Dr Maria Korochkina

Learning affixes through text experience: A new theoretical and computational framework

CBU @ Cambridge
9 October 2025

Economic and Social Research Council

ROYAL HOLLOWAY UNIVERSITY OF LONDON

What is morpheme knowledge for?

- Most English words are built by **recombining stems and affixes**

cleaner, cleanly, unclean
teacher, banker, builder

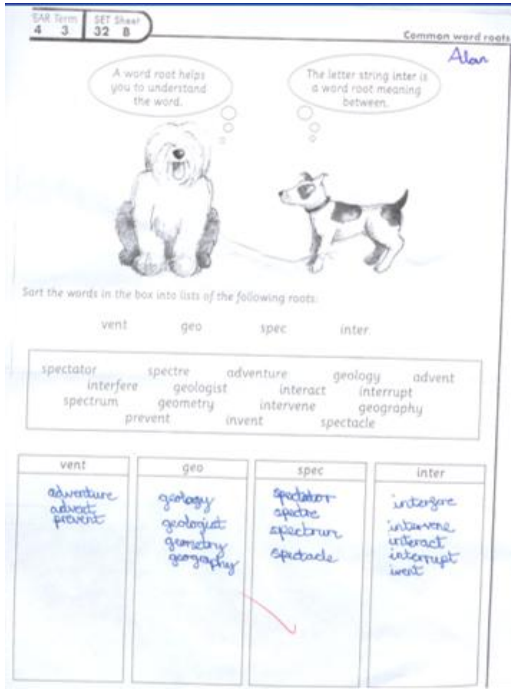
- Morpheme knowledge enables rapid access to the meanings of **familiar** words

- It is also crucial for computing the meanings of **unfamiliar** words

bright + -ify → brightify

How is morpheme knowledge acquired?

- **Limited time** for explicit instruction in school
- Teacher knowledge often **patchy**



- Form–meaning relationship more salient in written language
  - *bon<u>us</u>*, *atl<u>as</u>*, *serv<u>ice</u>*, *princ<u>ess</u>* vs. *hazard<u>ous</u>*

→ **Morpheme knowledge largely acquired via text experience**

Pre-requisites for morpheme learning

unknown	deactivate
unfair	decode
unafraid	decompose
unlikely	demand
unconvinced	deceive
unsure	depend
unwell	deliver (de- + -liberare)

- Must have **consistent meaning** transformation
- Must occur with a **high number of distinct stems** (type frequency)
- Must be **detectable**

Tamminen et al., 2015, *Cogn Psychol*
Ulicheva et al., 2020, *Cognition*
Korochkina & Rastle, 2025, *npj Sci Learn*

What’s children’s experience of morphology like in the wild?

A corpus linguistics approach

The **C**hildren & **Y**oung **P**eople's Books **Lex**icon

7-9 years

10-12 years

13+ years

- 1,200 popular books, 400 books per age band
- Over 70 mln words & over 105,000 distinct words

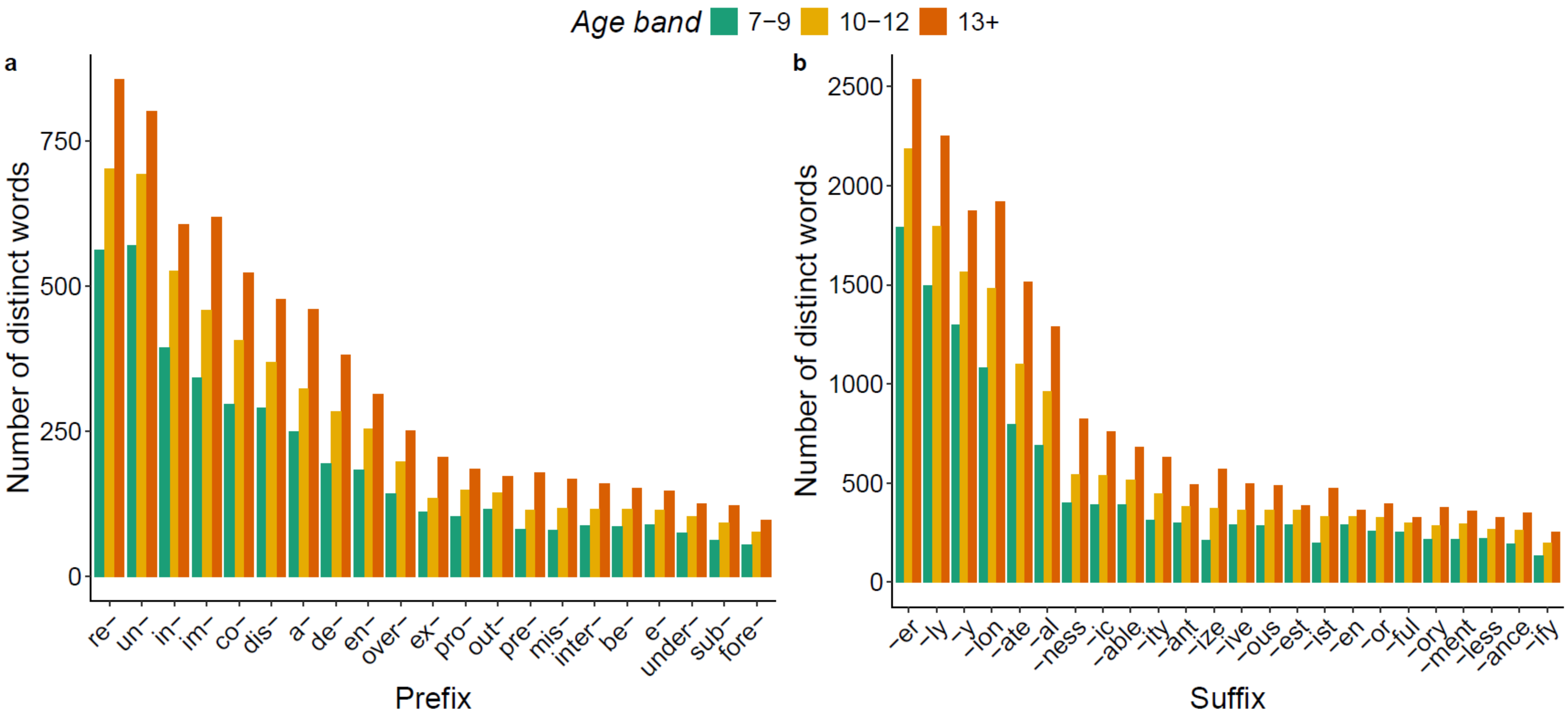
Korochkina et al., 2024, *QJEP*

Morphology in children's books

- Roughly **half of all distinct words** are complex
- **Few** complex words are **used repeatedly** or in many books
- Children are **likely to see** a complex word but **unlikely to see** this word **again**

- Only a **few affixes** have reasonably **high type frequency before 13+** texts

Korochkina & Rastle, 2025, *npj Sci Learn*



Korochkina & Rastle, 2025, *npj Sci Learn*

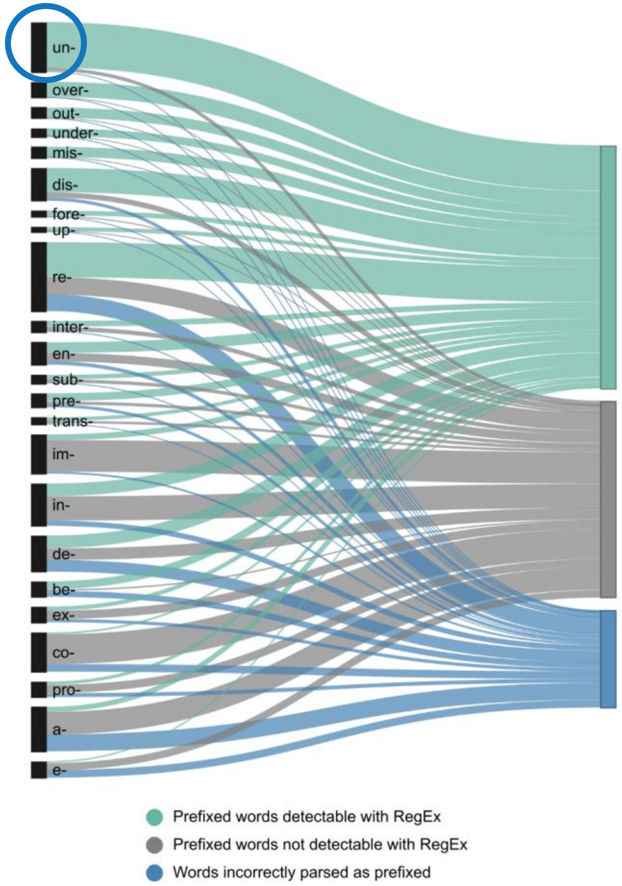
Morphology in children’s books

- Roughly **half of all distinct words** are complex
- **Few** complex words are **used repeatedly** or in many books
- Children are **likely to see** a complex word but **unlikely to see** this word **again**

- Only a **few affixes** have reasonably **high type frequency before 13+** texts
- Many **affixes difficult to detect**

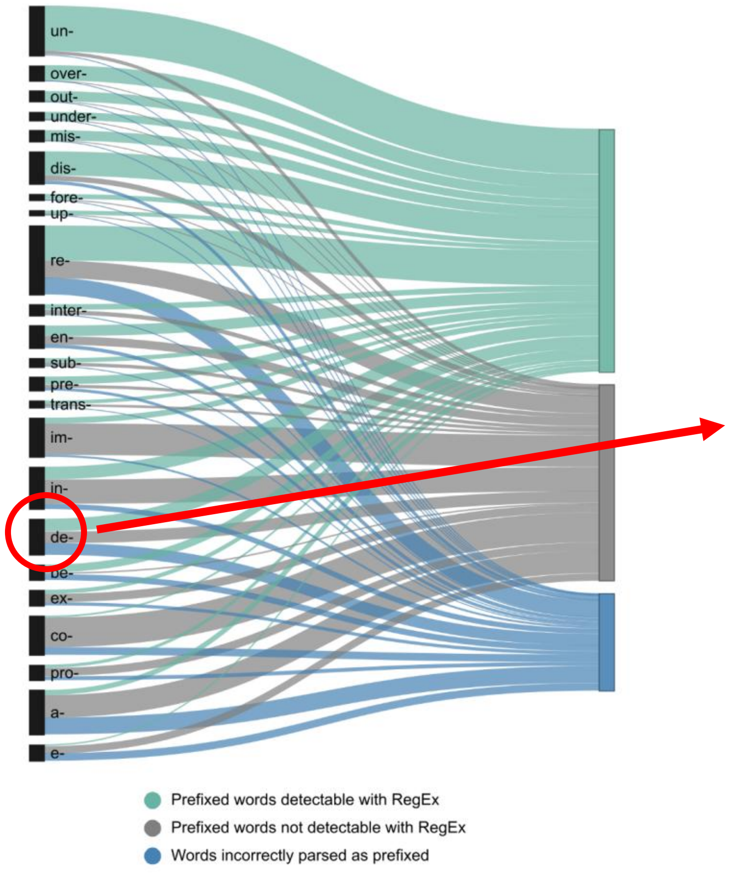
Korochkina & Rastle, 2025, *npj Sci Learn*

Many affixes are difficult to detect



Korochkina & Rastle, 2025, *npj Sci Learn*

Many affixes are difficult to detect



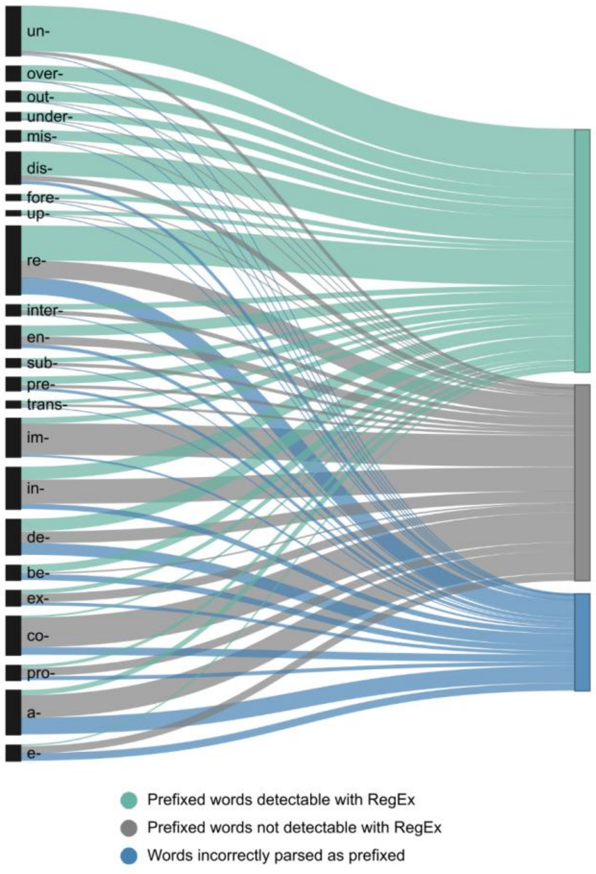
1/3 detectable
deactivate, decode, decompose

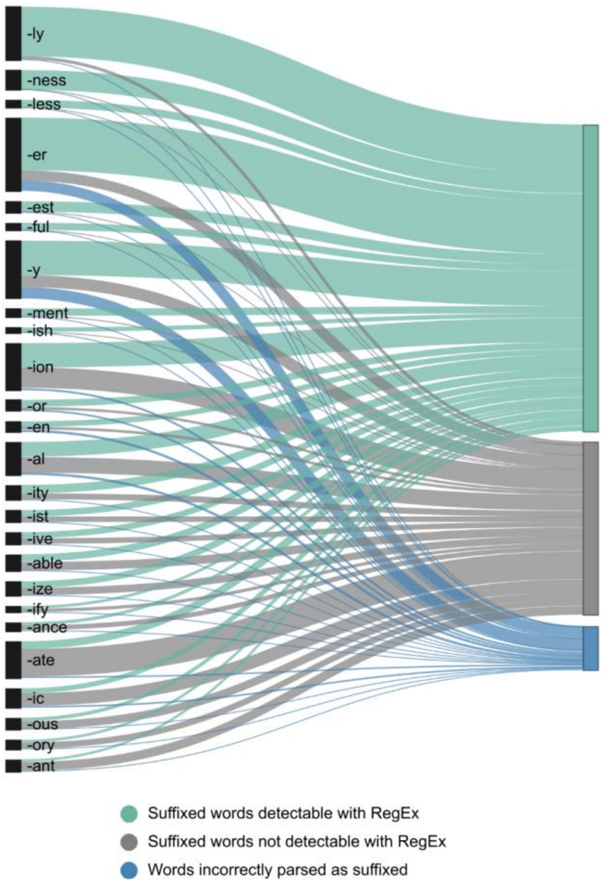
1/3 undetectable
demand, deceive, depend

1/3 false alarms
deliver, detail, defeat

Korochkina & Rastle, 2025, *npj Sci Learn*

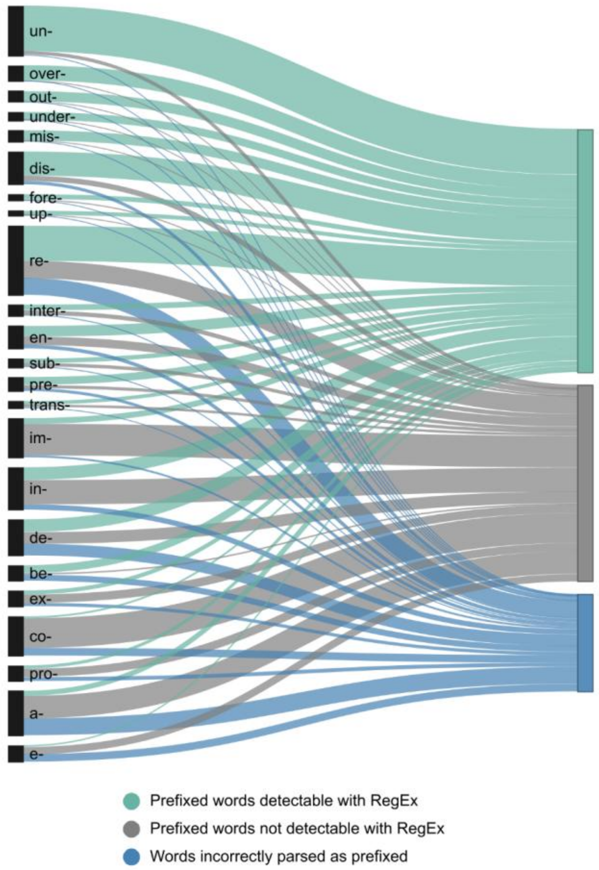
Many affixes are difficult to detect

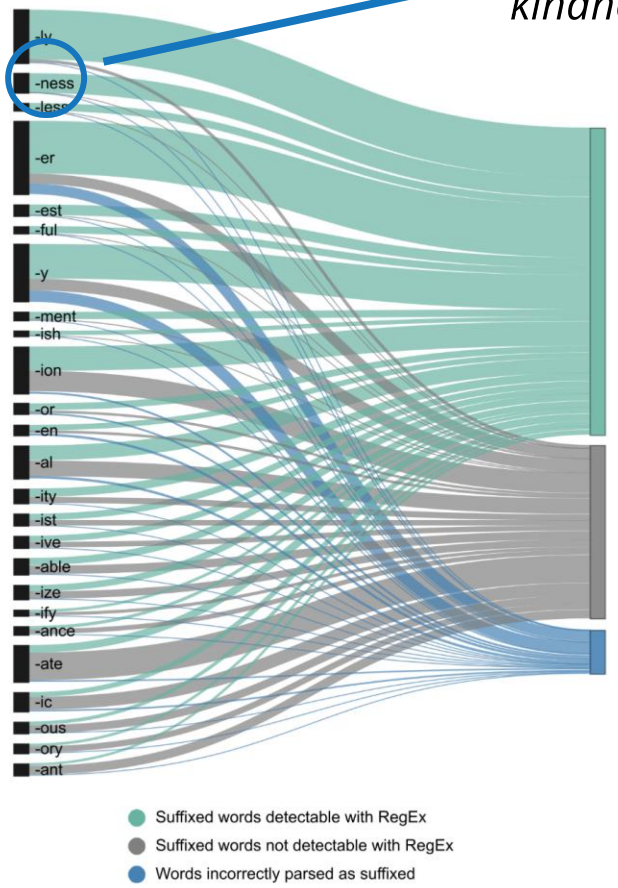




Korochkina & Rastle, 2025, *npj Sci Learn*

Many affixes are difficult to detect

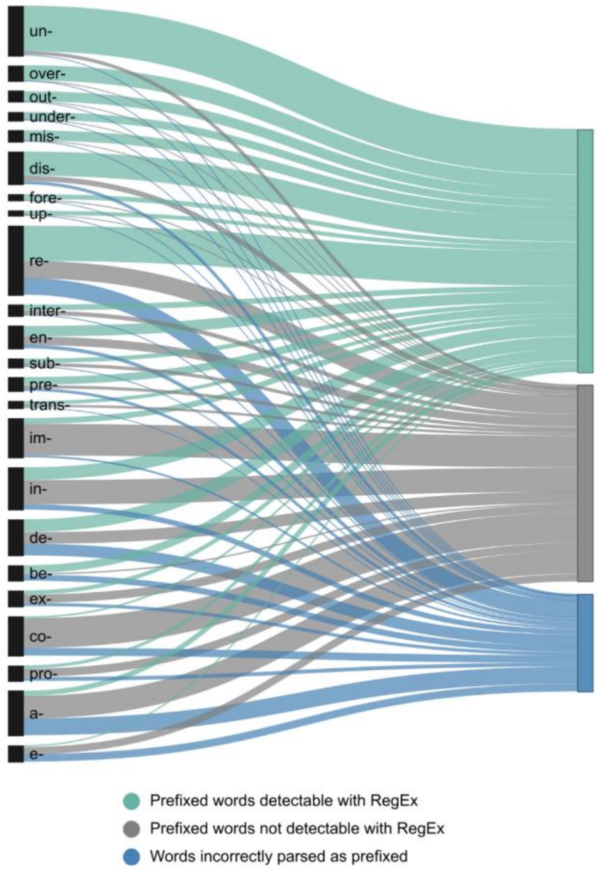


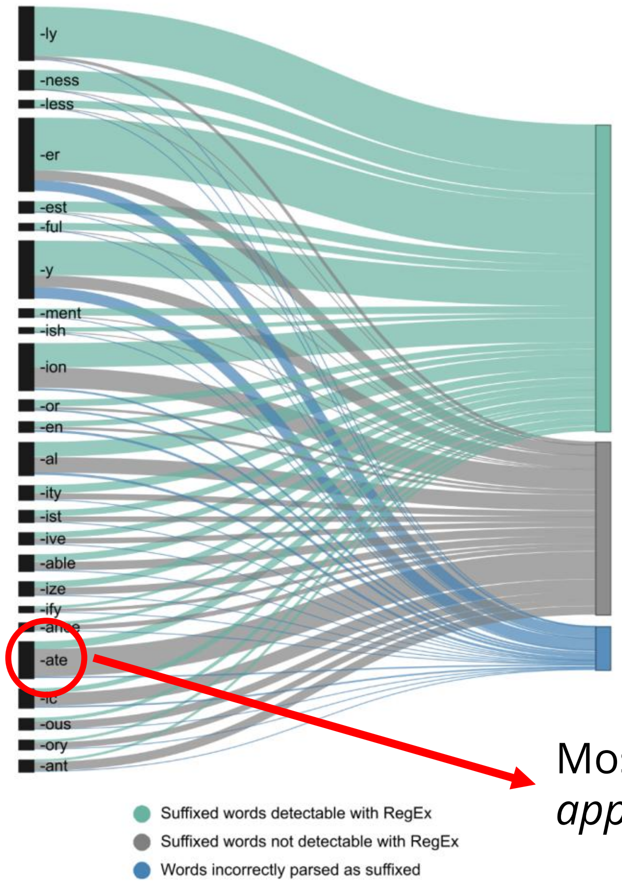


Easy to detect
kindness, weakness, sadness

Korochkina & Rastle, 2025, *npj Sci Learn*

Many affixes are difficult to detect

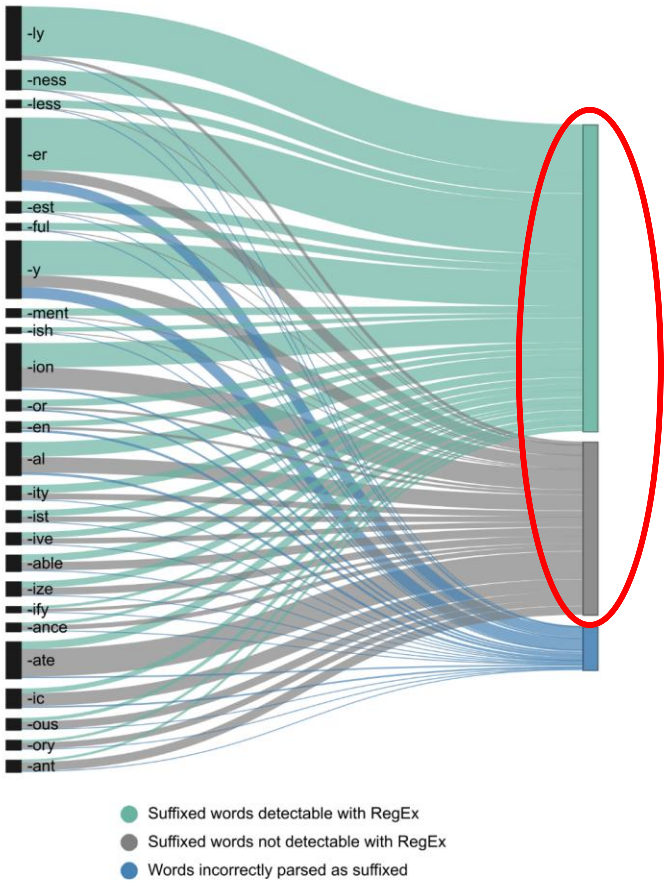




Mostly undetectable
appreciate, *generate*, *integrate*

Korochkina & Rastle, 2025, *npj Sci Learn*

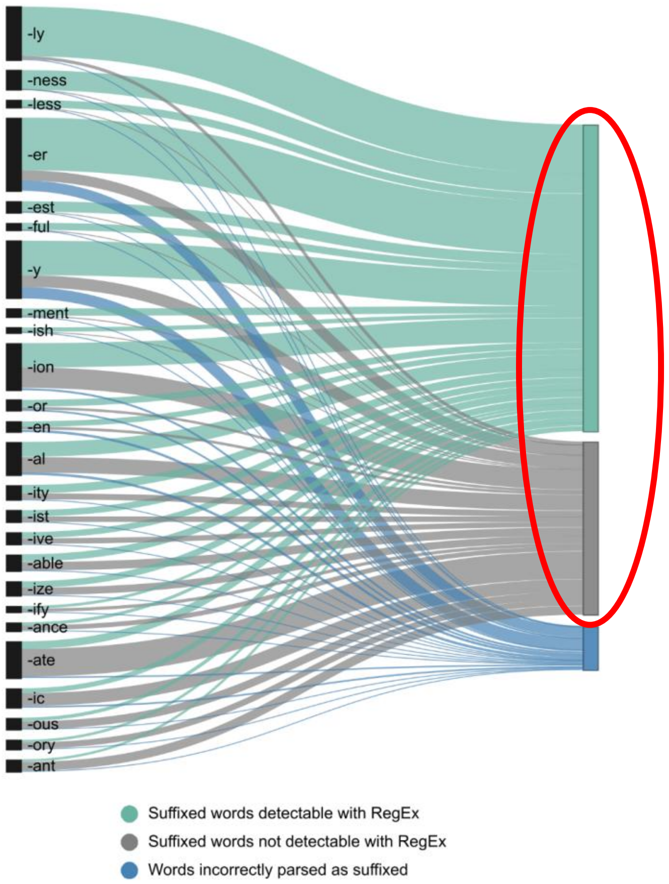
What constitutes morpheme experience?



1. All instances where a complex-looking word is **historically formed through derivation**

Korochkina et al., *In press*

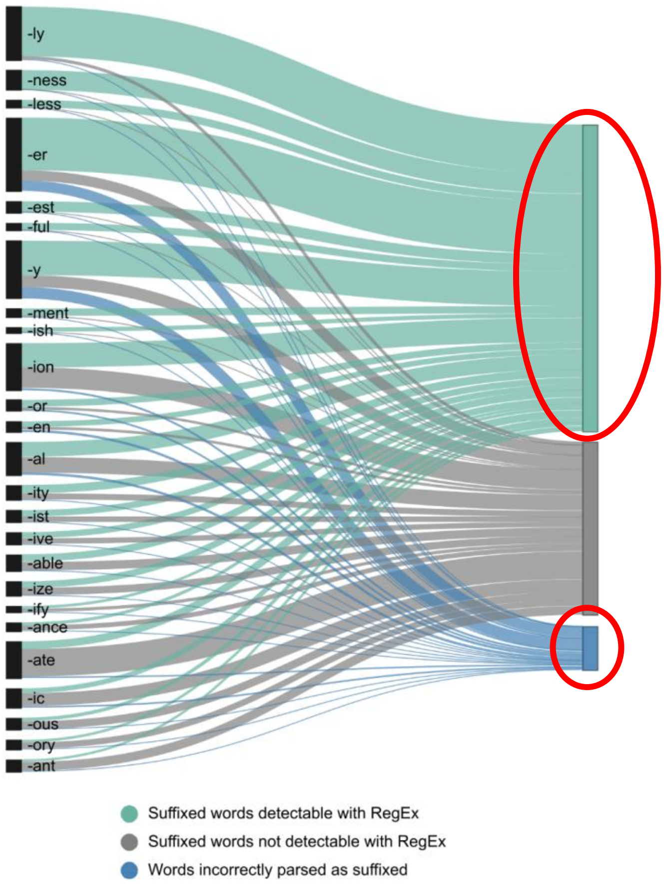
What constitutes morpheme experience?



1. All instances where a complex-looking word is **historically formed through derivation**
   dictionary-based type frequency

Korochkina et al., *In press*

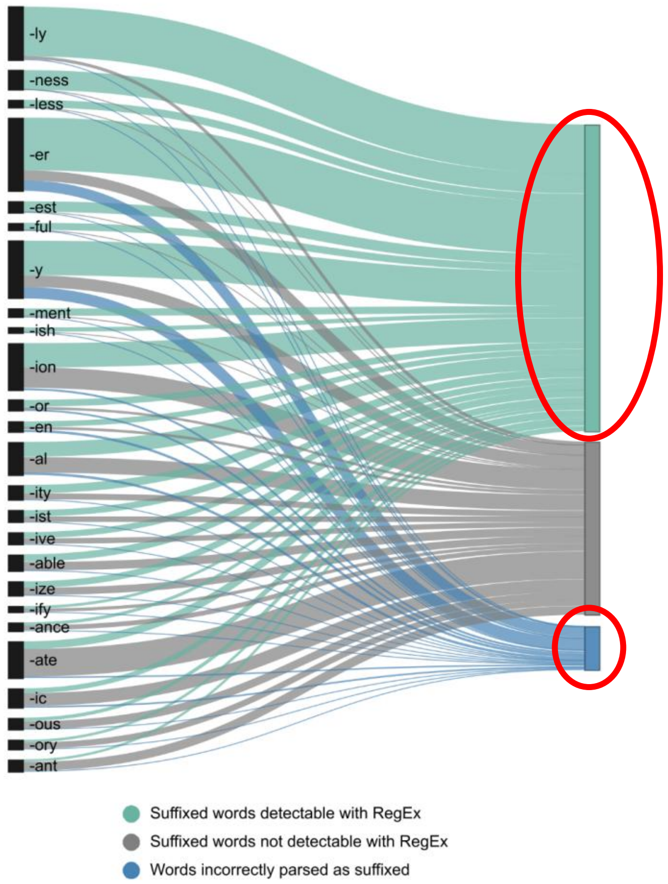
What constitutes morpheme experience?



1. All instances where a complex-looking word is **historically formed through derivation** *dictionary-based type frequency*
2. All instances where affixes are **identifiable without specialised knowledge**

Korochkina et al., *In press*

What constitutes morpheme experience?

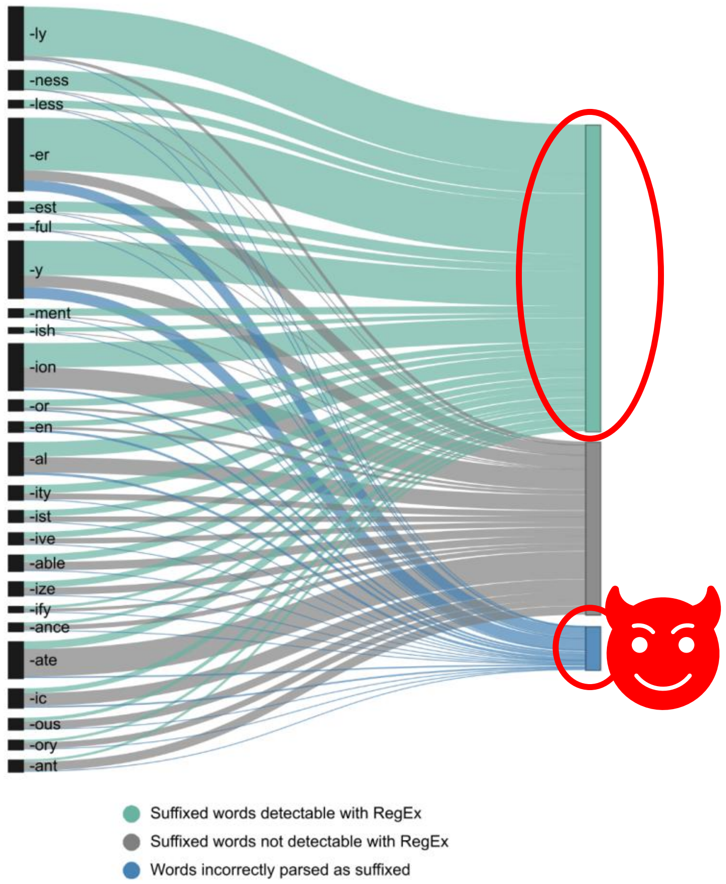


1. All instances where a complex-looking word is **historically formed through derivation**
   dictionary-based type frequency

2. All instances where affixes are **identifiable without specialised knowledge**
   orthography-based type frequency

Korochkina et al., *In press*

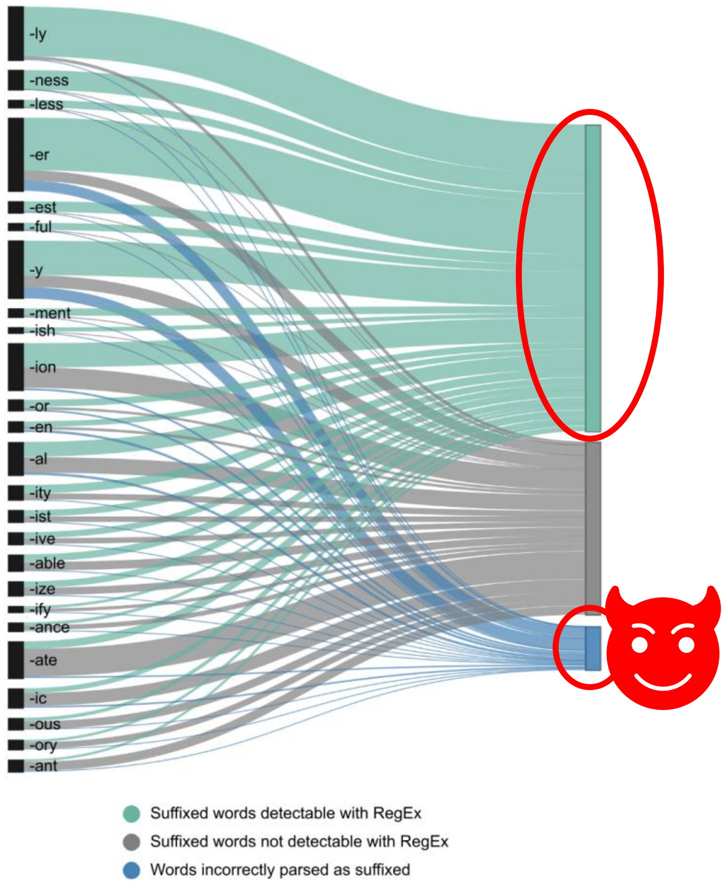
What constitutes morpheme experience?



1. All instances where a complex-looking word is **historically formed through derivation**
   dictionary-based type frequency
2. All instances where affixes are **identifiable without specialised knowledge**
   orthography-based type frequency
3. All instances where affixes are identifiable, but **false alarms incur a learning penalty**

Korochkina et al., *In press*

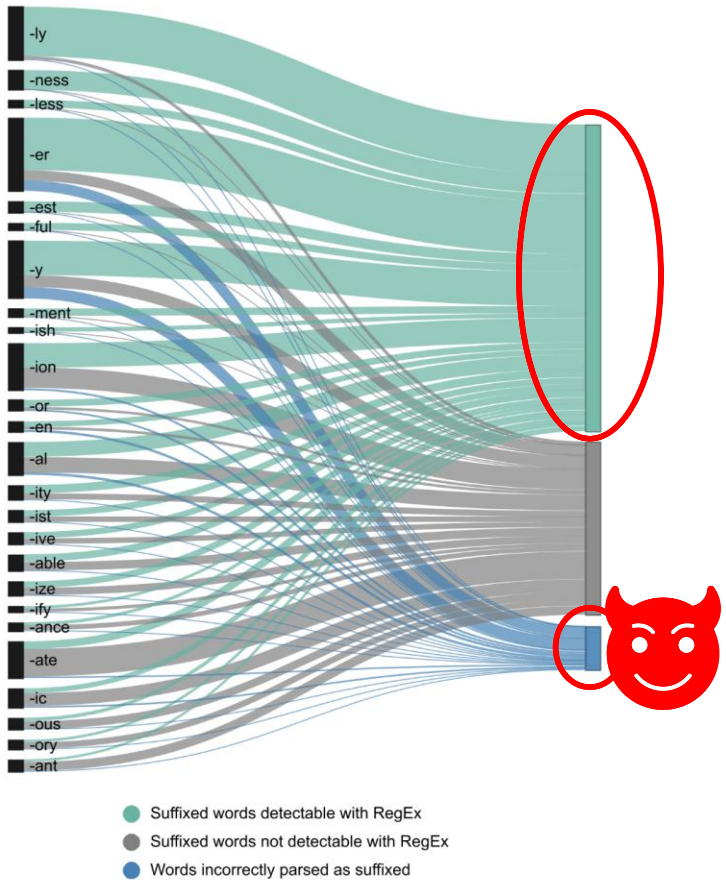
What constitutes morpheme experience?



1. All instances where a complex-looking word is **historically formed through derivation**
   dictionary-based type frequency
2. All instances where affixes are **identifiable without specialised knowledge**
   orthography-based type frequency
3. All instances where affixes are identifiable, but **false alarms incur a learning penalty**
   orthography-based type frequency + false alarm penalty

Korochkina et al., *In press*

The false alarm penalty

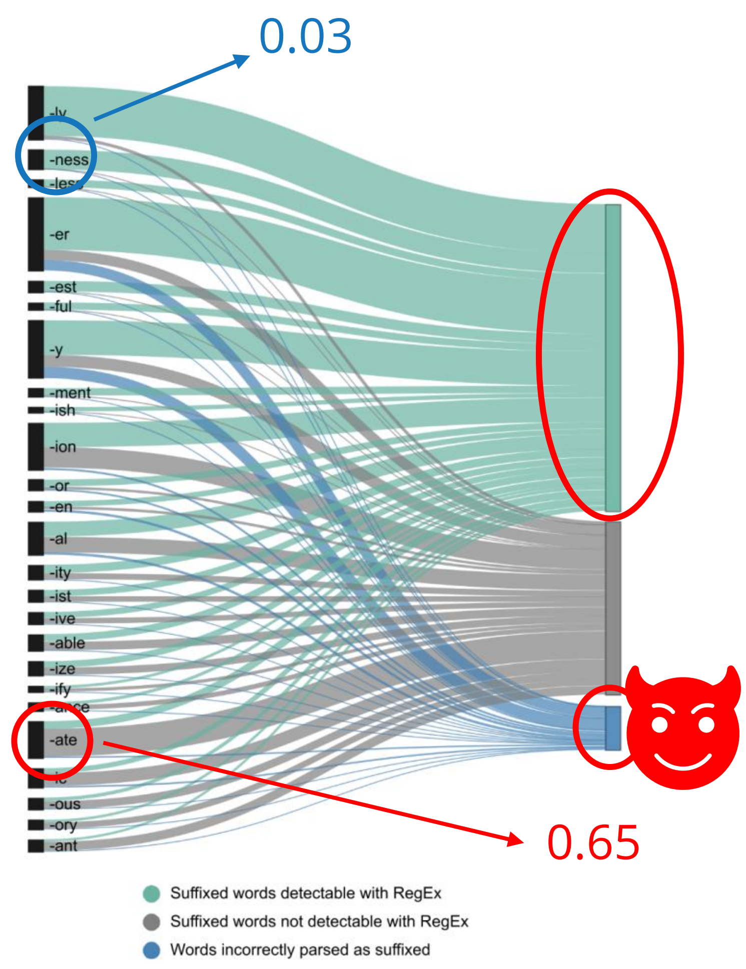


Shannon entropy

Quantifies the **uncertainty about the function** of the orthographic pattern associated with an affix

Shannon, 1948, *Bell Syst Tech J*

The false alarm penalty



Shannon entropy

Quantifies the **uncertainty about the function** of the orthographic pattern associated with an affix

Low entropy → little uncertainty → low penalty

High entropy → more uncertainty → high penalty

Korochkina et al., *In press*

Theories in action

Which definition best explains human behaviour?

The morpheme interference effect

woodness

word | not a word

woodnels

word | not a word

- Morphologically-structured nonwords are more difficult, and take longer, to reject
- Skilled readers segment complex-looking words into morphemes

Taft & Forster, 1975, *J of Verb Learn & Verb Behav*
Crepaldi et al., 2010, *Mem & Cogn*

Stimuli

- 6 prefixes
  - *un-*, *mis-*, *dis-*, *pre-*, *de-*, *re-*
- 6 suffixes
  - *-ness*, *-ly*, *-able*, *-er*, *-ic*, *-ate*

- Morphologically structured nonwords
  - *unwood*, *woodness*
- Nonwords without morphological structure
  - *ubwood*, *woodnels*

- Each participant saw…
  - Each affix with 10 stems (120 morphologically structured nonwords)
  - Orthographic controls (120 nonwords with no morphological structure)
  - 120 morphologically complex + 120 morphologically simple words

Korochkina et al., *In press*

Stimuli

- 6 prefixes
  - *un-, mis-, dis-, pre-, de-, re-*
- 6 suffixes
  - *-ness, -ly, -able, -er, -ic, -ate*
- Morphologically structured nonwords
  - *<u>un</u>wood, wood<u>ness</u>*
- Nonwords without morphological structure
  - *<u>ub</u>wood, wood<u>nels</u>*

- Each participant saw 480 letter strings
  - Each affix with 10 stems (120 morphologically structured nonwords)
  - Orthographic controls (120 nonwords with no morphological structure)
  - 120 morphologically complex + 120 morphologically simple words

Korochkina et al., *In press*

Participants

120 participants

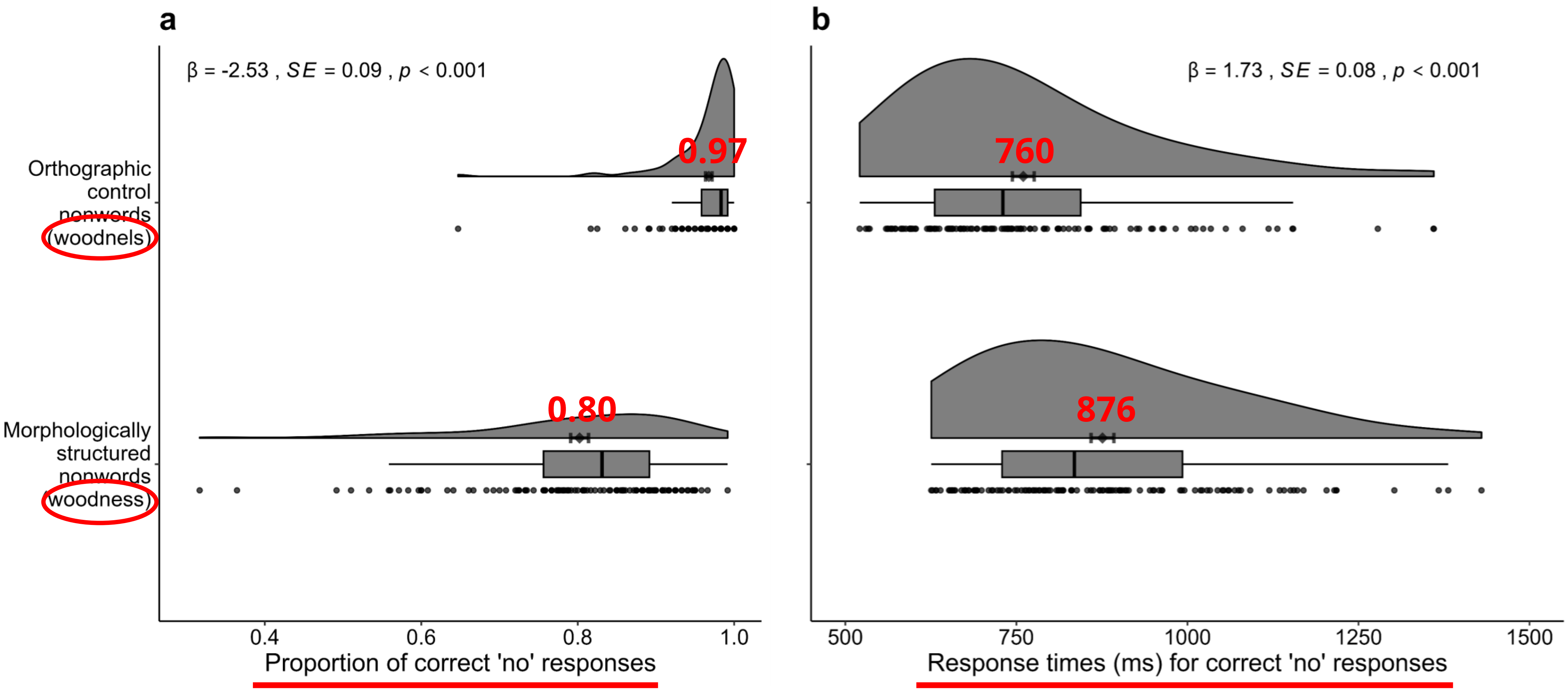
18 – 40 years old

63 female
56 male
1 non-binary

UK based
English as a first language
No language disorders

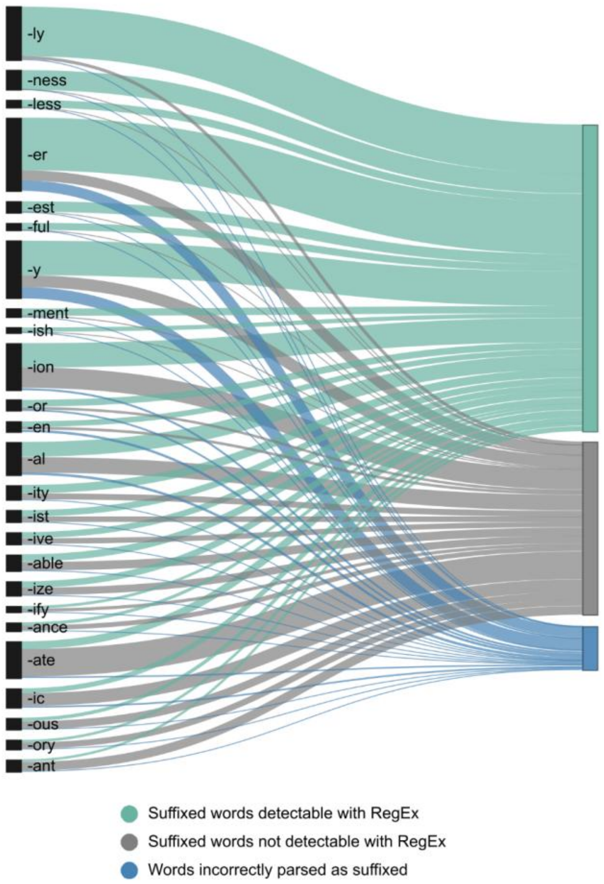
Korochkina et al., *In press*

Readers are sensitive to morphological structure



Korochkina et al., *In press*

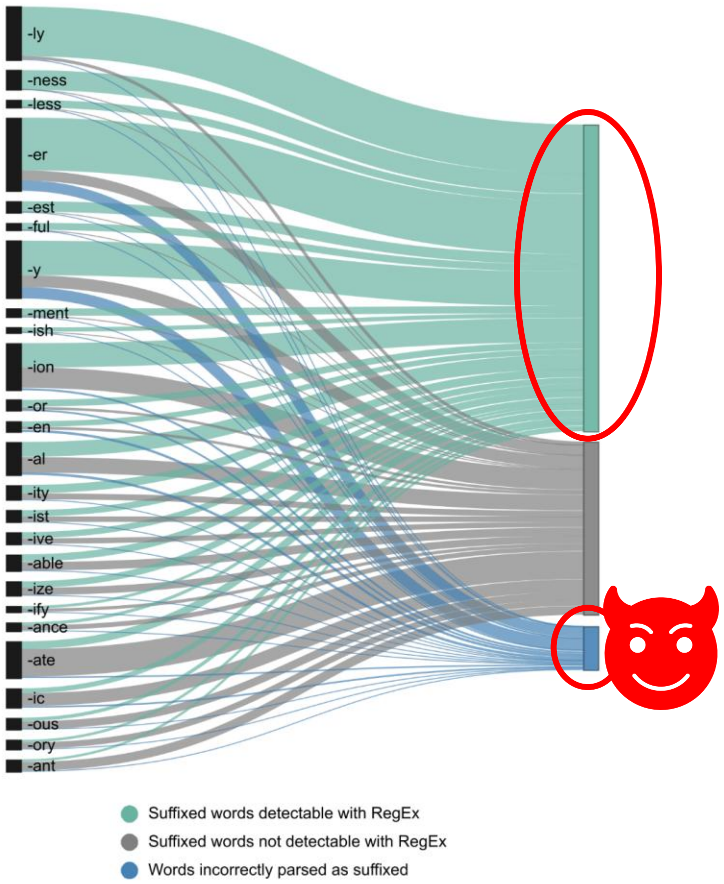
What constitutes morpheme experience?



1. All instances where a complex-looking word is **historically formed through derivation** *dictionary-based type frequency*
2. All instances where affixes are **identifiable without specialised knowledge** *orthography-based type frequency*
3. All instances where affixes are identifiable, but **false alarms incur a learning penalty** *orthography-based type frequency + false alarm penalty*

Korochkina et al., *In press*

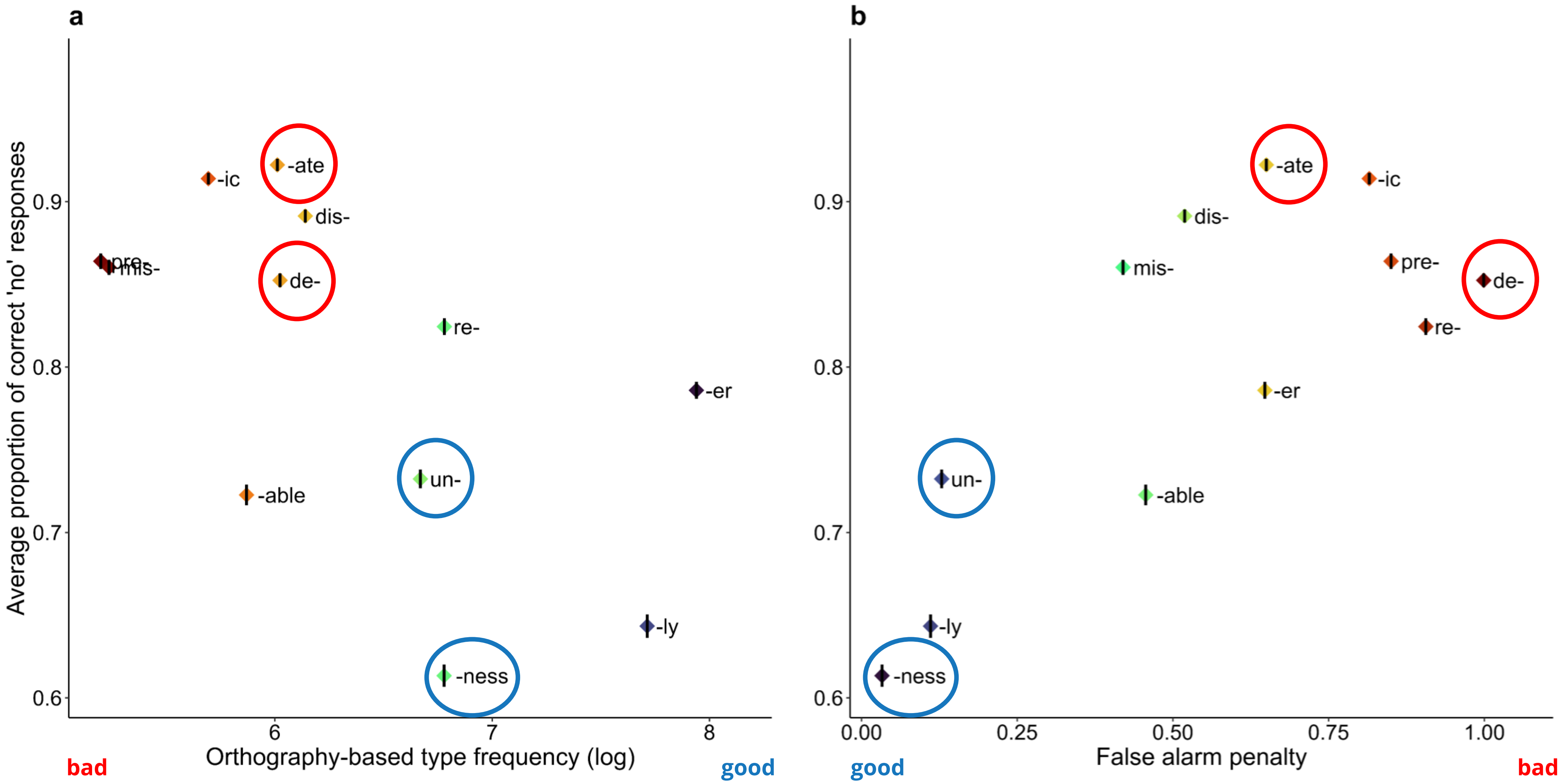
Theory 3 explains data best!



1. All instances where a complex-looking word is **historically formed through derivation**
   dictionary-based type frequency
2. All instances where affixes are **identifiable without specialised knowledge**
   orthography-based type frequency
3. All instances where affixes are identifiable, but **false alarms incur a learning penalty**
   orthography-based type frequency + false alarm penalty

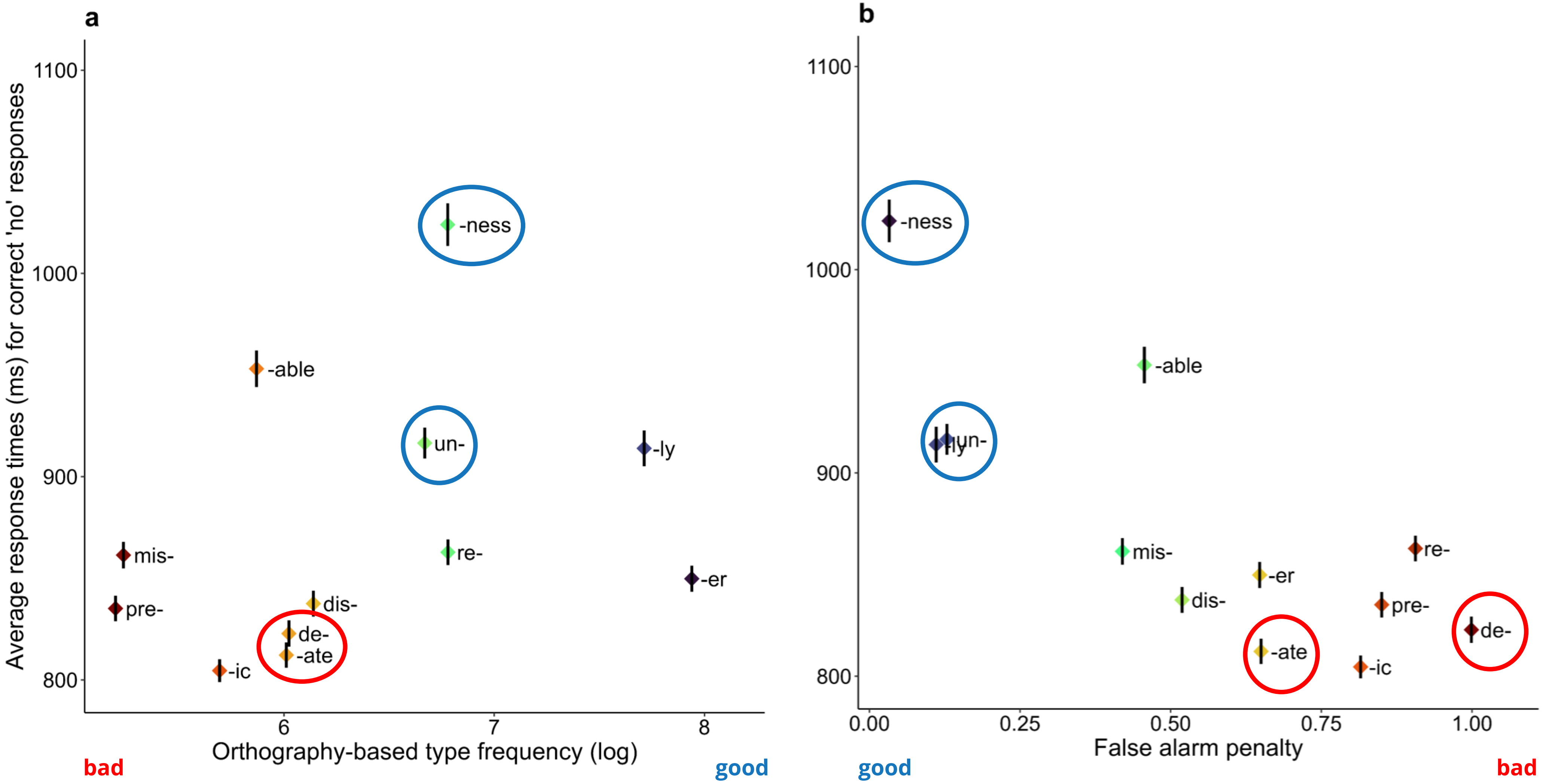
Korochkina et al., *In press*

Nonwords with “good” affixes are hard to reject…



Korochkina et al., *In press*

… and these rejections take time



Korochkina et al., *In press*

Interim summary

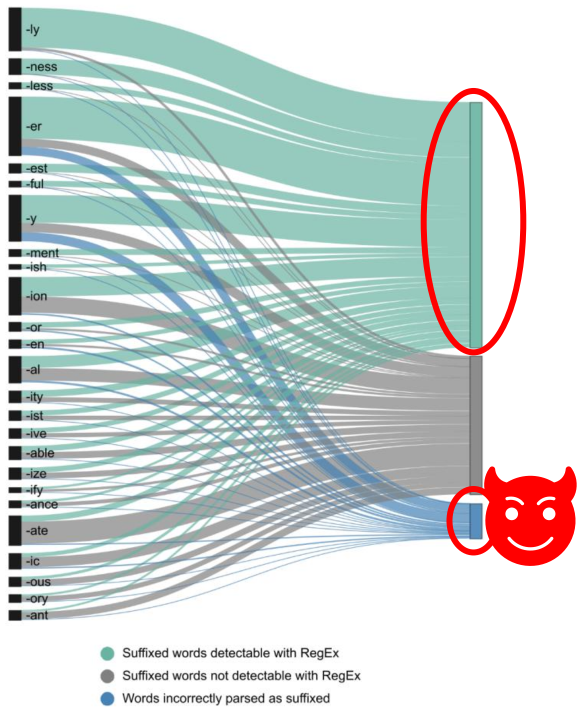
Quantified morpheme **experience** in print

↓

Proposed a new definition of morpheme experience

↓

Tested this definition against human data



- Critical step toward a **psychologically valid theory** of morpheme learning
- However, this approach is still a workaround: needs expert input and reduces affix meaning to a binary distinction

Modelling affix learning

… through compositional distributional semantic models

Distributional semantics

- A word's meaning can be inferred from contexts in which it appears
  - Similar contexts → similar meanings
  - Distinct contexts → more divergent meanings

boat

ship

Distributional semantics

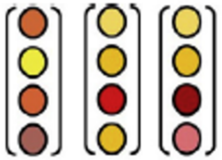
- A word's meaning can be inferred from contexts in which it appears
  - Similar contexts → similar meanings
  - Distinct contexts → more divergent meanings

	water	**passenger**	**sea**
boat			
ship			

Distributional semantics

- A word's meaning can be inferred from contexts in which it appears
  - Similar contexts → similar meanings
  - Distinct contexts → more divergent meanings

	water	passenger	sea
boat	23	15	40
ship	25	20	50

- Co-occurrence matrix (e.g., LSA) / neural embeddings (e.g., word2vec) → vector 
- Collection of vectors for a large number of words – **semantic space** 

Compositional distributional semantics
CAOSS

snowman

Marelli et al., 2017, *Cognition*

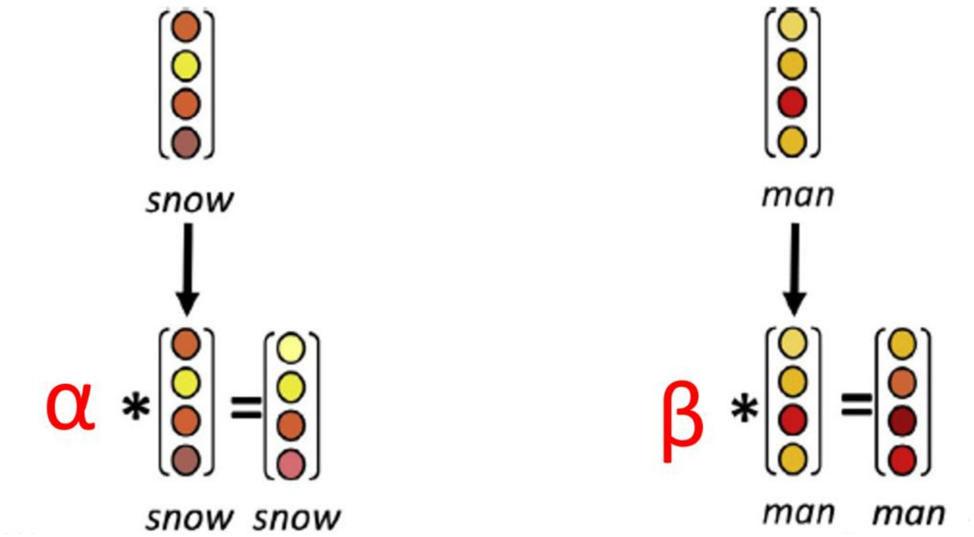
Compositional distributional semantics
CAOSS

snowman

Marelli et al., 2017, *Cognition*

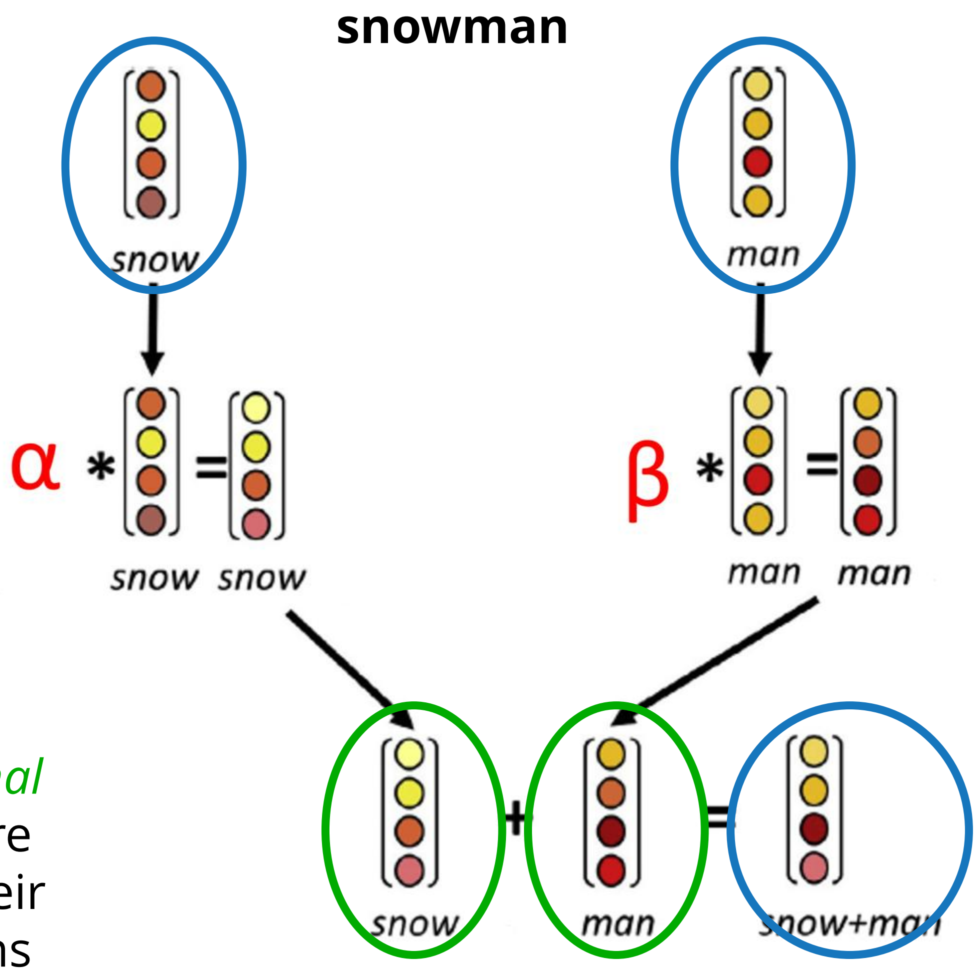
Compositional distributional semantics
CAOSS

snowman



Marelli et al., 2017, *Cognition*

Compositional distributional semantics
CAOSS

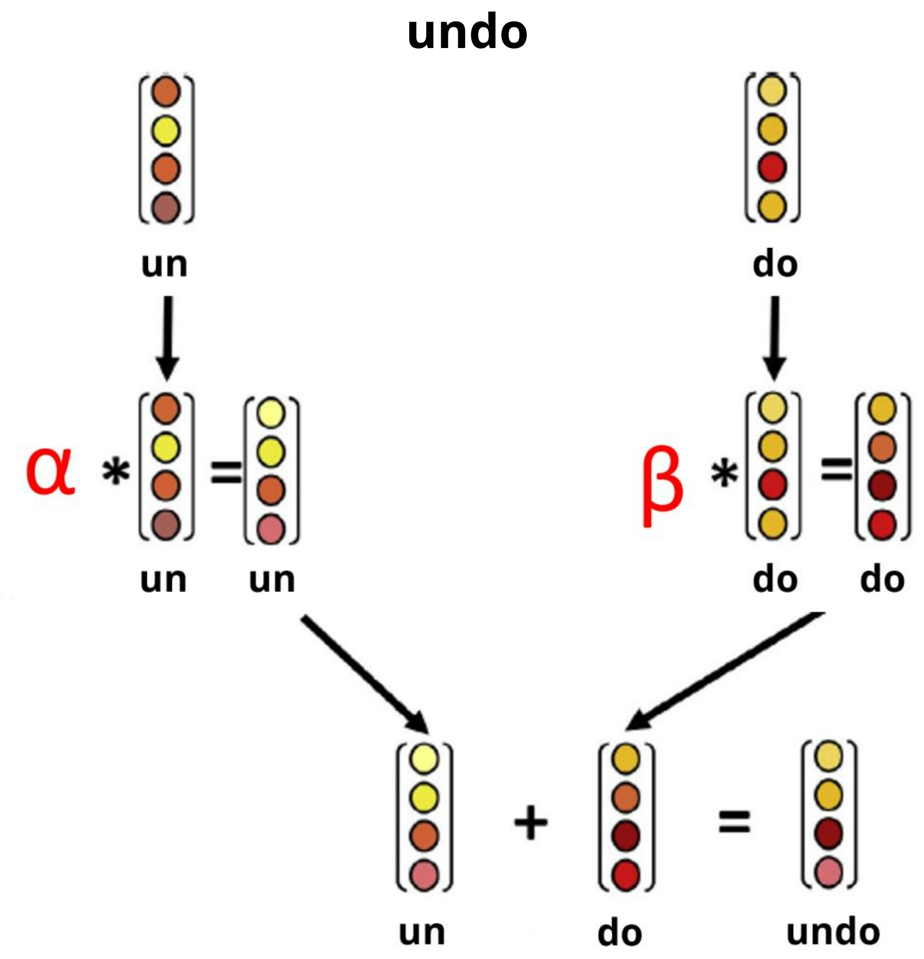


Model receives *lexicalised* representations of stems and compounds

Model learns the *transformation matrices* such that the *compositional* vectors of compounds are as close as possible to their *lexicalised* representations

Marelli et al., 2017, *Cognition*

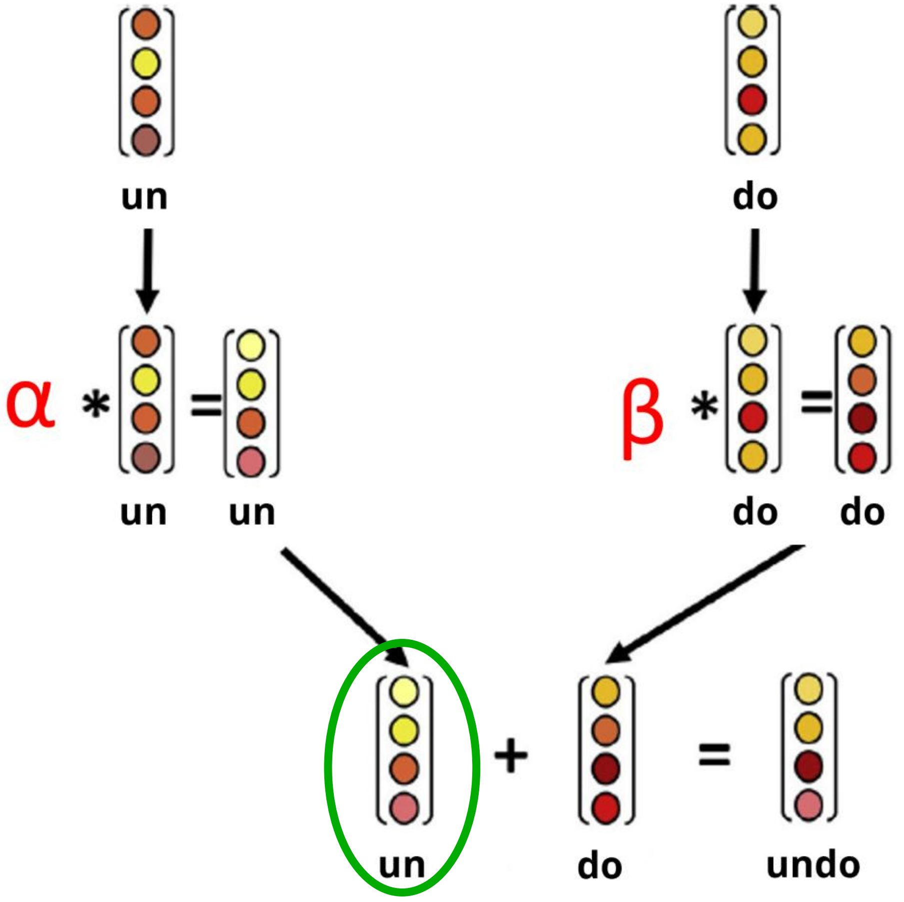
CAOSS applied to affixation



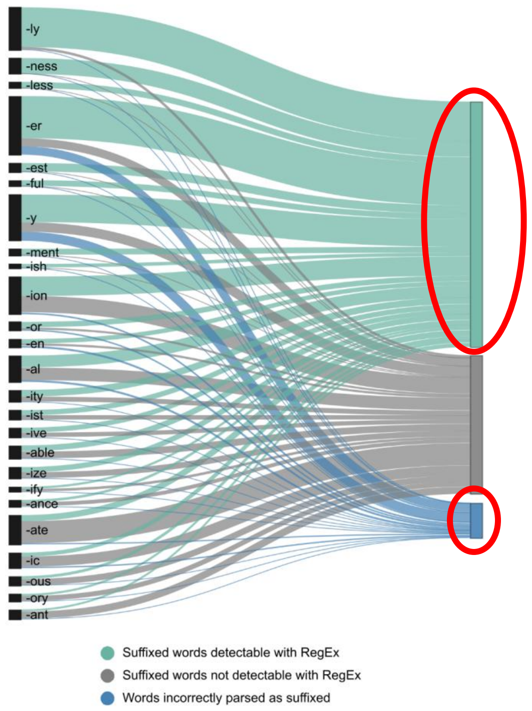
Our modelling approach

Lexicalised representations
for words & affixed words from *subs2vec* (van Paridon & Thompson, 2021)

Affix representations
Average of vector representations of all words with this affix (Westbury & Hollis, 2019)



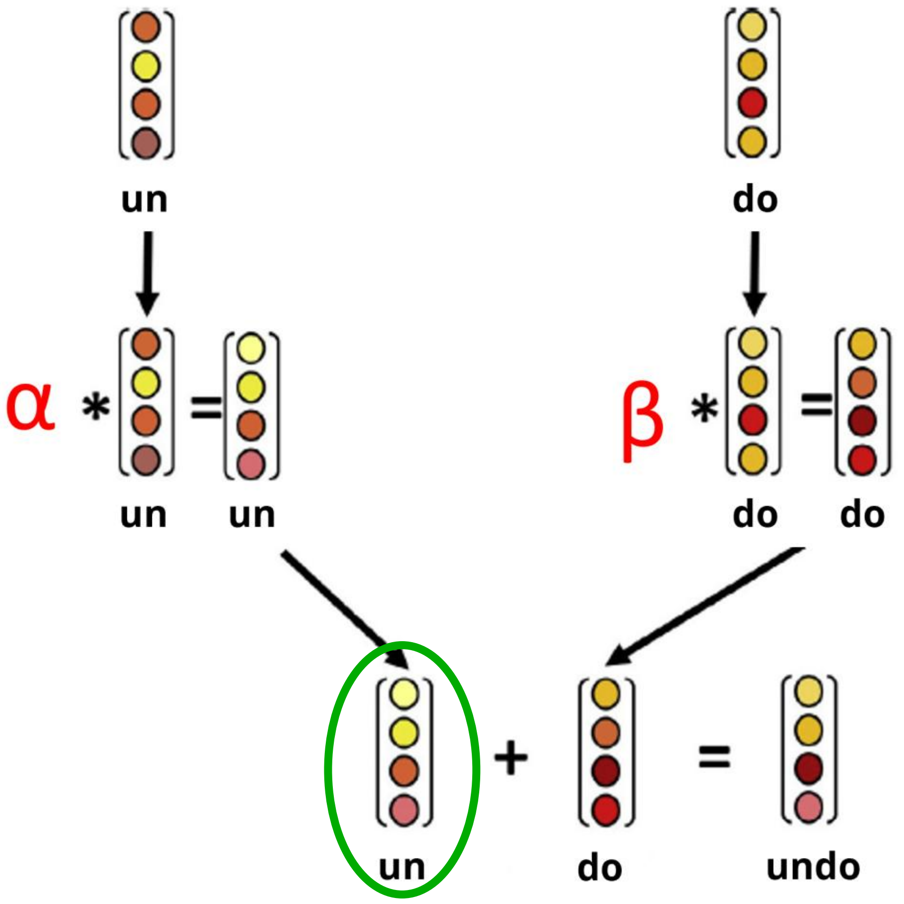
Training set



CAOSS metrics

Does model knowledge of affixes account for patterns in human lexical processing?

Morpheme interference data from Korochkina et al., *In press*



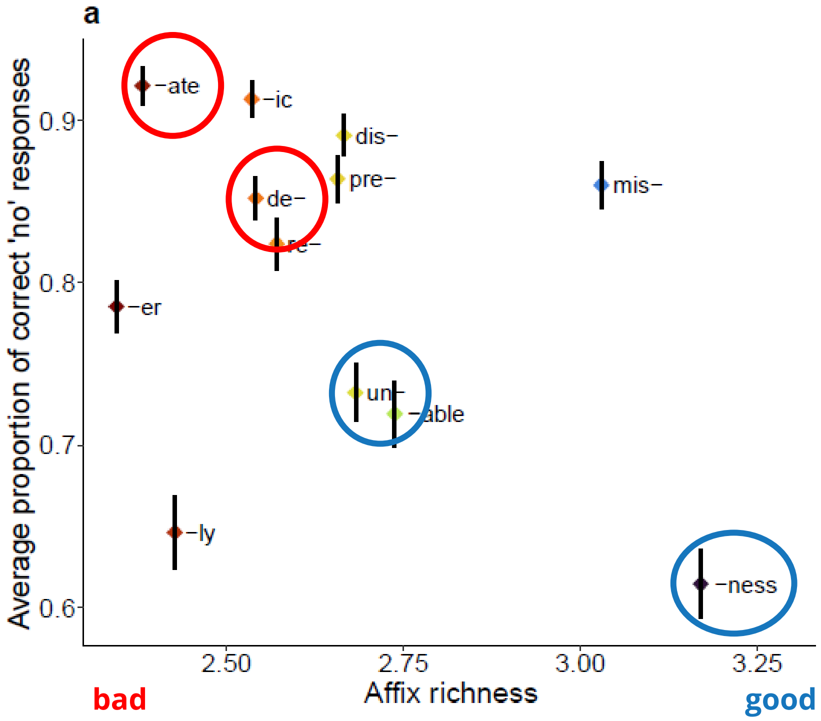
Affix diffuseness
Degree of diffusion and uncertainty in the affix meaning

Affix richness
Richness and complexity of affix meaning

Affix coherence
Similarity between the meaning of the affix and the meanings of words that contain it

More errors when rejecting nonwords with affixes with…

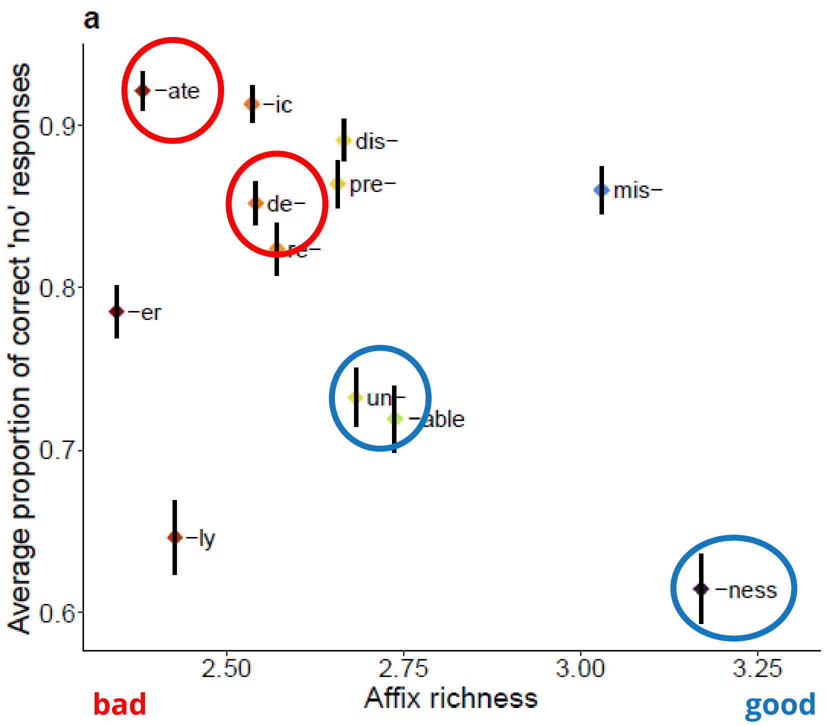
richer

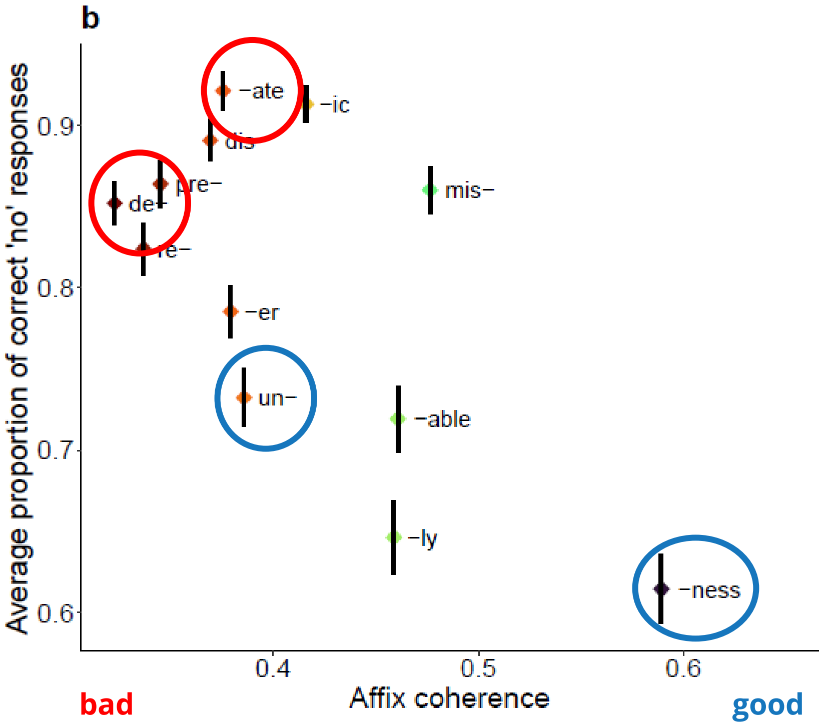


More errors when rejecting nonwords with affixes with…

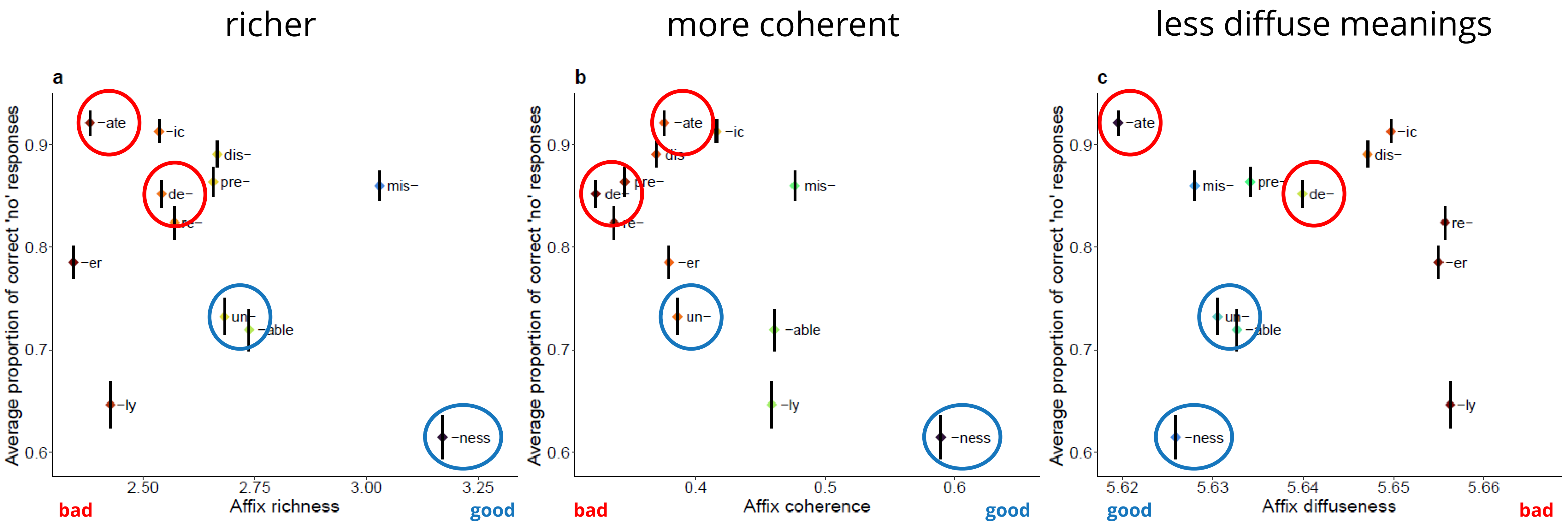
richer

more coherent

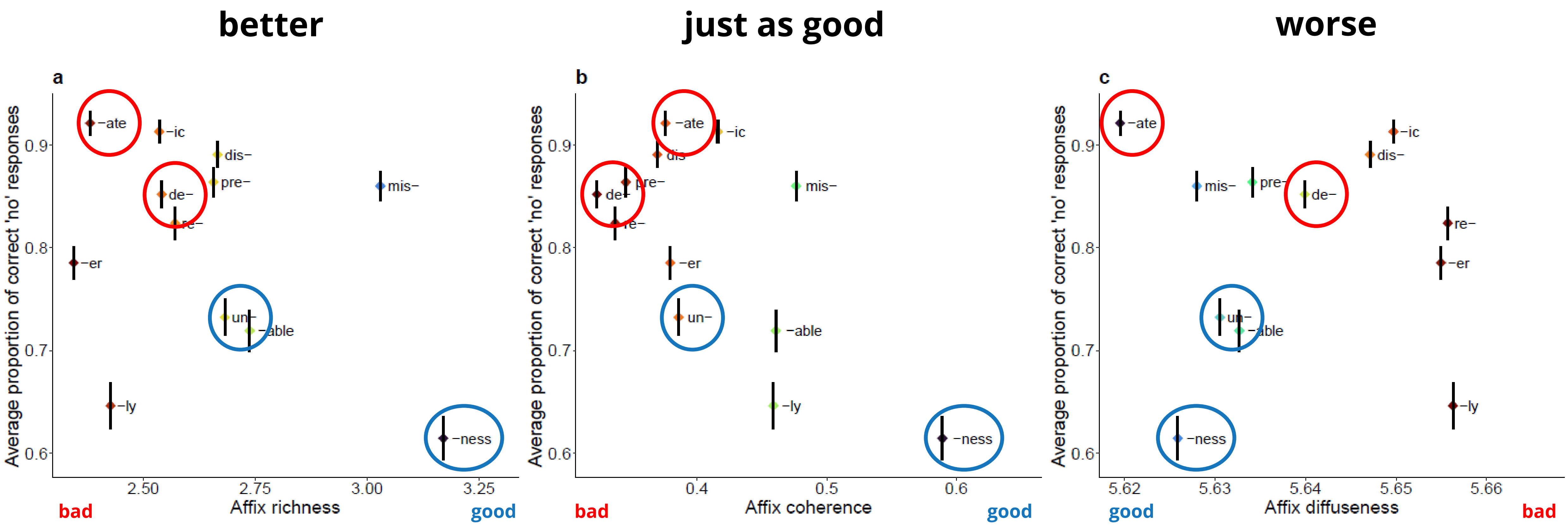




More errors when rejecting nonwords with affixes with…

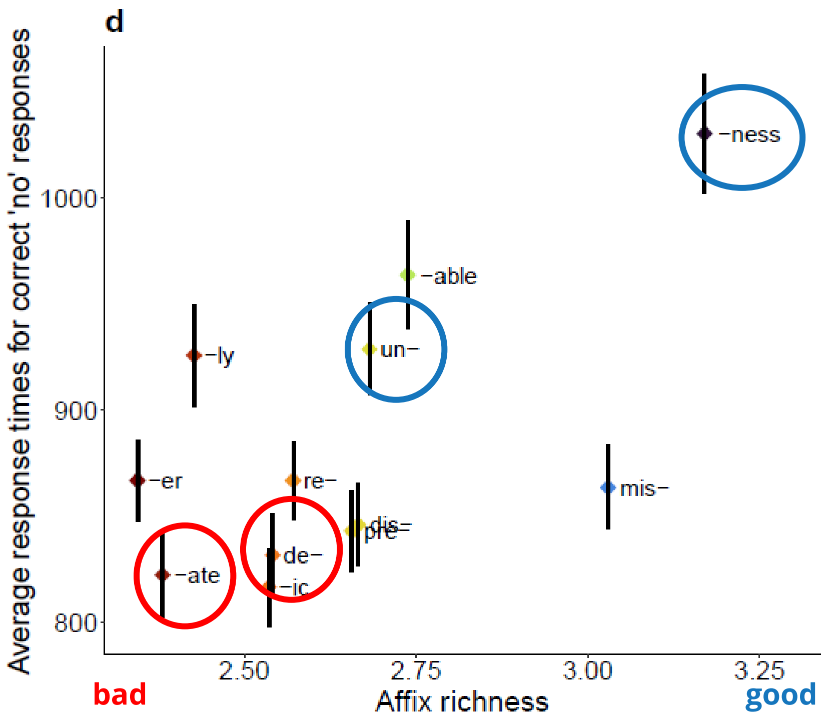


Compared to the false alarm penalty model…

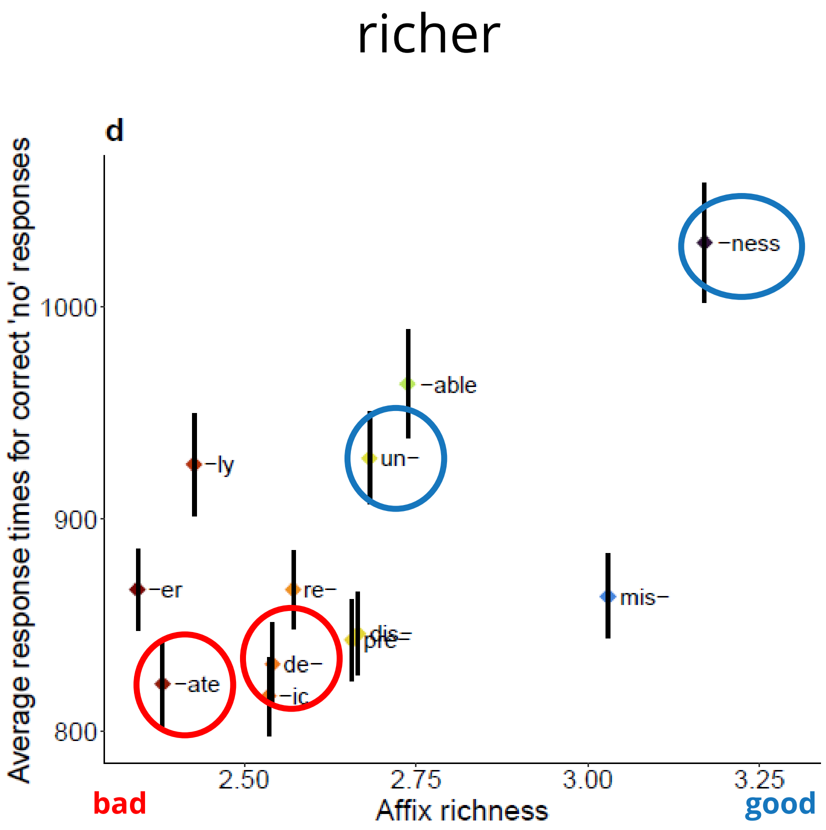


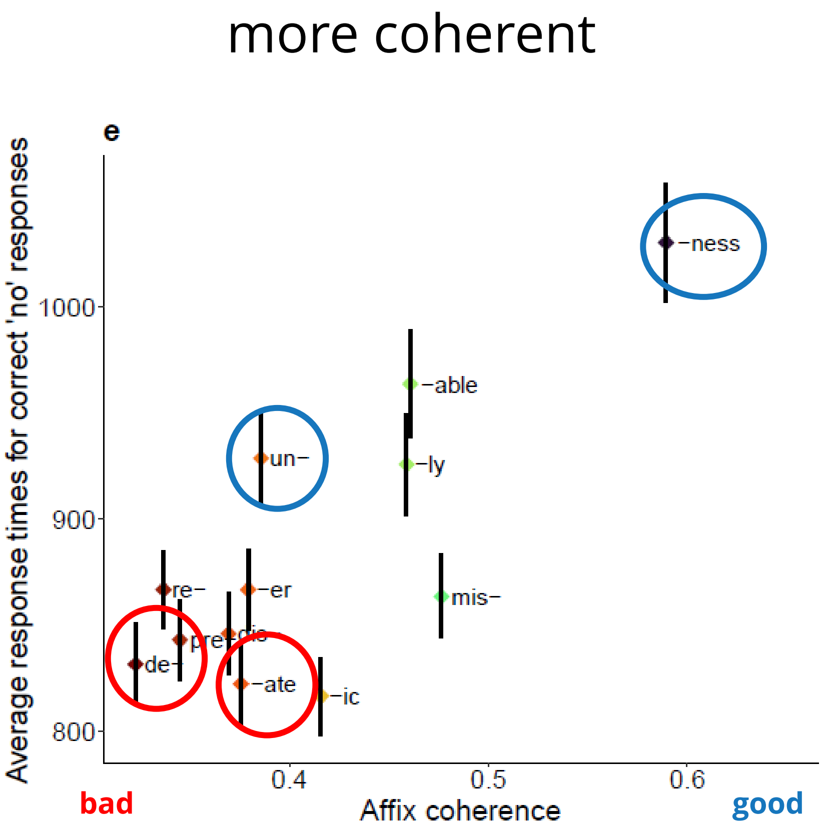
Slower when rejecting nonwords with affixes with...

richer



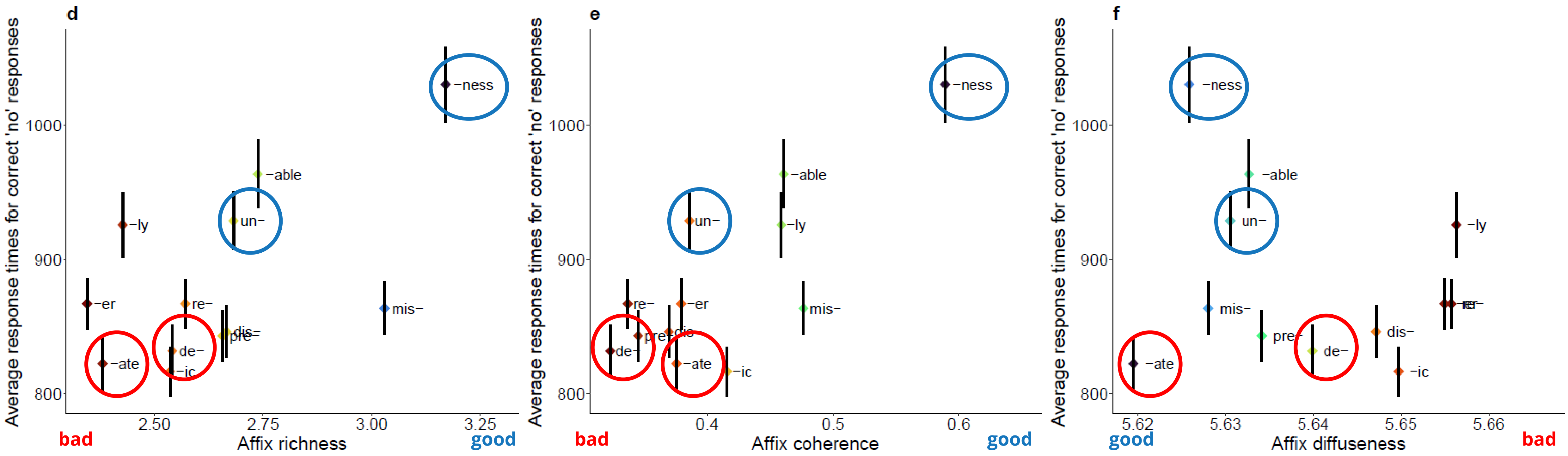
Slower when rejecting nonwords with affixes with…





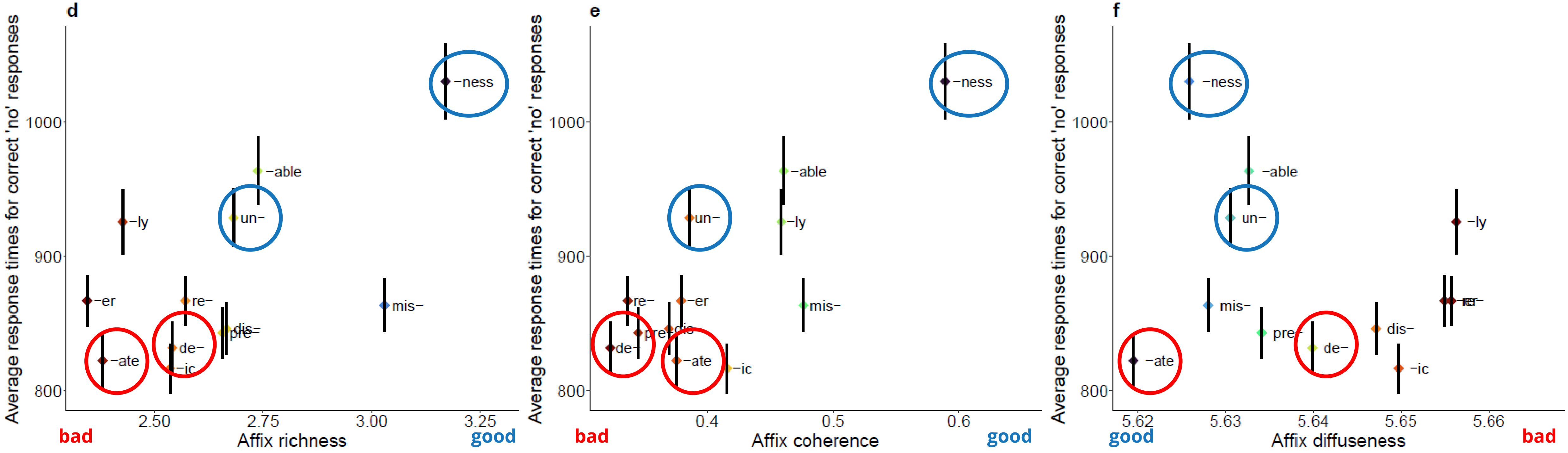
Slower when rejecting nonwords with affixes with…

richer | more coherent | less diffuse meanings



Compared to the false alarm penalty model…

better | **just as good** | **worse**



Recall our interim summary…

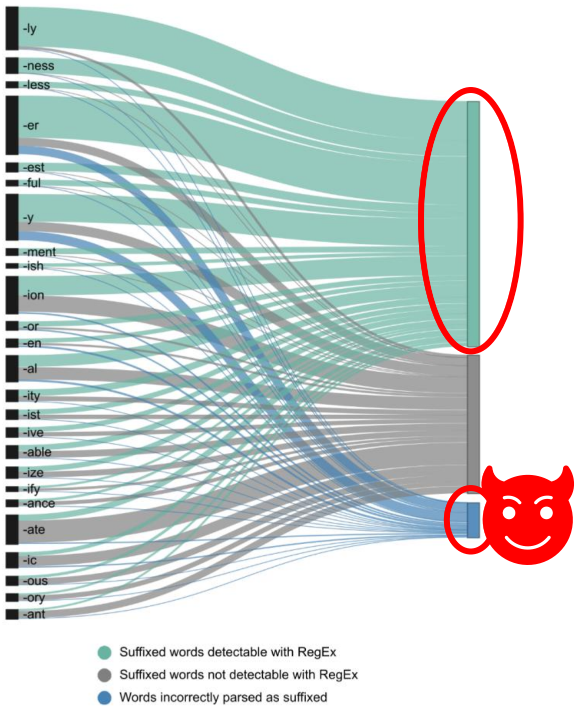
Quantified morpheme **experience** in print

↓

Proposed a new definition of morpheme experience

↓

Tested this definition against human data



- Critical step toward a **psychologically valid theory** of morpheme learning
- However, this approach is still a workaround: needs expert input and reduces affix meaning to a binary distinction

Conclusions

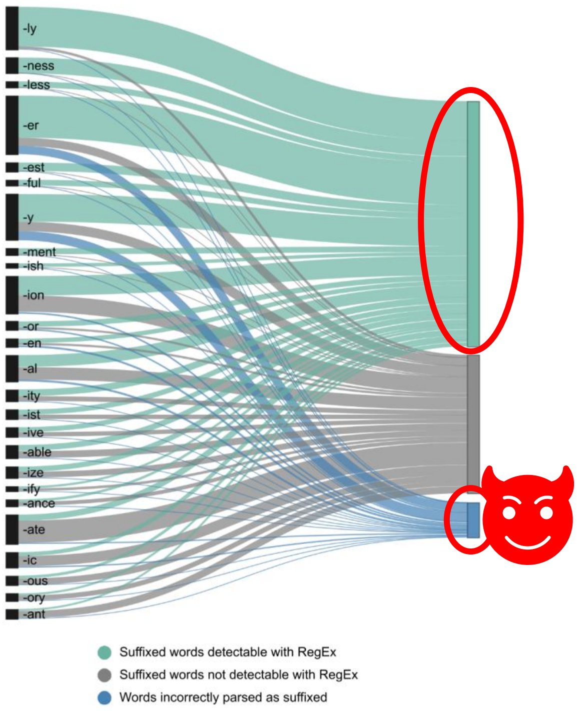
Quantified morpheme **experience** in print

↓

Proposed a new definition of morpheme experience

↓

Tested this definition against human data



- **Principled, scalable** account of **morpheme learning in the wild**
- **Innovation** in computational **modelling** of affix semantics
  - **First** use of CAOSS with **"noisy"** input
  - **First** attempt to model **prefix semantics**

Readers' text experience shapes perception of both affix meaningfulness and plausibility of novel morphemic combinations

Further reading

Article | Open access | Published: 05 May 2025

Morphology in children's books, and what it means for learning

Maria Korochkina & Kathleen Rastle

npj Science of Learning **10**, Article number: 22 (2025) | Cite this article

https://doi.org/10.1038/s41539-025-00313-6

Further reading

Article | Open access | Published: 05 May 2025

Morphology in children's books, and what it means for learning

Maria Korochkina & Kathleen Rastle

npj Science of Learning **10**, Article number: 22 (2025) | Cite this article

4985 Accesses | **23** Altmetric | Metrics

https://doi.org/10.1038/s41539-025-00313-6

Morpheme knowledge is shaped by information available through orthography

Maria Korochkina[1], Holly Cooper[1], Marc Brysbaert[2], and Kathleen Rastle[1]

In press in *Psychon. Bul. Rev.*, pre-print at:

https://doi.org/10.31219/osf.io/ad3jh_v2

Further reading

Article | Open access | Published: 05 May 2025

Morphology in children's books, and what it means for learning

Maria Korochkina & Kathleen Rastle

npj Science of Learning **10**, Article number: 22 (2025) | Cite this article

https://doi.org/10.1038/s41539-025-00313-6

Morphemes in the wild: Modelling affix learning from the noisy landscape of natural text

Maria Korochkina[1], Marco Marelli[2], and Kathleen Rastle[1]

Under review, pre-print at:

https://doi.org/10.31234/osf.io/yzcqm_v1

Morpheme knowledge is shaped by information available through orthography

Maria Korochkina[1], Holly Cooper[1], Marc Brysbaert[2], and Kathleen Rastle[1]

In press in *Psychon. Bul. Rev.*, pre-print at:

https://doi.org/10.31219/osf.io/ad3jh_v2


Holly Cooper


Marco Marelli


Marc Brysbaert


Kathy Rastle

Thank you!

maria.korochkina@rhul.ac.uk
https://mariakna.github.io/

Economic and Social Research Council

ROYAL HOLLOWAY UNIVERSITY OF LONDON

Additional slides

Nonword-based metrics

Role of stem-affix **combination**

3 nonword-based metrics

- **Nonword diffuseness**: how well-defined or vague a nonword's meaning is
- **Nonword richness**: semantic richness of a nonword's meaning
- **Nonword neighbourhood density**: proximity of a nonword to its nearest semantic neighbours

→ Does the inclusion of these metrics into the models with affix-based metrics improve model fit?

Role of stem-affix **combination**

3 nonword-based metrics

- **Nonword diffuseness**: how well-defined or vague a nonword's meaning is
- **Nonword richness**: semantic richness of a nonword's meaning

Yes, for the affix richness (accuracy only) and affix diffuseness models

- **Nonword neighbourhood density**: proximity of a nonword to its nearest semantic neighbours

Yes, for all response times models

→ Does the inclusion of these metrics into the models with affix-based metrics improve model fit?

Summing up

- Morphologically structured nonwords most difficult to reject when…
  - they are semantically rich,
  - closely related in meaning to their semantic neighbours,
  - contain affixes with richer, more coherent, and less diffuse meanings

- Affix meaningfulness influenced processing **more** than the overall nonword meaning
- Skilled readers' judgments of affixed nonwords are driven **mainly by the properties of the affixes** they contain, rather than by the specific meaning of the stem-affix combination in each individual nonword