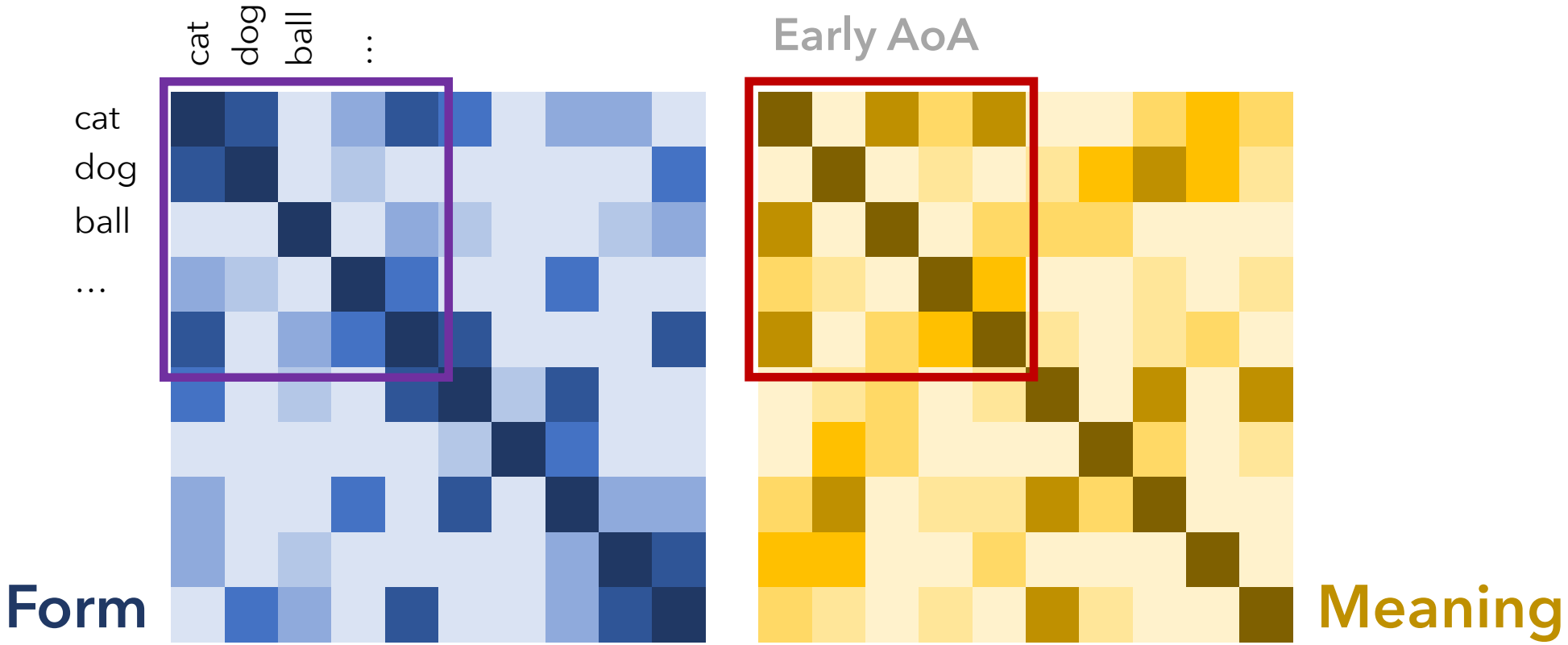
# Form-Semantic Consistency (FSC)



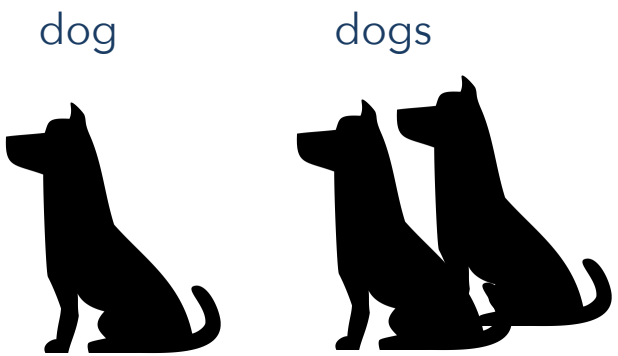
# Why AoA?

# Systematicity and acquisition



# Systematicity

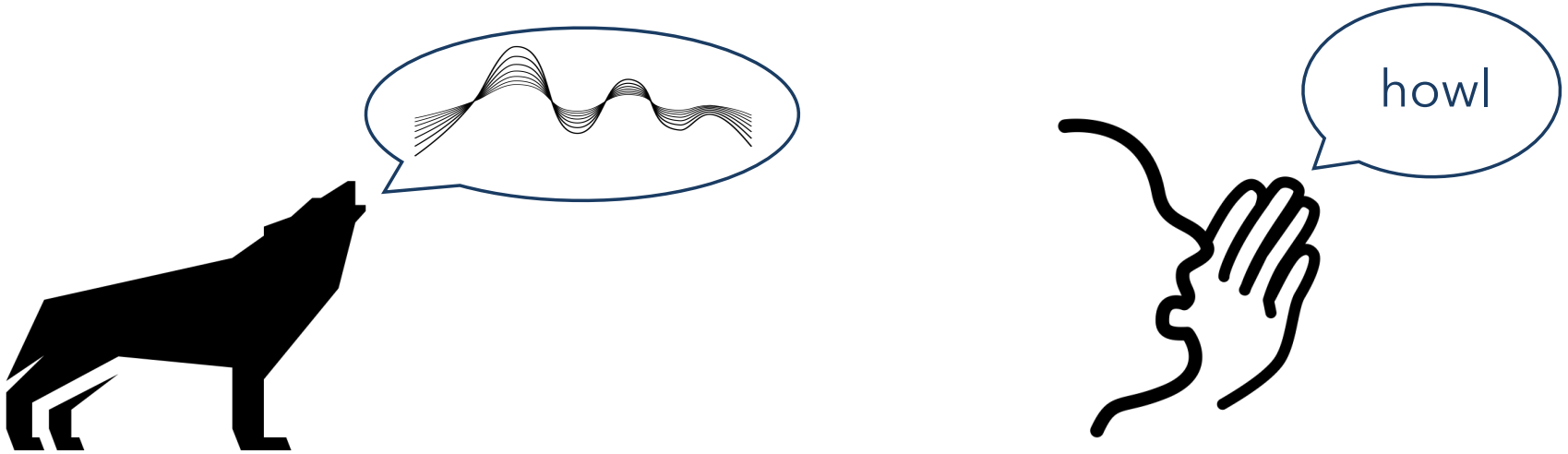
What about the interface between word form and language internal patterns?






# Iconicity

Several studies have shown that there is a **learning advantage** for iconic words, where word forms reflect perceptual information.



# But…

What if we don't have a meaning representation for a word and want to check its degree of systematicity?

What if systematicity's effect in learning boils down to words being entrenched in the language network in both form and meaning?

What if systematicity isn't really capturing what we think it is?

# The working hypothesis

Learning precise semantic representations from exposure to language and context takes time.

Maybe the word form itself bootstraps this process by giving hints, such that **words for which it's easier to guestimate the meaning from the form are learned earlier.**

# Why FSC?

# Tested and proven: RTs in LD

- Marelli, Amenta, Crepaldi [2015]: OSC predicts RTs in visual LD (linear models)
- Amenta, Marelli, Sulpizio [2017]: PSC also predicts RTs in visual LD (mediation analysis)
- Hendrix & Sun [2020]: OSC predicts RTs in visual LD, especially early on (piece-wise GAMs)

# Tested and proven: priming effects

Amenta, Crepaldi, Marelli [2020]:

OSC explains independent variance in priming magnitude, suggesting that word semantics is accessed already at early stages of word processing and that crucially **semantic access is constrained by word orthography**.

# Tested and proven: reading times

Amenta, Hasenäcker, Crepaldi, Marelli [*submitted*]:

surprisal and OSC exert an influence on several eye-movement measures and interact in a non-linear way. When sentence context is moderately surprising, word **orthography becomes a precious source of additional information to activate word meaning**.

# A generative model unchained

So far systematicity has been construed as representational similarity analysis: are similarity patterns consistent across feature spaces? What about novel words?

FSC however works for pseudowords as well, and **all words start as non-words for every learner**!

# Then what is this study about?

1. Can FSC also predict AoA patterns?
2. Is the effect of systematicity robust when controlling for neighborhood effects in both form and meaning?
3. Is FSC robust to changes in the size and composition of the reference vocabulary?
4. Is FSC really capturing systematicity?

# How?

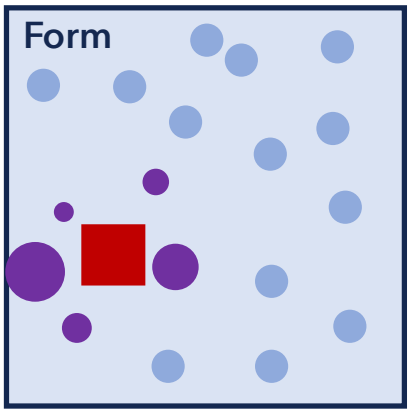
# Two ways to find neighbors

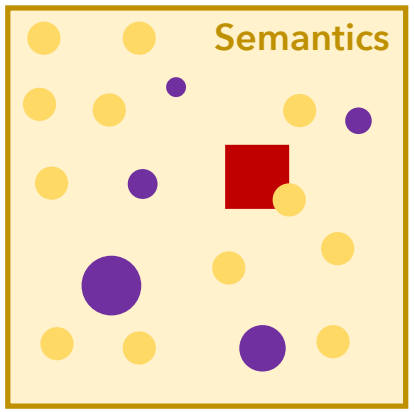
**Words including the target word**: *credit* is a neighbor of *red*

Levenshtein distance: *rad* and *rod* are neighbors of *red* (5 nn)

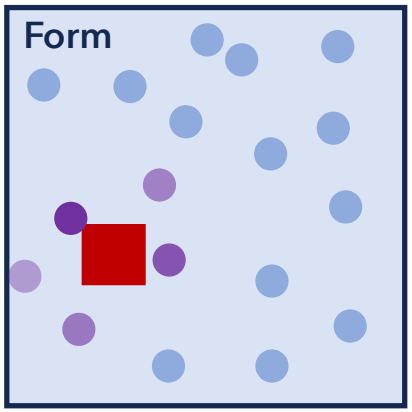
# Two ways of scaling semantic similarities

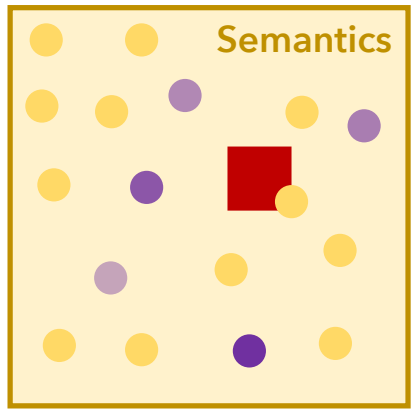
**Frequency weighting**: more frequent form-based neighbors weigh more




Inverse distance weighting: closer neighbors in form space weigh more



# Several datasets

- AoA norms (test-based and subjective)
- SUBTLEX-US (frequency and reference vocabulary)
- concreteness norms
- valence norms
- MALD dataset (Coltheart's N)
- iconicity norms
- MorphoLEX (mono v. polymorphemic words)
- CELEX for phonological transcriptions
- CBoW embeddings from Mandera et al [2017]

# Experiments

# Main analysis

- outcome variable: objective AoA norms
- method: Linear regression and Random Forests
- control variables: frequency, concreteness, valence, length in phonemes, PND (Coltheart's N), SND (20 nn), morphological complexity (Box-Cox transformed and z-standardized)
- fit baseline statistical model with control variables only
- add target variables/interactions and measure the change in AIC

# PSC (n = 6,407)

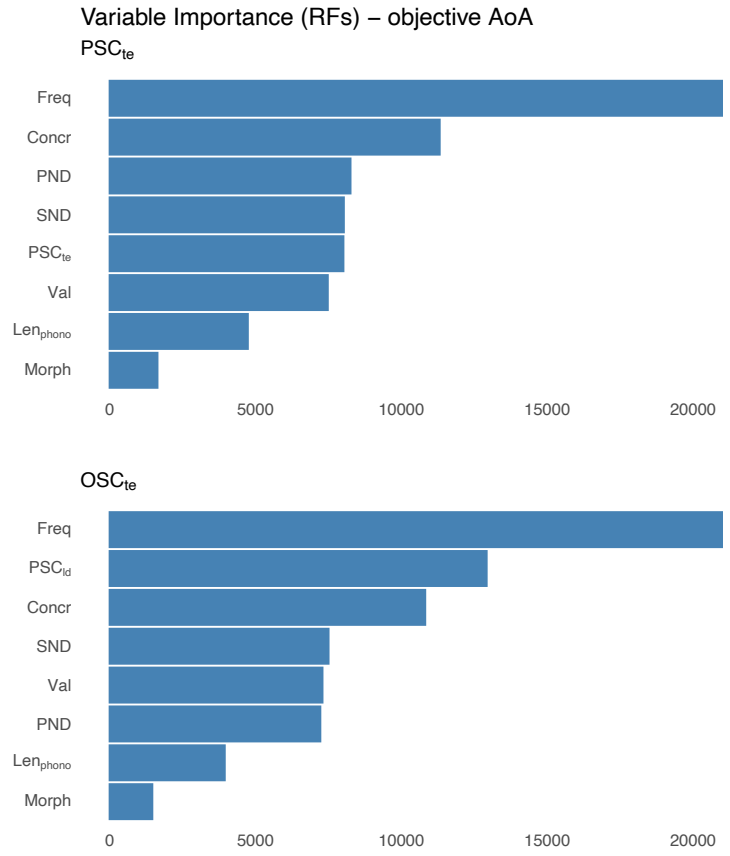
| Measure | β | se | t | p | ΔAIC |
|---|---|---|---|---|---|
| $PSC_{te}$ | -0.145 | 0.034 | -4.263 | < .001 | 16.17 |
| $PSC_{ld}$ | -0.738 | 0.046 | -15.968 | < .001 | 248.38 |

PSC has a negative effect on AoA: more systematic words are acquired earlier.

# Random Forests (n = 6,407)

$PSC_{ld}$ is a very important variable in RF, confirming collinearity did not bias the conclusion.

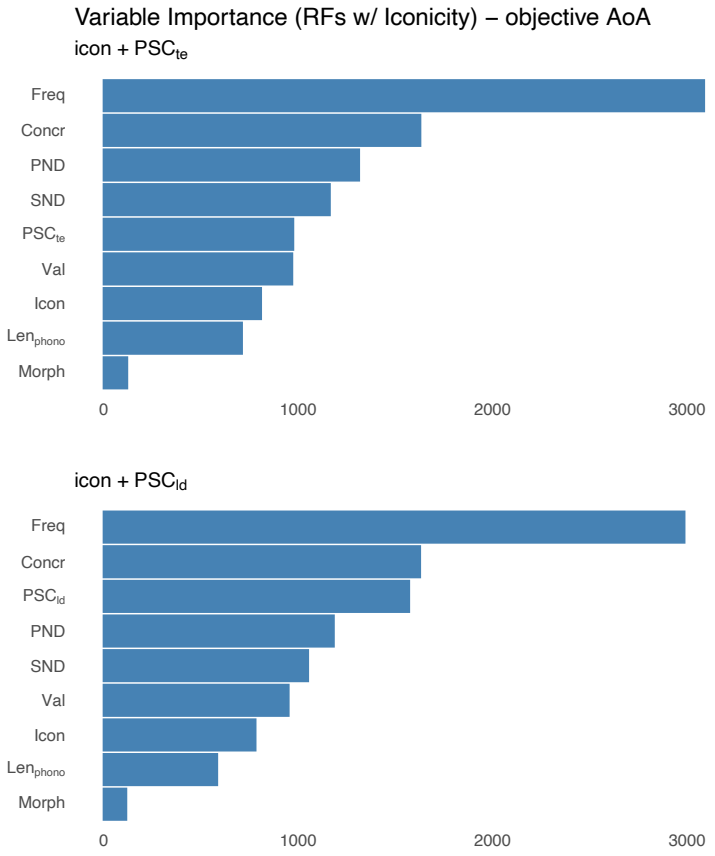
$PSC_{te}$ not as much, in line with the linear regression.



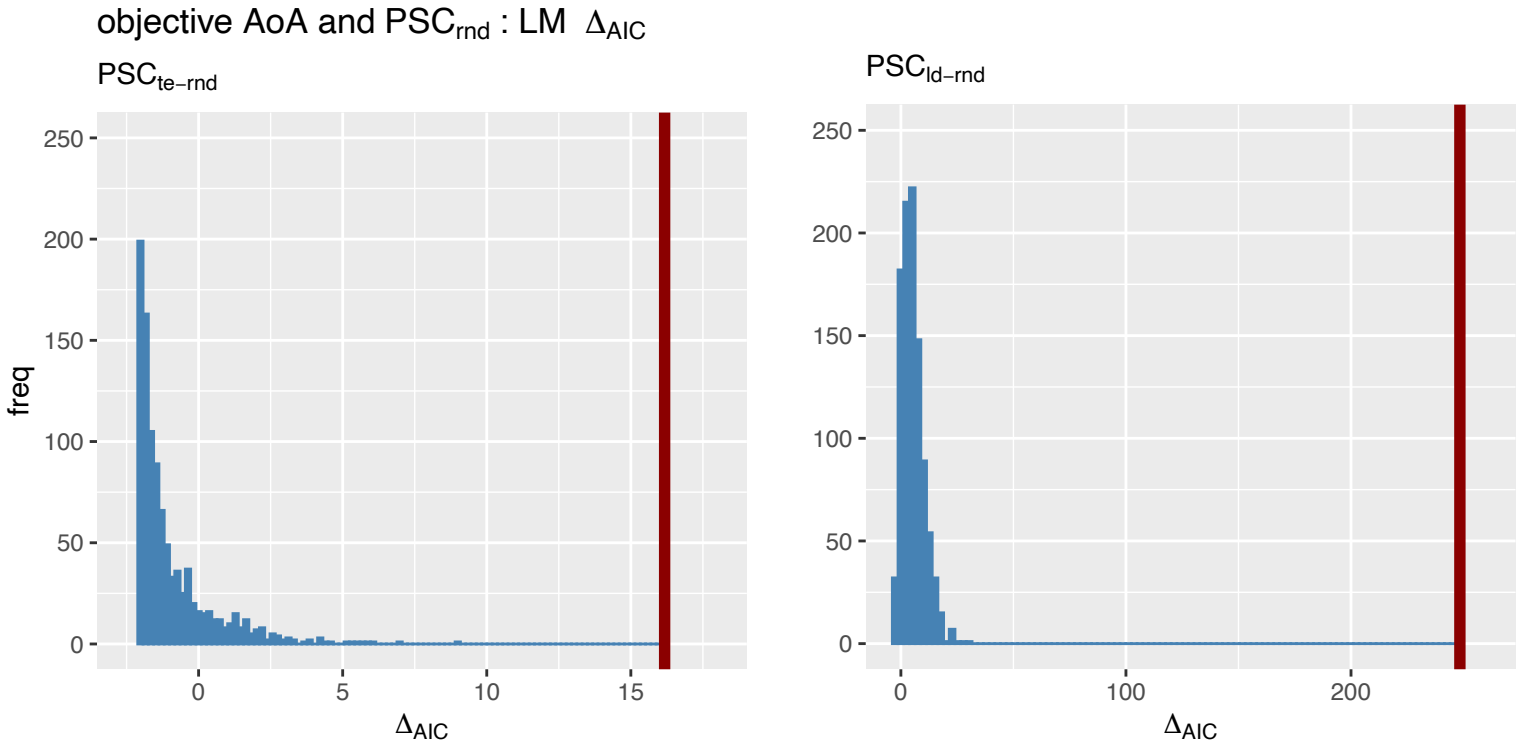
# Iconicity control (n = 1,771)

PSC improves model fit also when controlling for iconicity (**β** = -0.295, se = 0.069, t = -4.293, p < .001; **Δ**AIC = 16.44).

PSC is also a more important variable than iconicity to predict AoA.

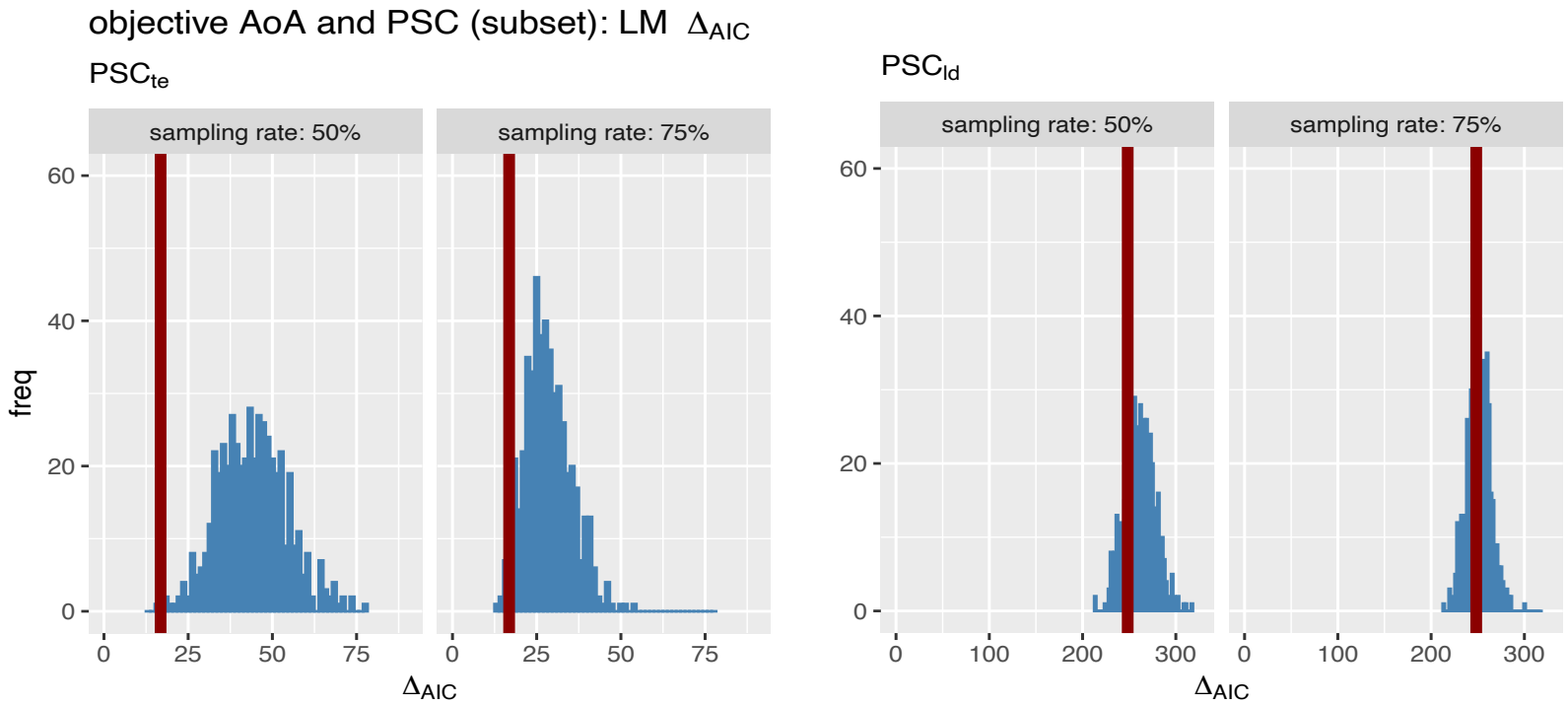


# Robustness check: random form-meaning




Non-zero correlation between AoA and $PSC_{ld}$ when computed on random scrambles of the vocabulary. But only true PSC explains additional variance!

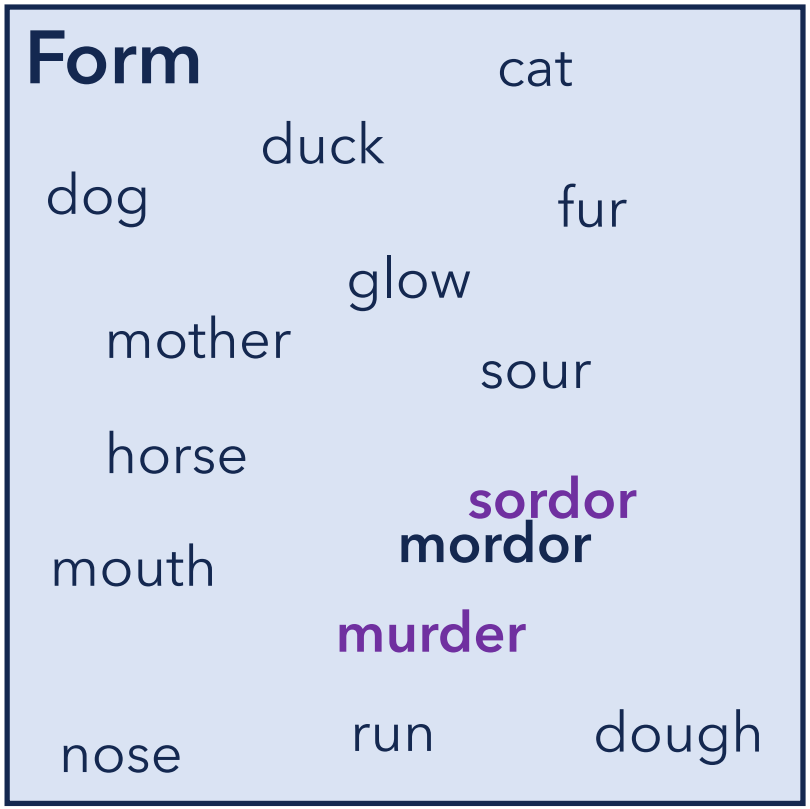
# Robustness check: vocabulary subsets

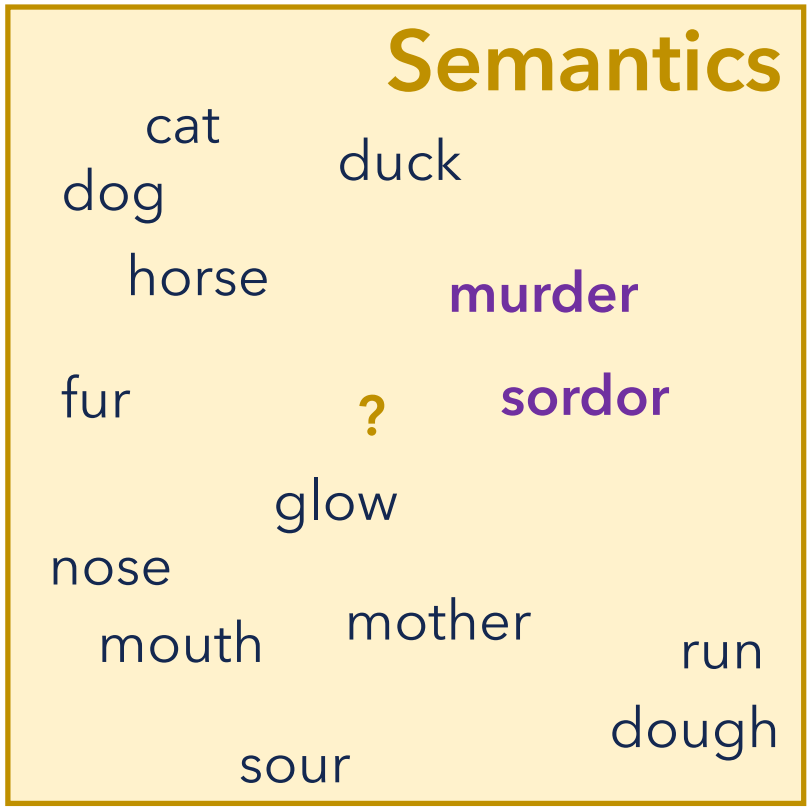



The improvement in model fit brought by PSC is similar if not stronger when sampling from the entire vocabulary: the effect is robust!

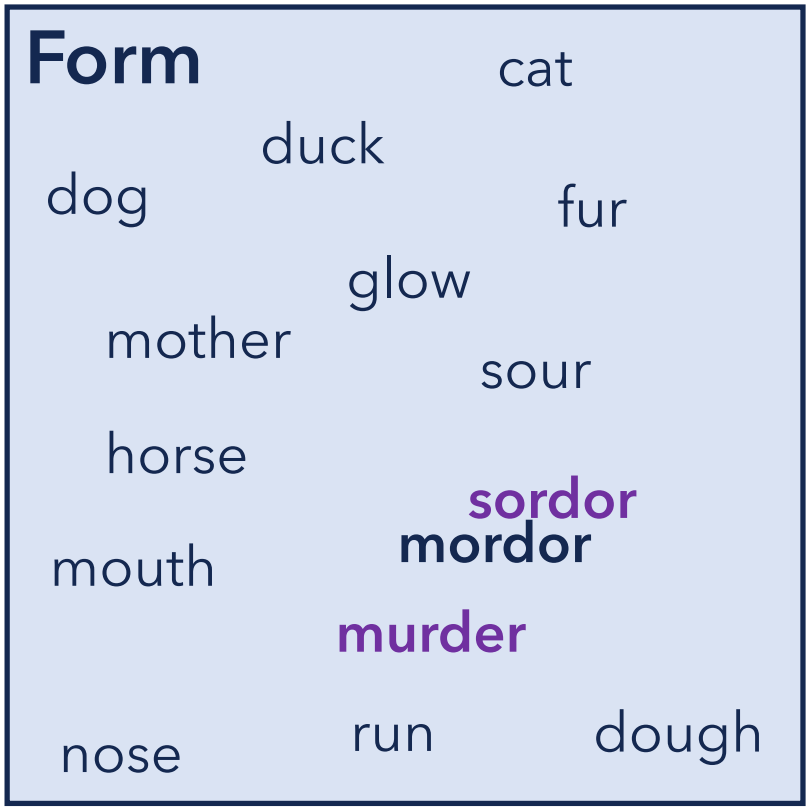
# What does it mean?

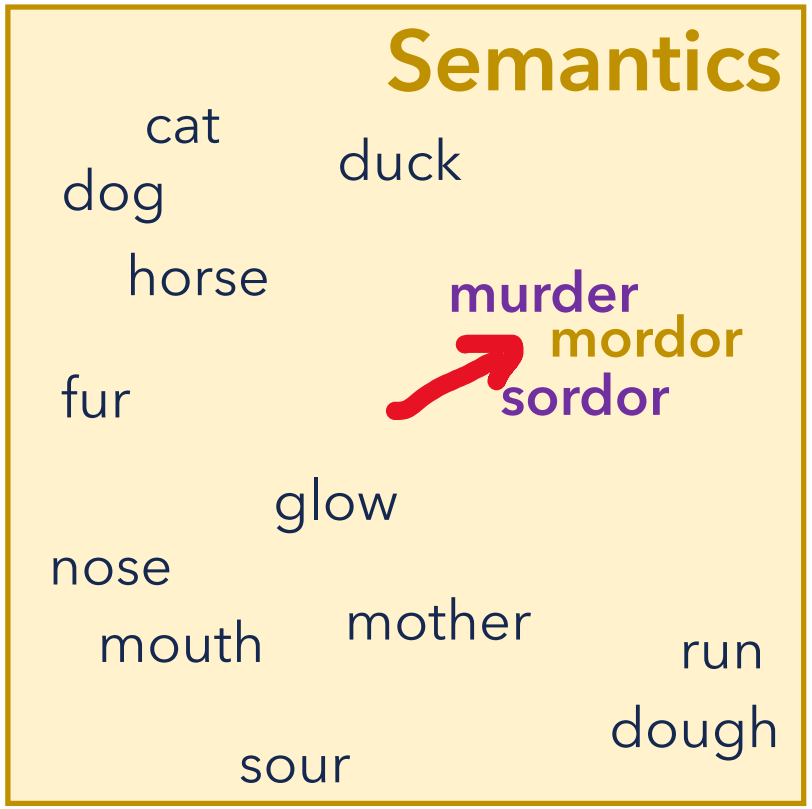
# How could FSC work in acquisition?



# How could FSC work in acquisition?



# How can this work?

Inferring meaning by analogizing over form similarities only works if learners can approximate lexical semantics at least most of the times, i.e., if the process works for many words.

For it to work there needs to be **some degree of form-meaning systematicity**.

# How does it fit in the literature?

FSC may be an account of how learners **exploit systematicity productively**, to learn novel words.

FSC doesn't only capture activation patterns in the mental lexicon. It also captures learning patterns: systematicity plays a role during learning **and** processing.

# A wilder hypothesis

Iconicity works on an **analogical process between word forms and perceptual semantics**, and it is hypothesized to contribute to establish the referential nature of language.

FSC could exploit a similar analogical process between word forms and lexical semantics, once learners know enough words to analogize.

# Caveats

Learners need to know some words for FSC to work: **FSC cannot anything about earliest language learning.**

We controlled for morphological complexity, but we didn't quantify **how much of FSC's effect is due to morphological productivity, etymology, or sound symbolism**.

No indications about **other languages**: is there a role for morphological complexity?

# Where do we go from here?

- How do learners compute form similarities? Is there a role for **phonotactics**?
- Is the role of FSC dependent on **vocabulary knowledge**?
- Is there a **time-course** for form-based v. context-based semantics during learning? Does **frequency** modulate it?
- Other methods to derive systematicity measures (e.g., LDL [Baayen et al, 2018])?
- Is FSC really a mapping function?

# Thank you!

g.cassani@tilburguniversity.edu

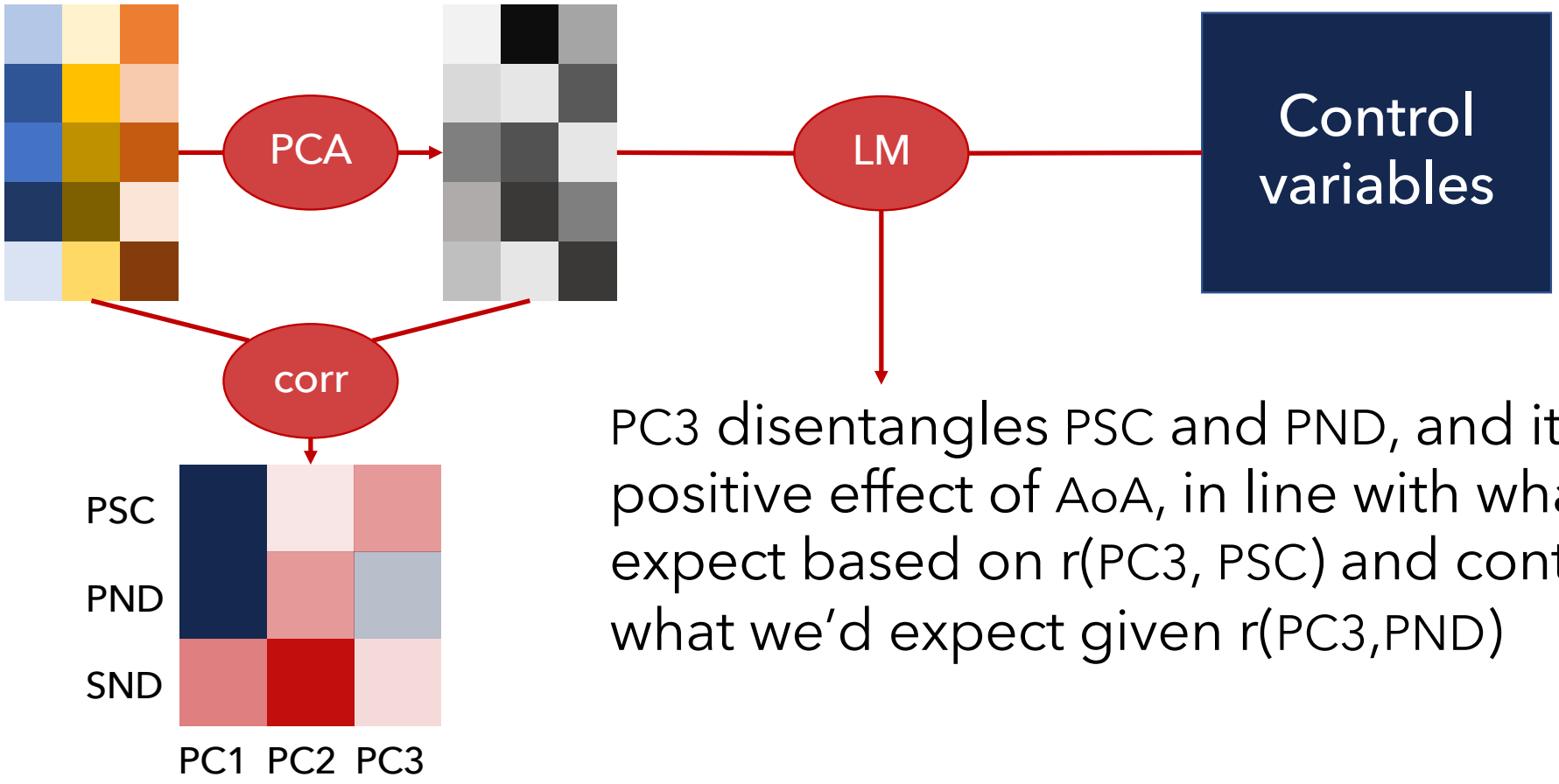
# Extras

# Robustness check: linearity

Same as main analysis but replace linear and random forest regression with **GAMs** to make sure effects are not just by-products of wrong assumptions.

Effect of PSC is robust!

# Robustness check: FSC or neighborhood?




PC3 disentangles PSC and PND, and it has a positive effect of AoA, in line with what we expect based on r(PC3, PSC) and contrary to what we'd expect given r(PC3,PND)

# Robustness check: OSC

Repeat the main analysis using OSC instead of PSC to ensure what was observed is not a by-product of the chosen encoding of word-form.

All patterns replicated, but PSC generally has stronger results.

# Robustness check: subjective AoA

Replace objective with subjective AoA norms to start checking whether the effect of PSC is specific to one way of construing AoA (vocabulary tests) or not.

The effect of PSC is replicated.