Emergent morpho-phonological representations in models of spoken word recognition

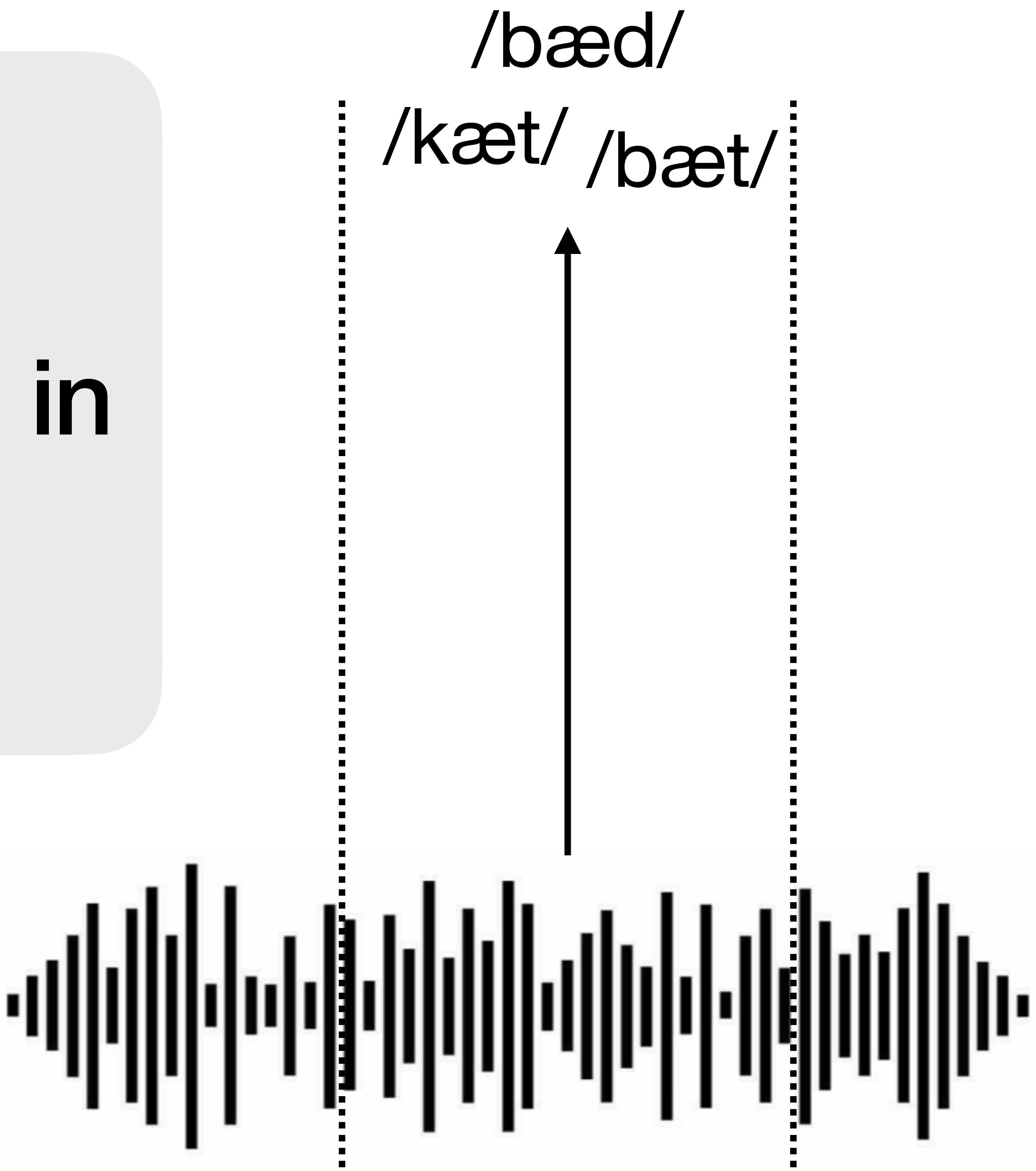
Jon Gauthier[1]
Matthew Leonard[1]
Canaan Breiss[2]
Edward Chang[1]

[1] UCSF [2] USC

AMP 2025
UC Berkeley

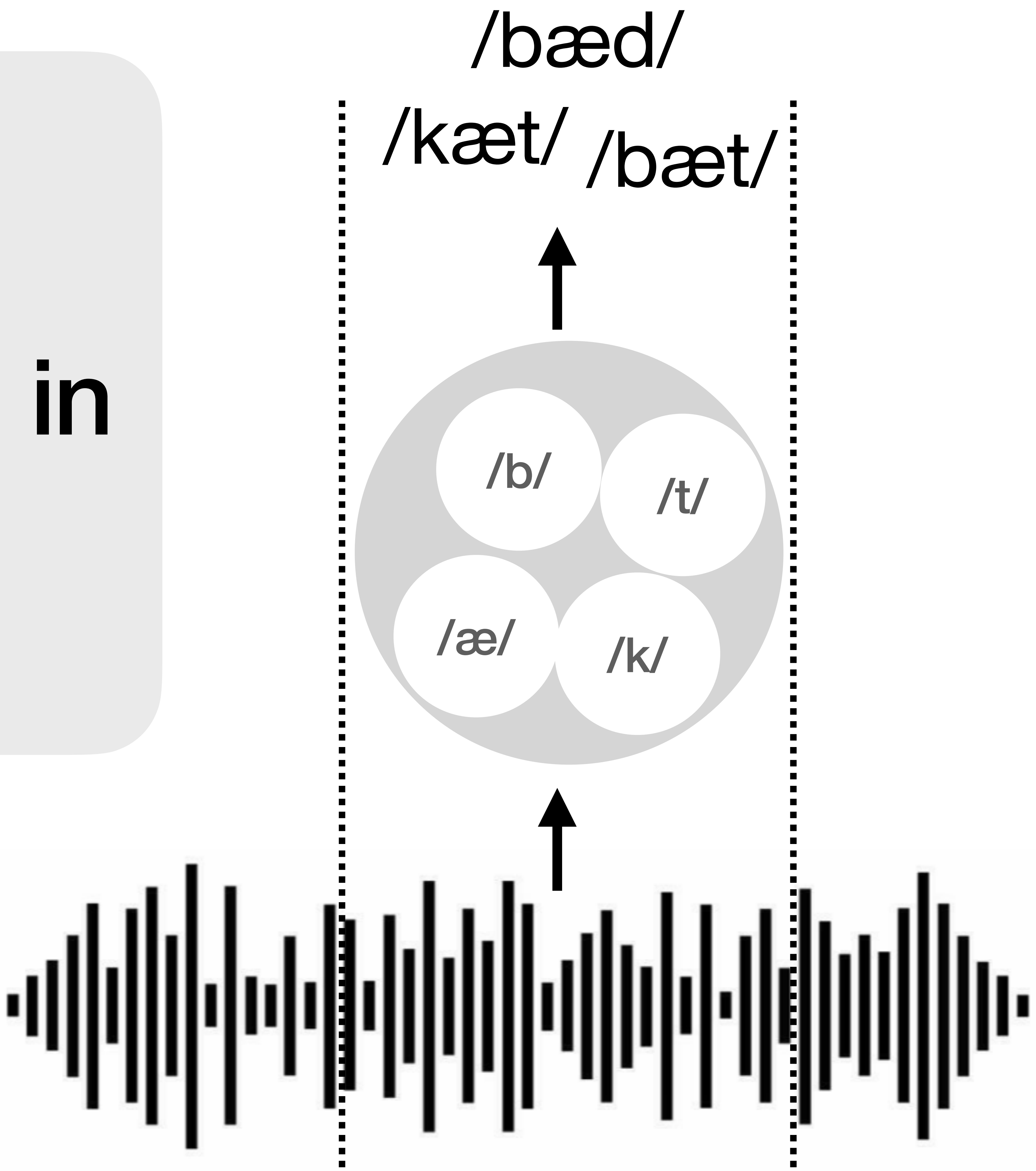
Spoken word recognition

What kinds of linguistic representations are recruited in spoken word recognition?



Spoken word recognition

What kinds of linguistic representations are recruited in spoken word recognition?



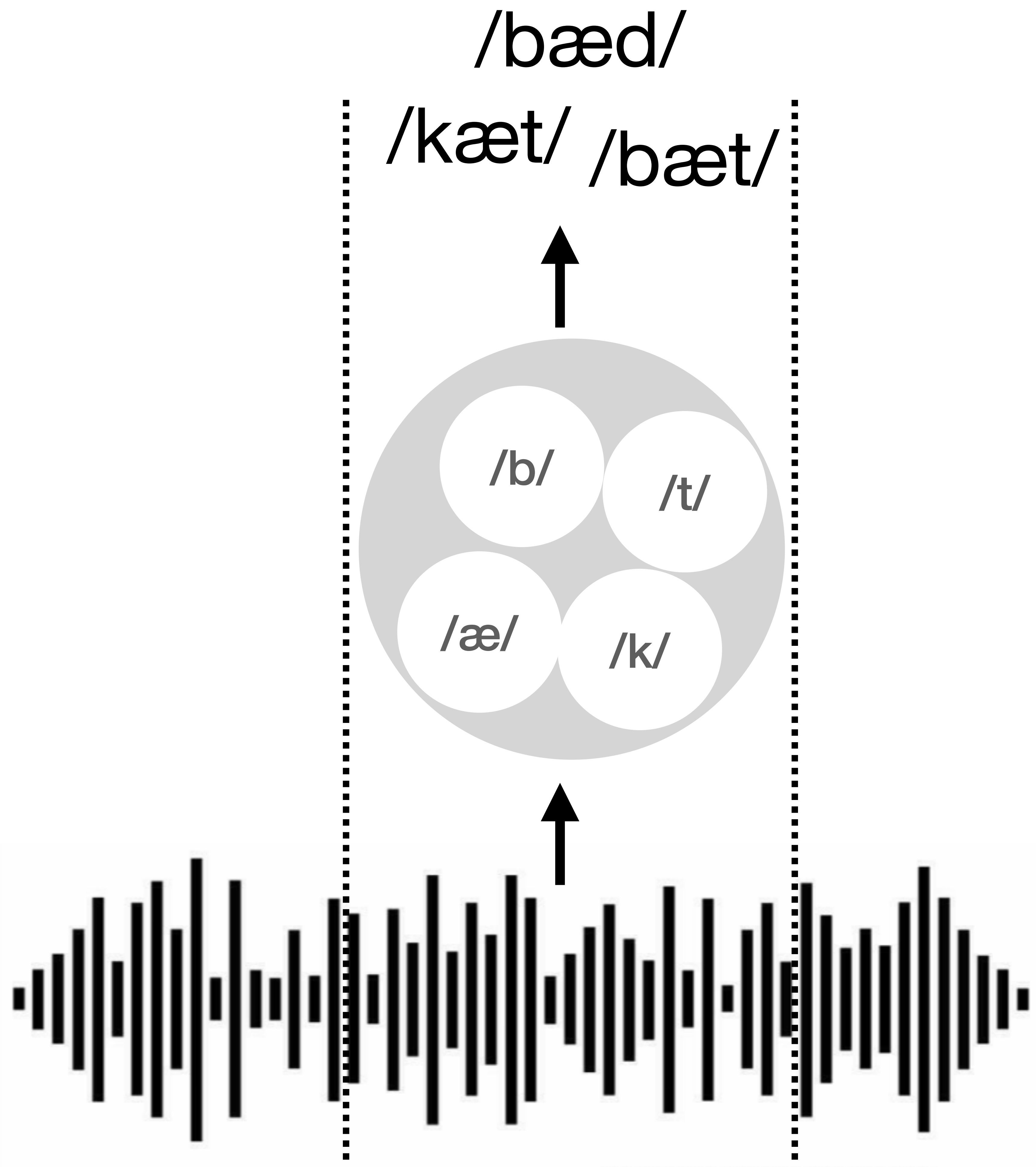
Spoken word recognition

Classic computational theories

TRACE, Shortlist, Merge, DCM, …

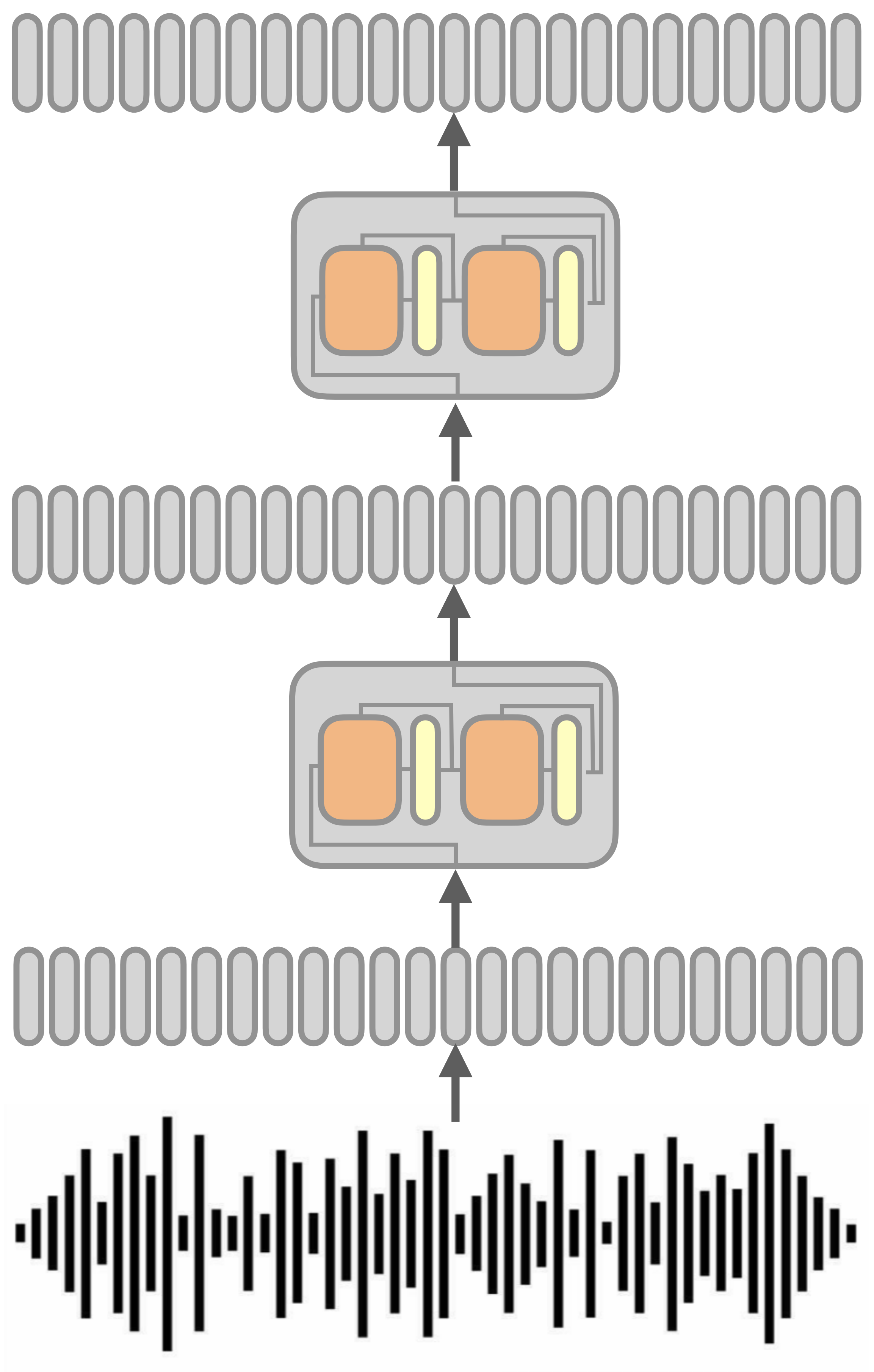
Explicit levels of linguistic representation

Explain spoken word recognition at small scales



TRACE (McClelland & Elman, 1986)

Spoken word recognition



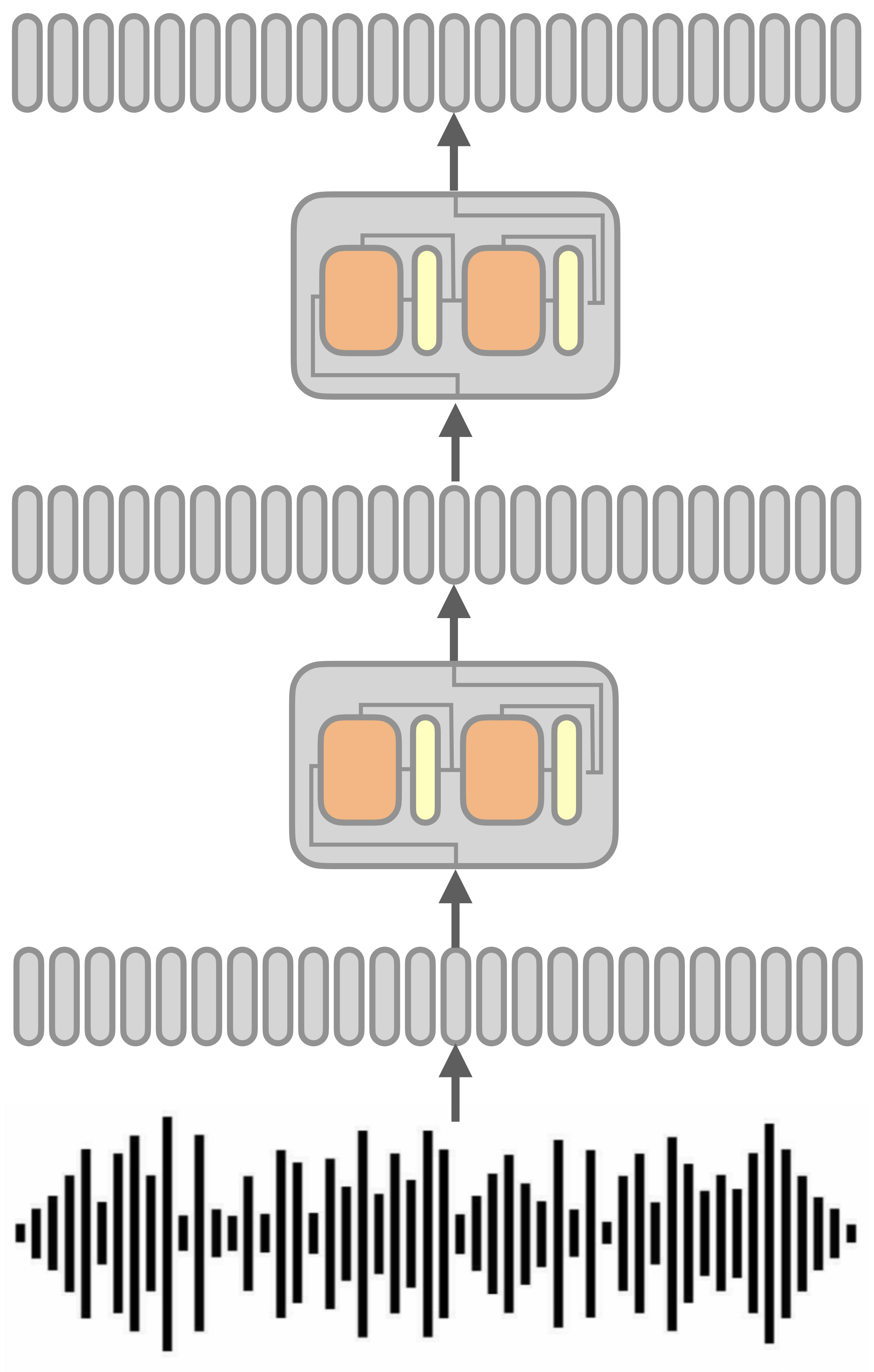
Self-supervised models

Word recognition at large scale, from raw audio

No explicit linguistic representation

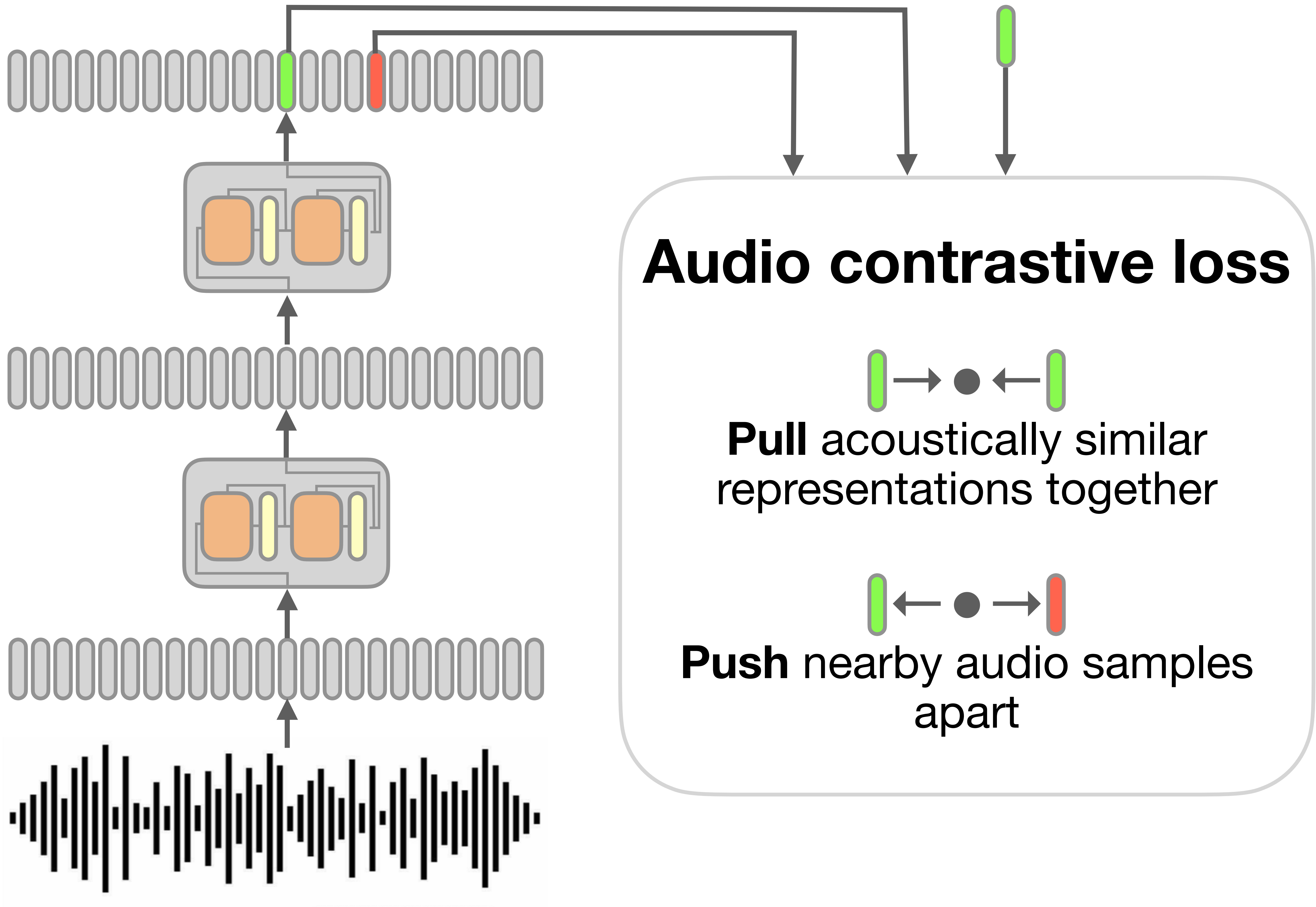
Not (yet) interpretable as a cognitive theory

Spoken word recognition



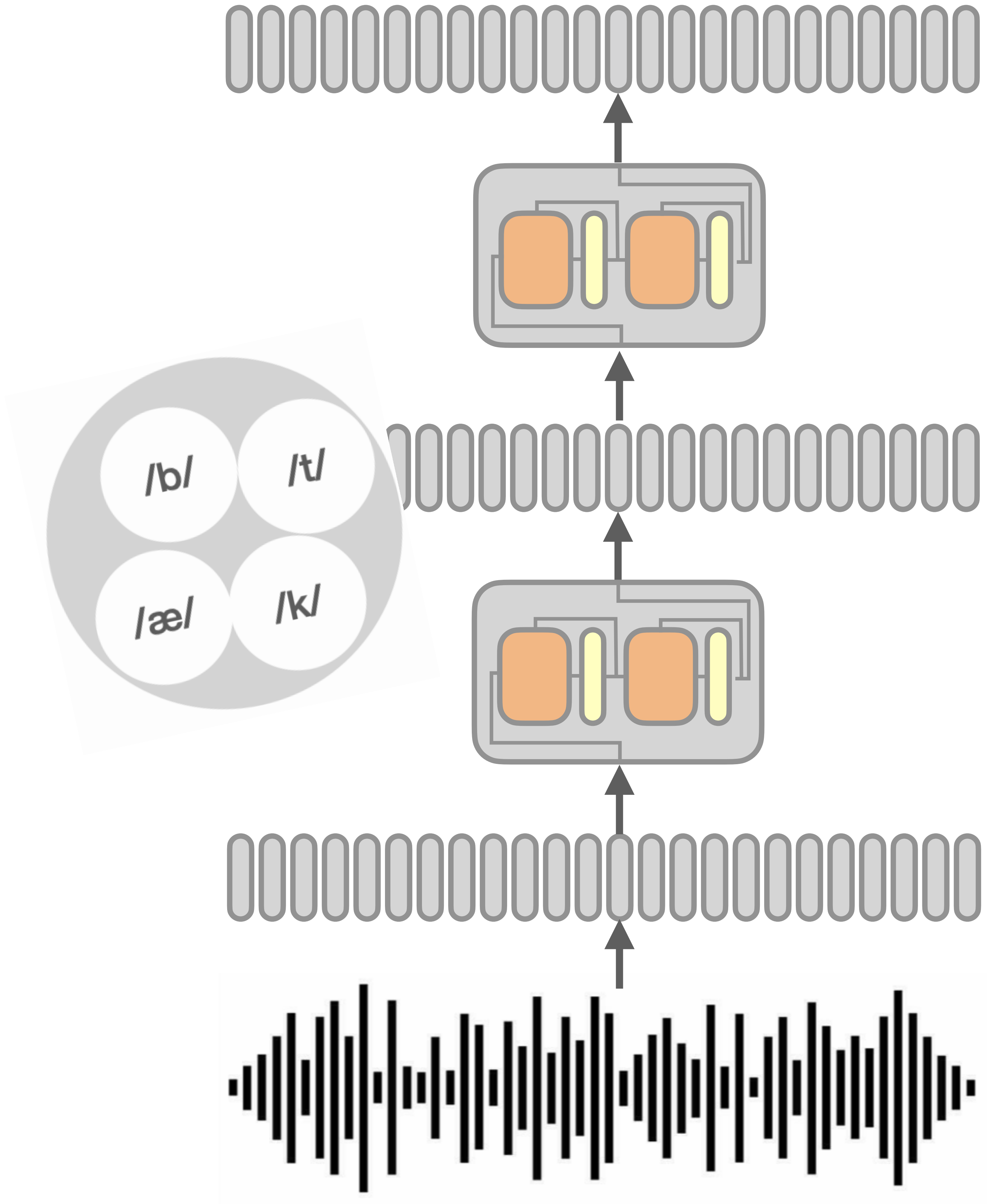
What kinds of linguistic representations are recruited in models of spoken word recognition?

A self-supervised model: `wav2vec2`



Baevski et al. (2020)

A self-supervised model: `wav2vec2`

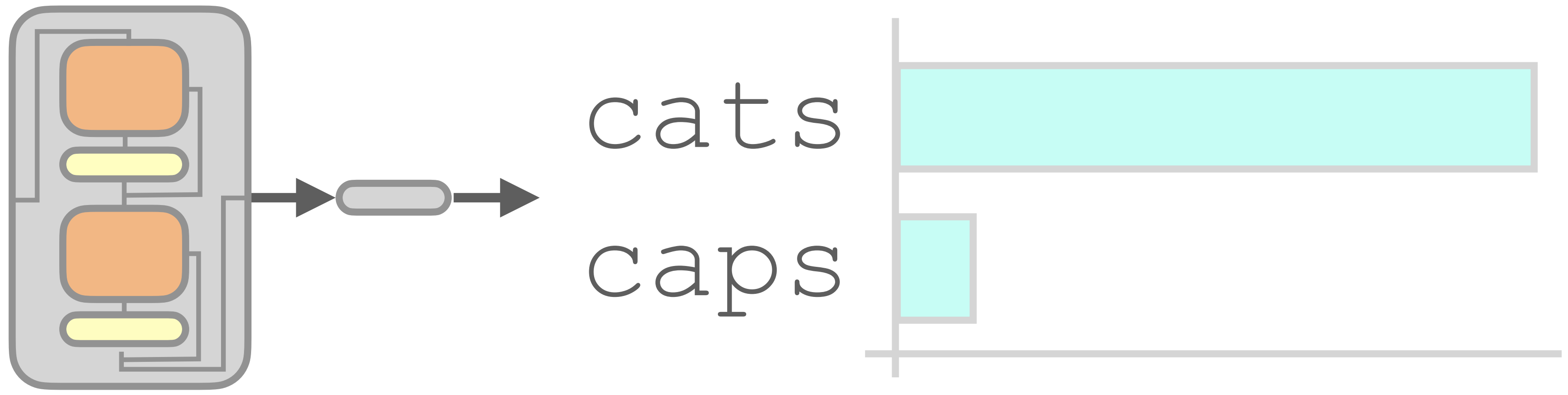


- Self-supervised models encode basic phonological categories
  … but these may serve many functions beyond word recognition

  (Pasad et al. 2021, 2023; Martin et al. 2023; Abdullah et al. 2023; Choi et al. 2024, 2025)

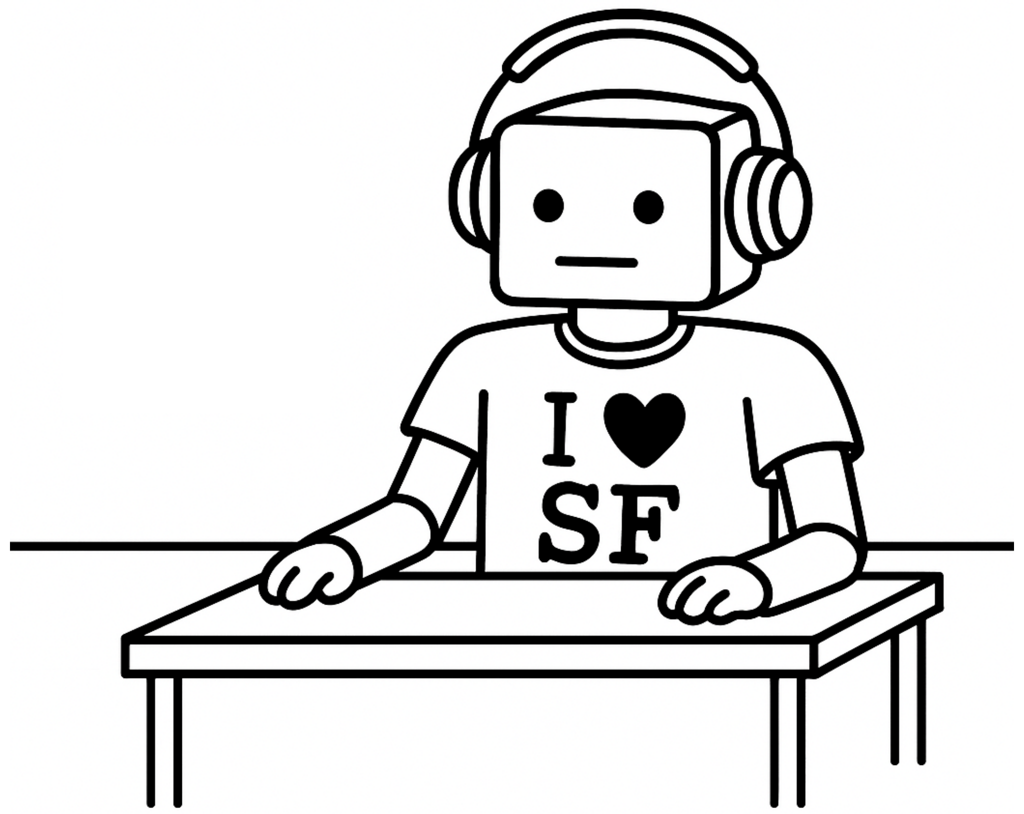
Plan

Model



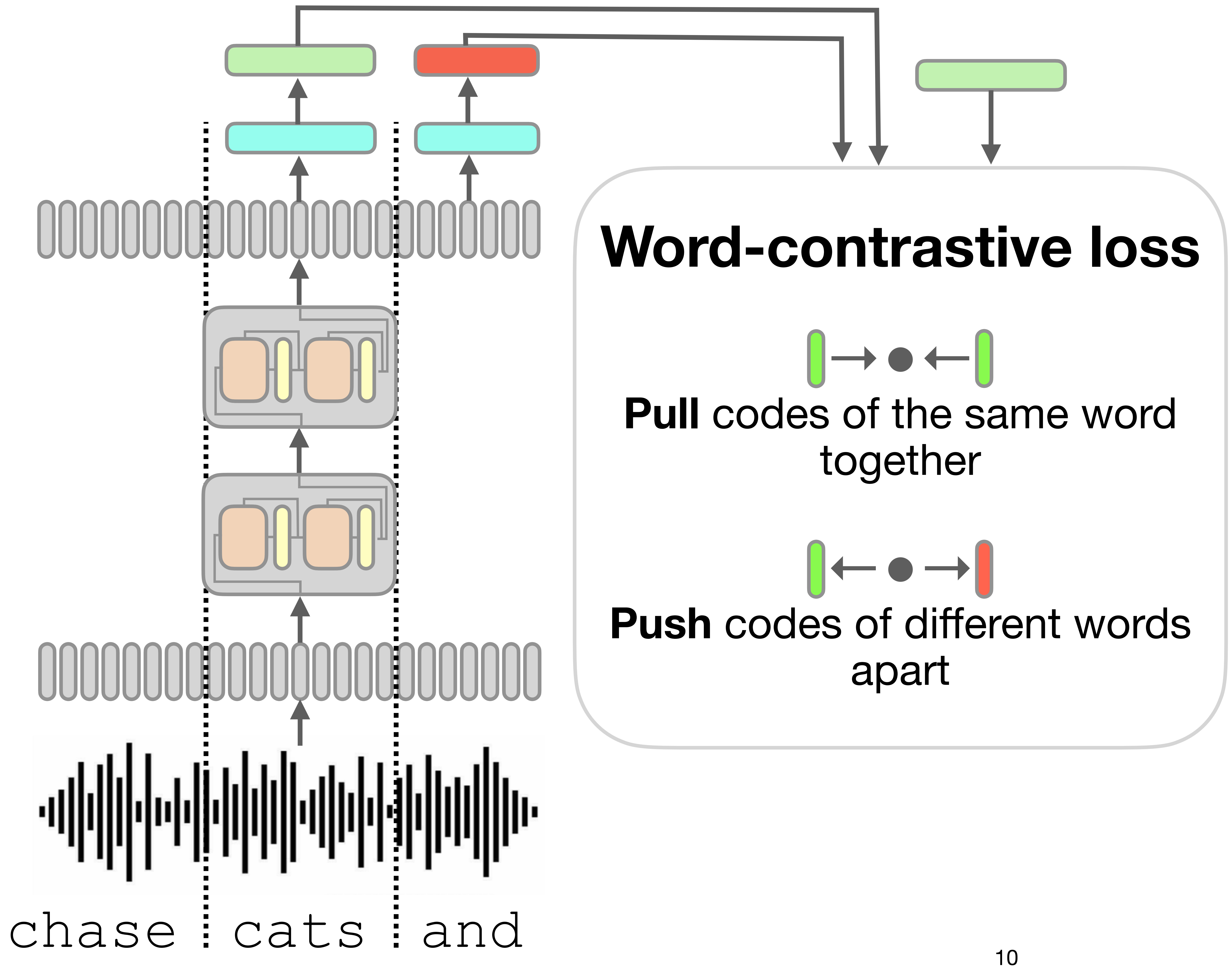
Derive a **word recognition** model from a self-supervised model

Experiment

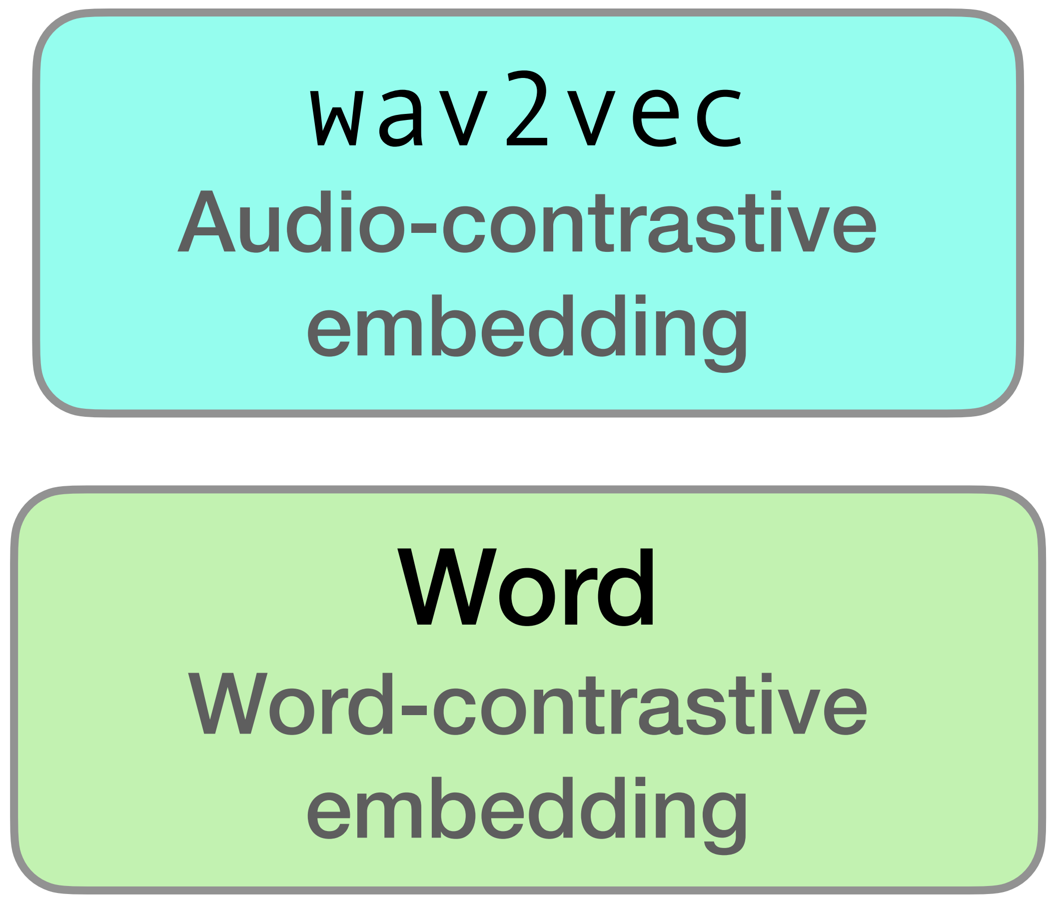


Dissect its computations by treating it as an experimental subject

Word recognition model

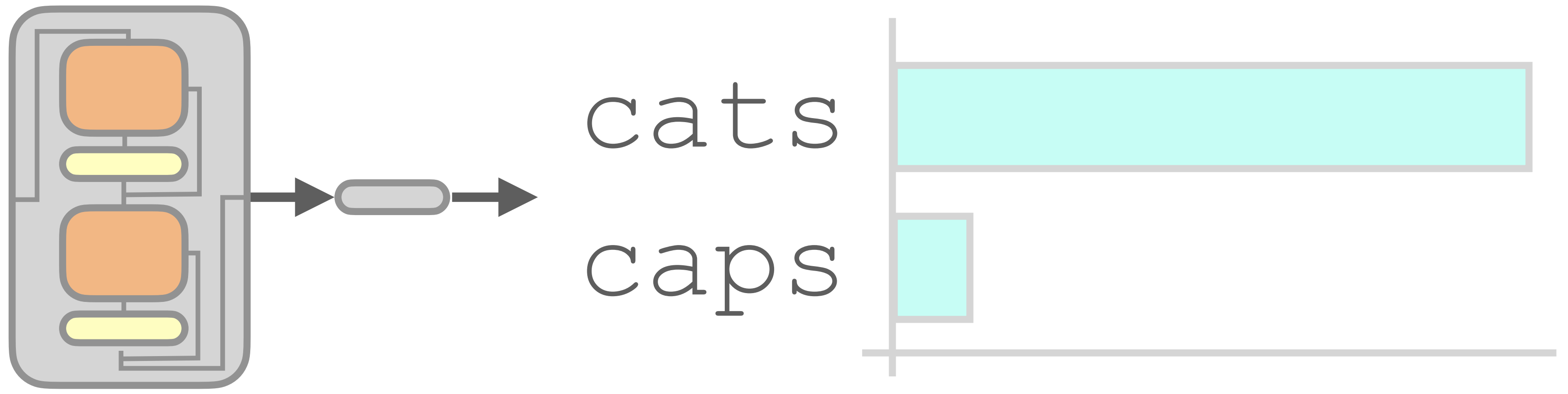


- We compute embeddings for every word token in a test corpus:



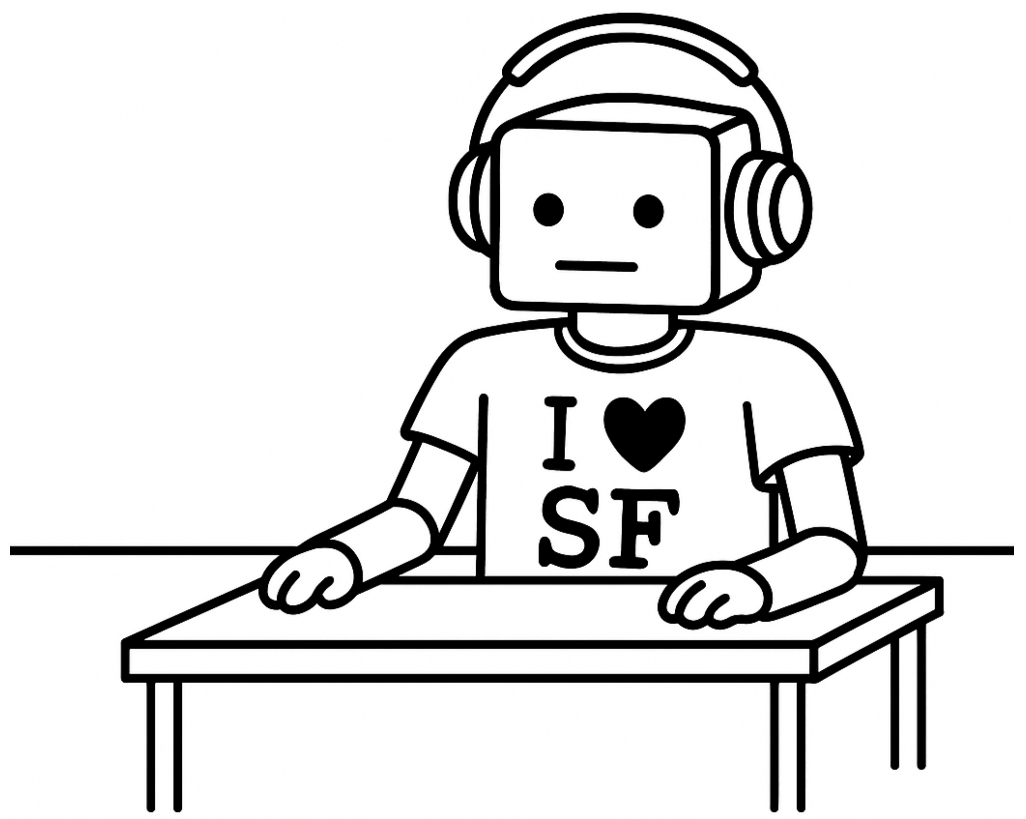
Plan

Model



Derive a **word recognition** model from a self-supervised model

Experiment



Dissect its computations by treating it as an experimental subject

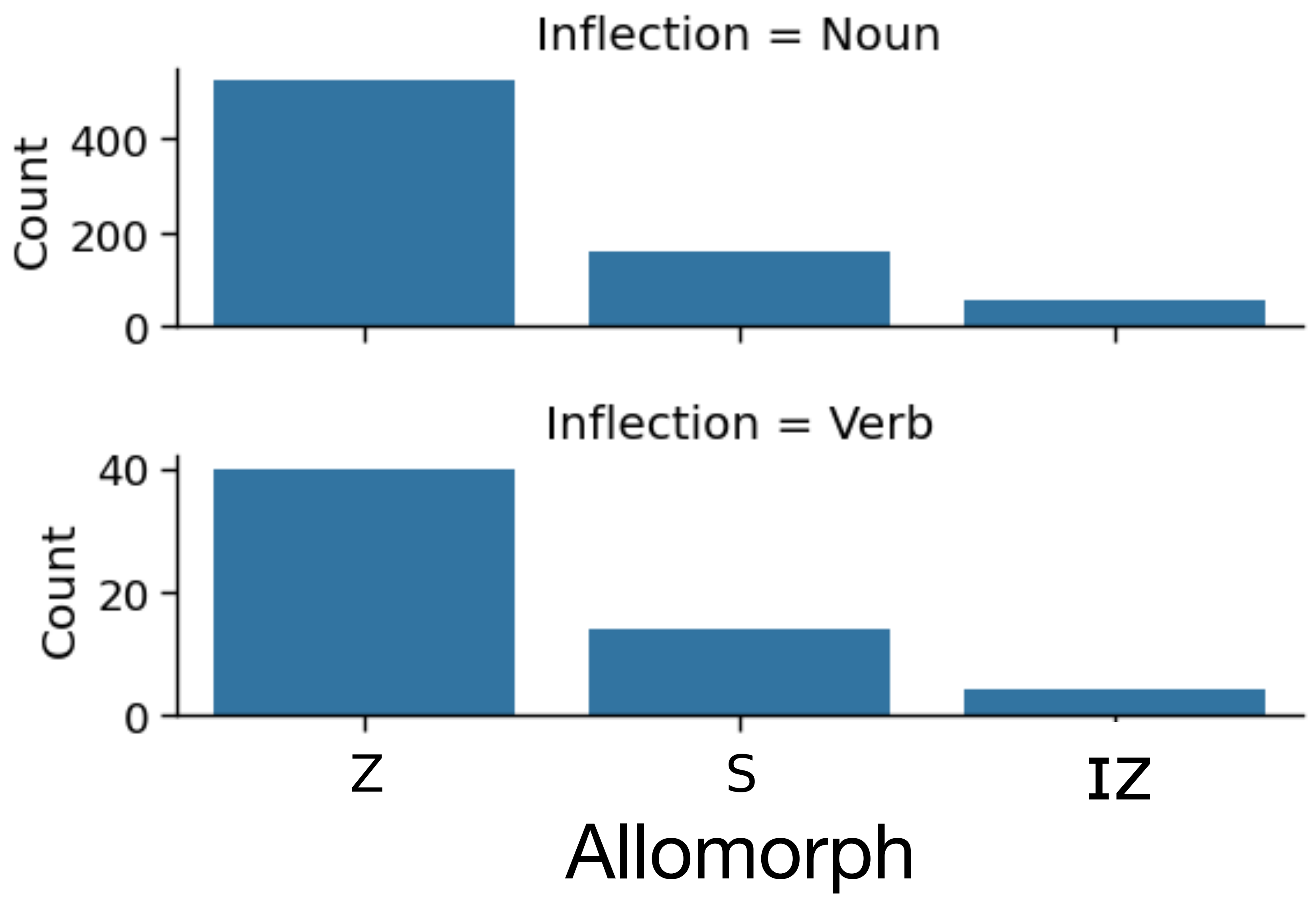
Phenomenon

- Word-final [z], [s], [ɪz]
- Distributed by multiple **morphological** processes
- Governed by **phonological rules:**
  - [ɪz] after sibilants
  - [z] after voiced segments
  - [s] after voiceless segments

		Base	**Inflected**
		daughter	daughters
		lip	lips
		age	ages
		bring	brings
		speak	speaks
		please	pleases

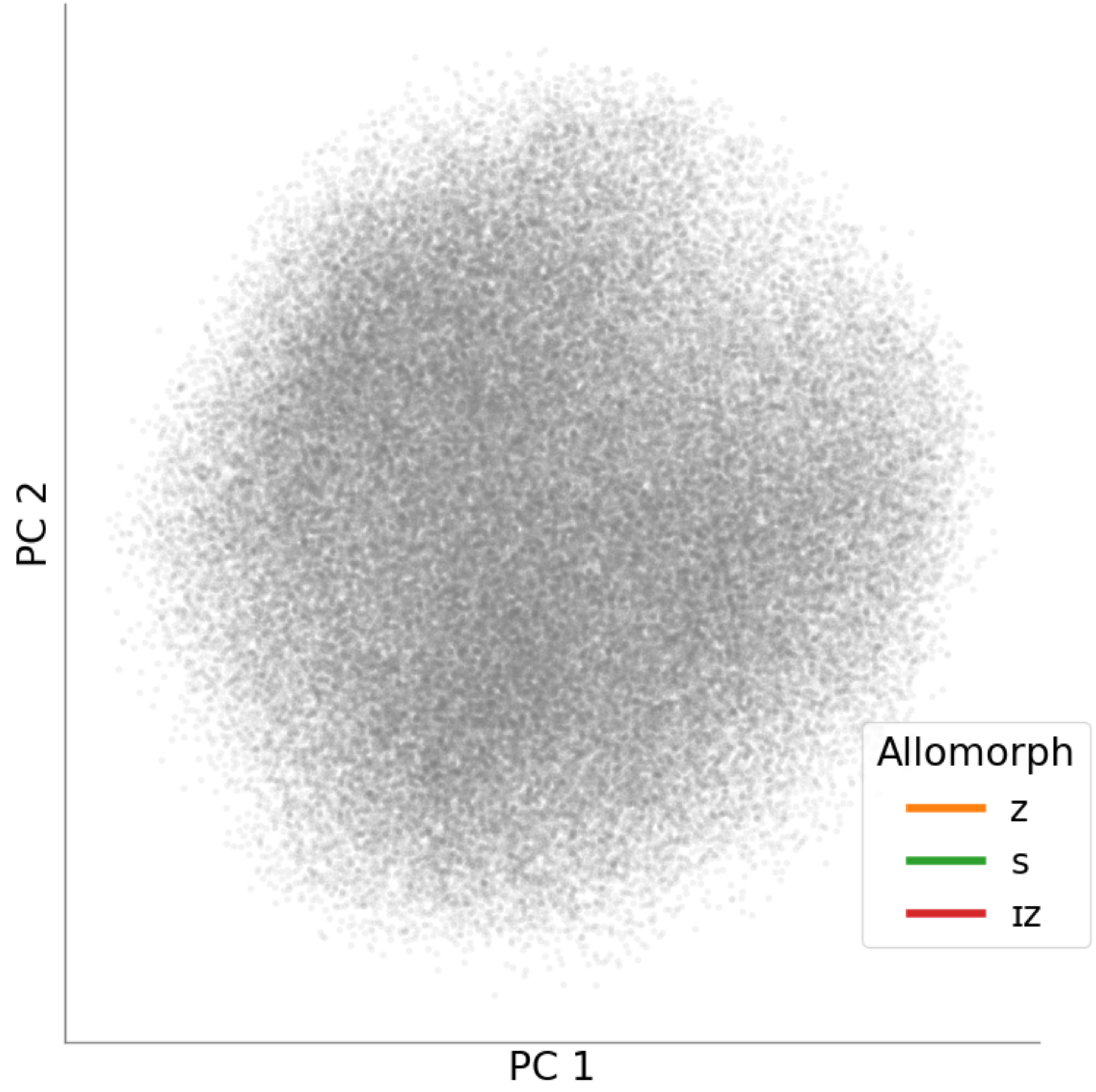
Corpus

- LibriSpeech corpus: 960 hours of amateur audiobook recordings (AmE, BrE)
- Source 786 regular nouns and 61 regular verbs whose inflected forms are **unambiguous**, e.g.
    - *belongs* is only a 3SG verb and not a plural noun
    - *currents* is only a plural noun and not a 3SG verb



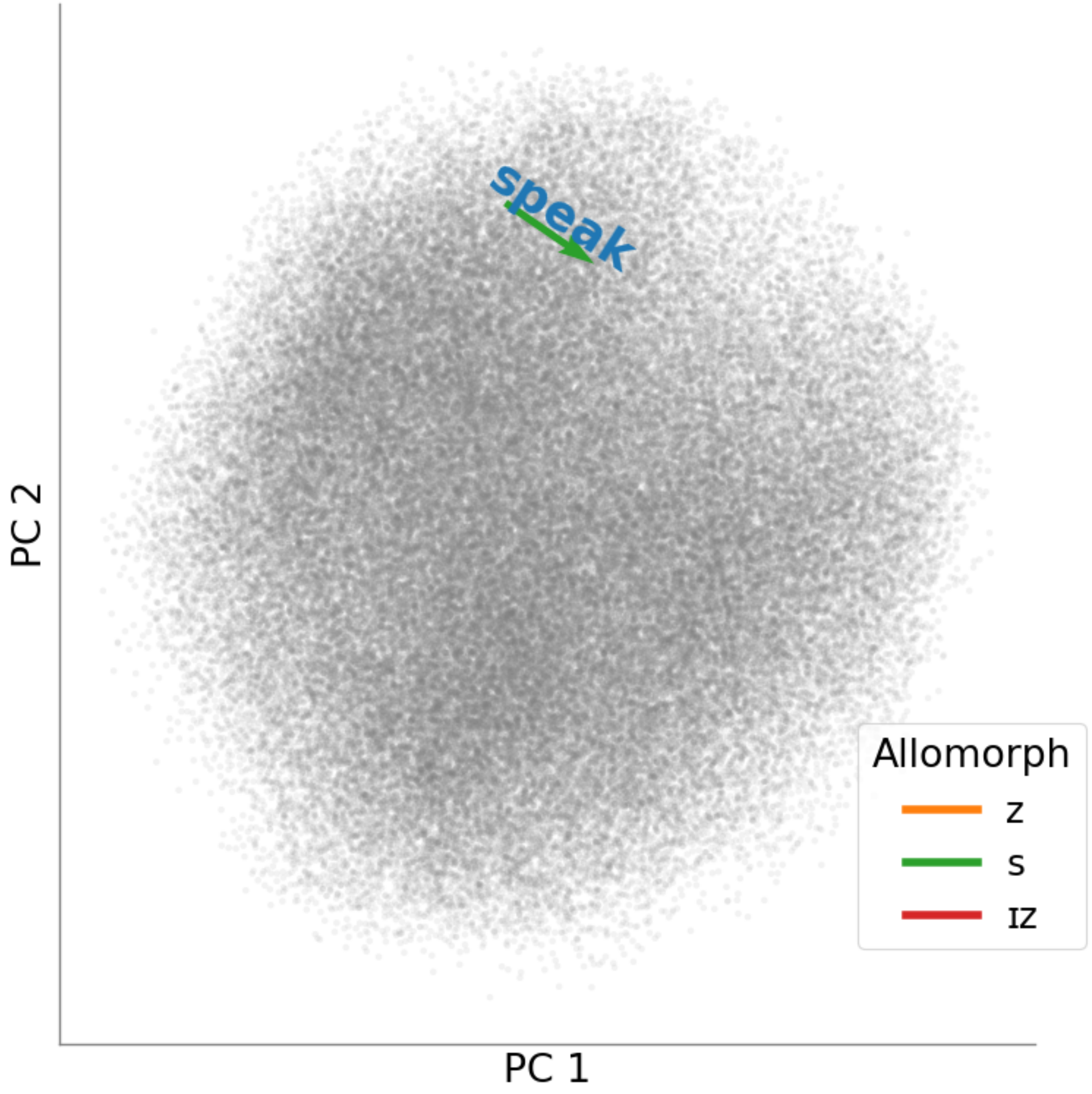
Global linear geometry

Word
Word-contrastive embedding



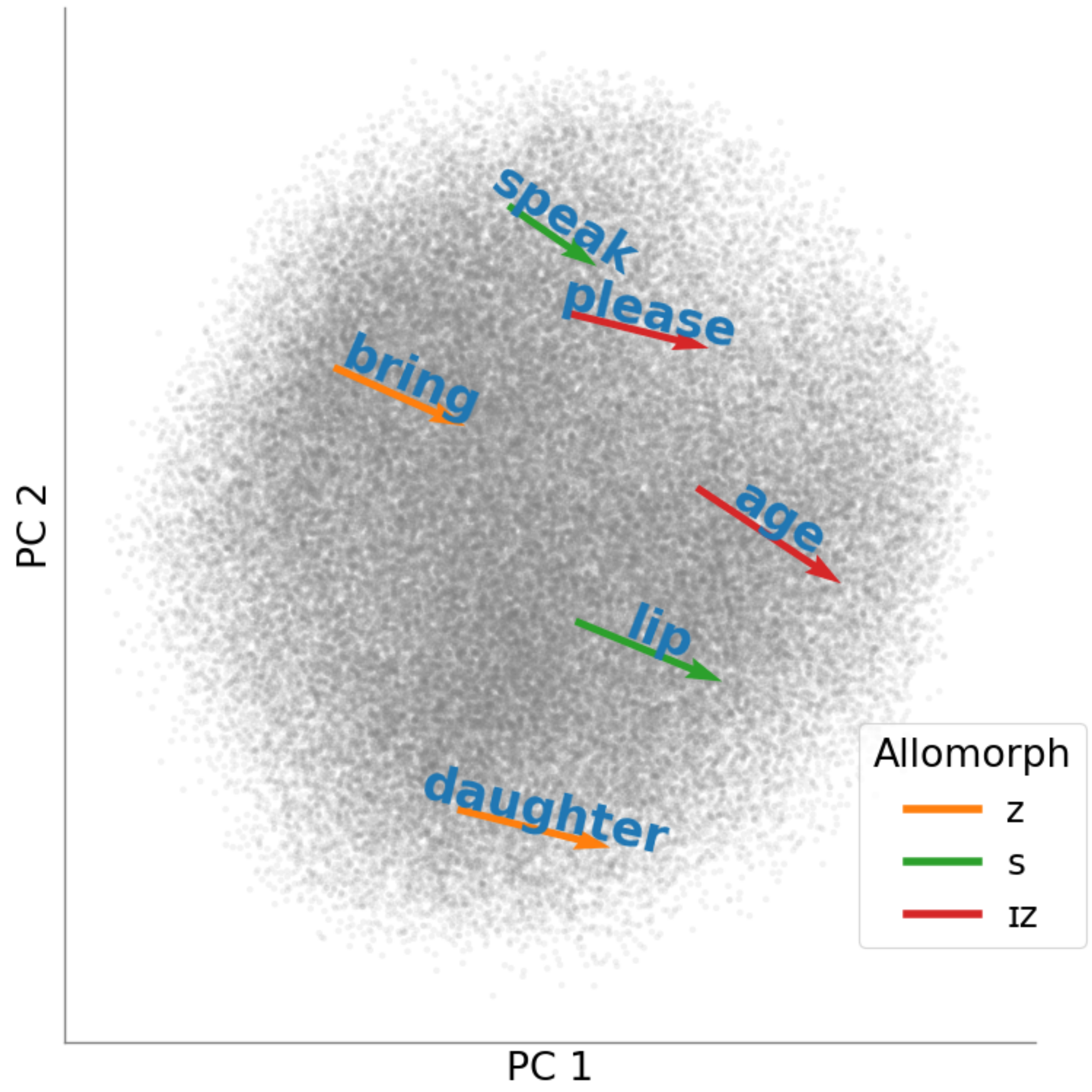
Global linear geometry

Word
Word-contrastive embedding



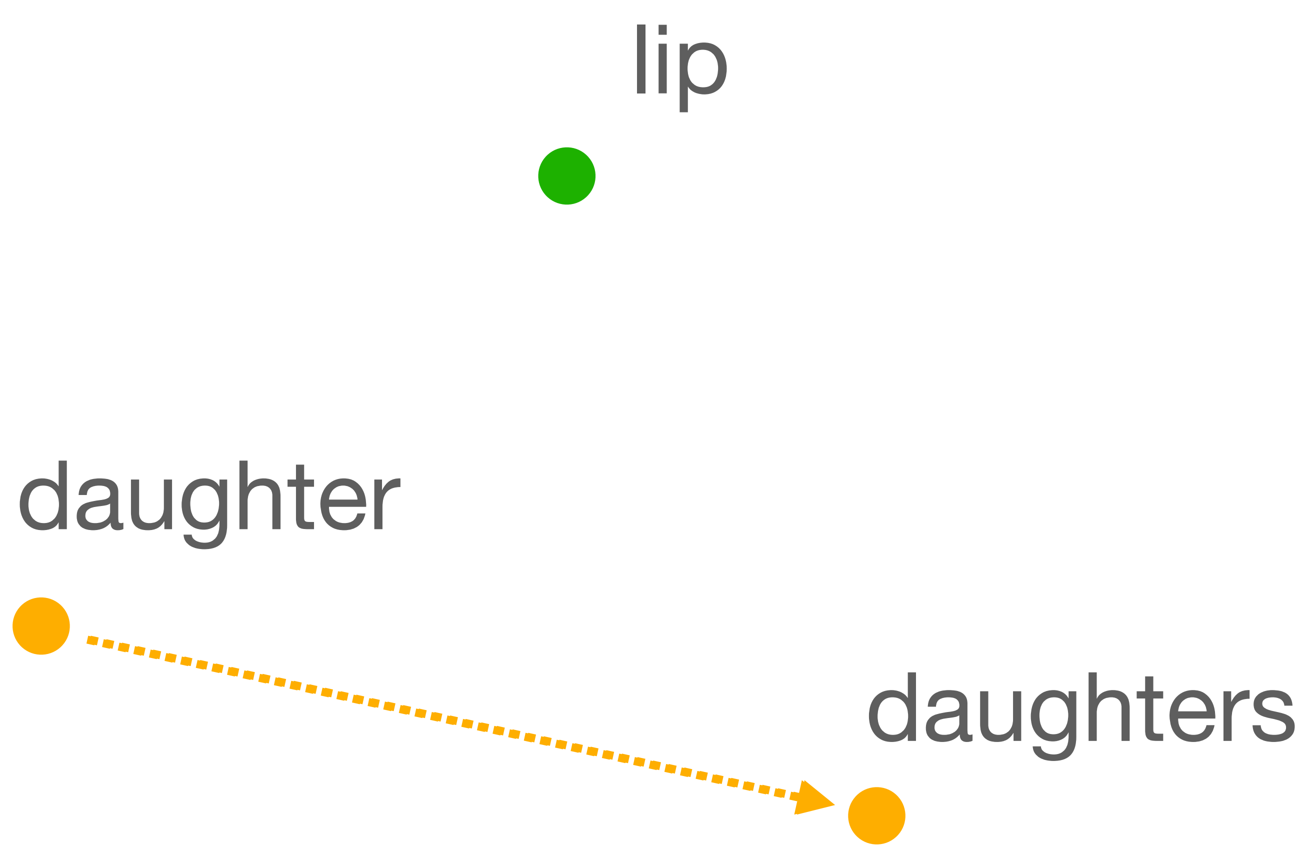
Global linear geometry

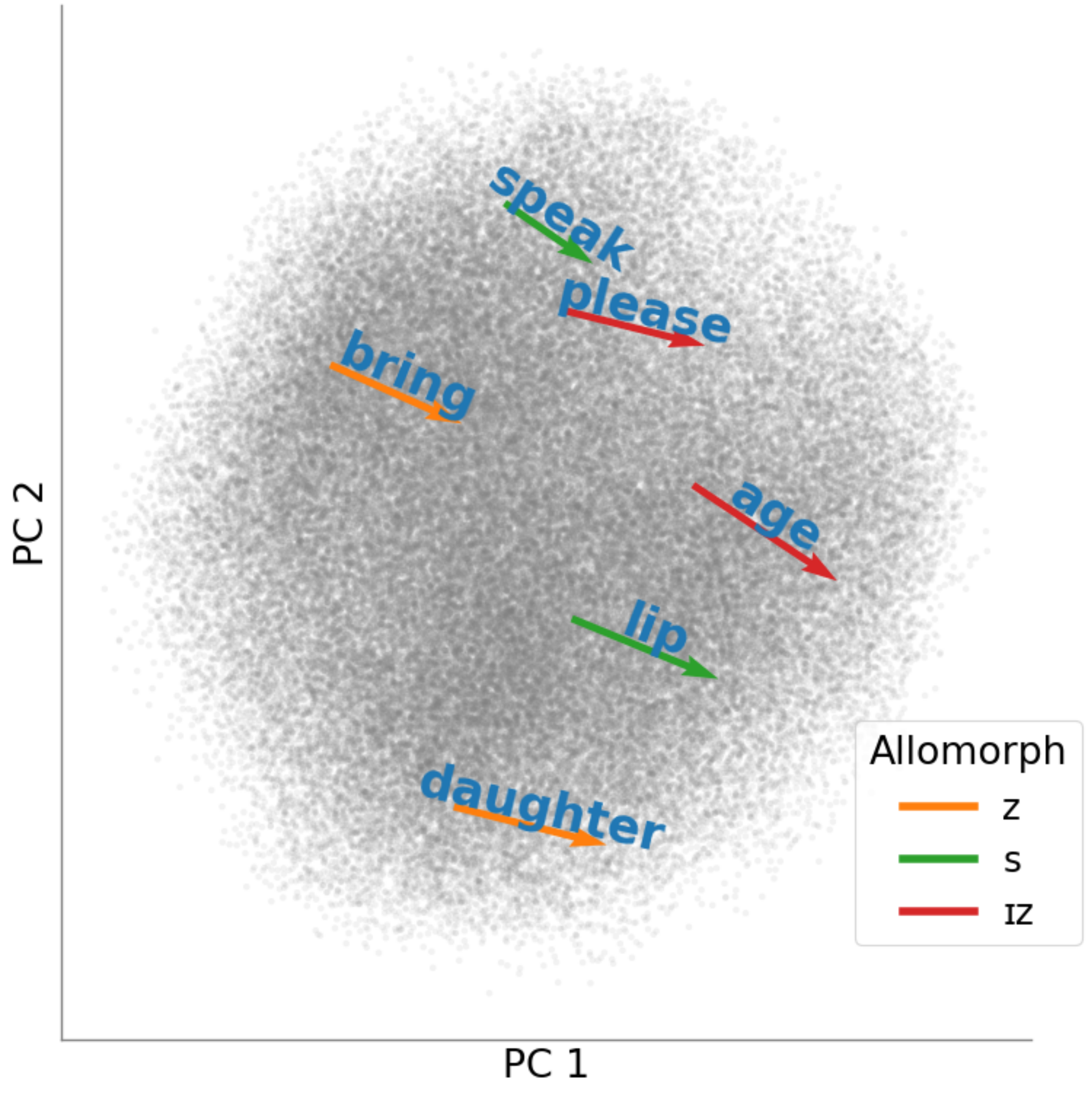
Word
Word-contrastive embedding



Hypothesis

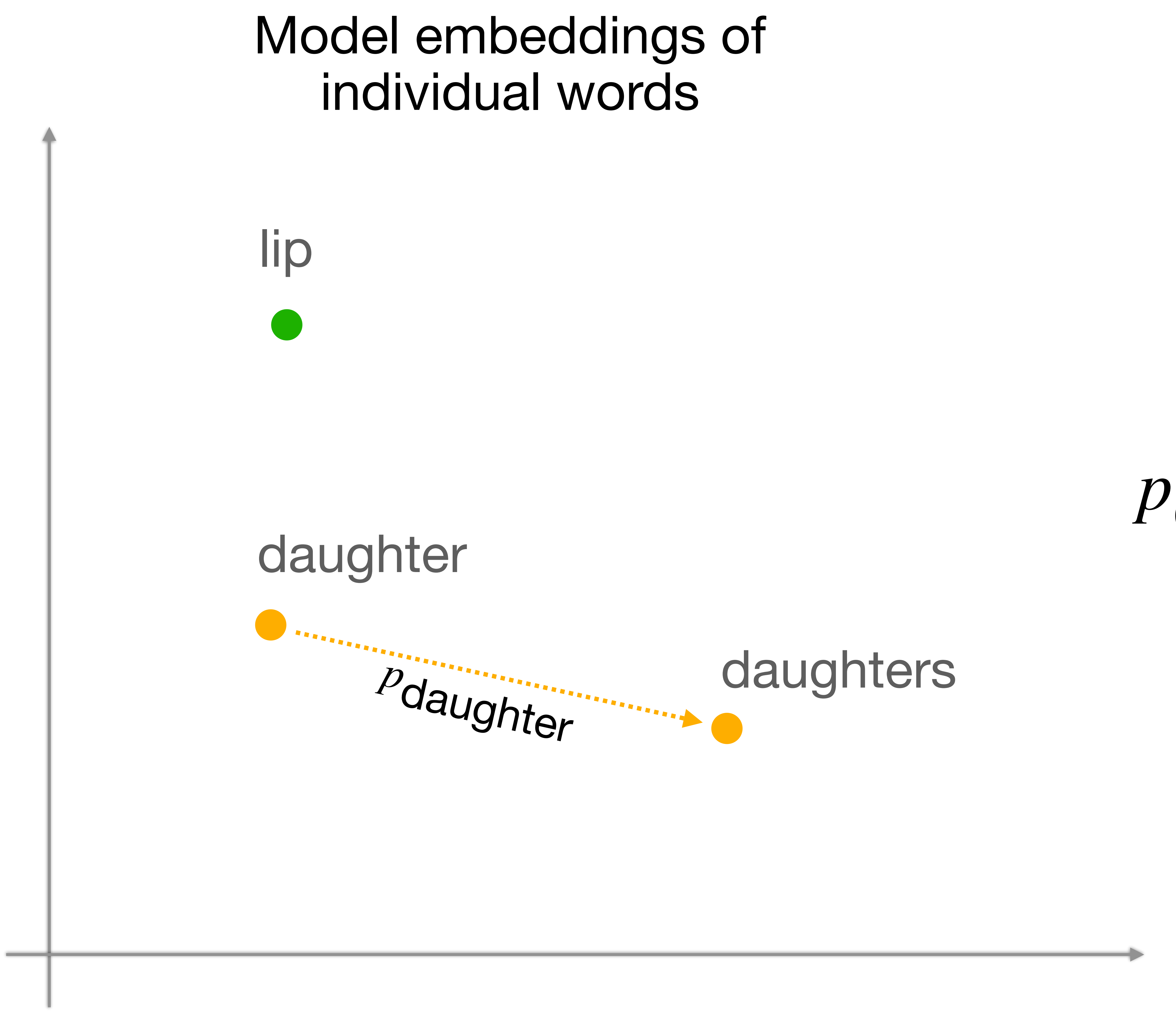
A global **linear** translation links the representations of **base** and **inflected** forms





Prediction

`daughter : daughters :: lip : _____`



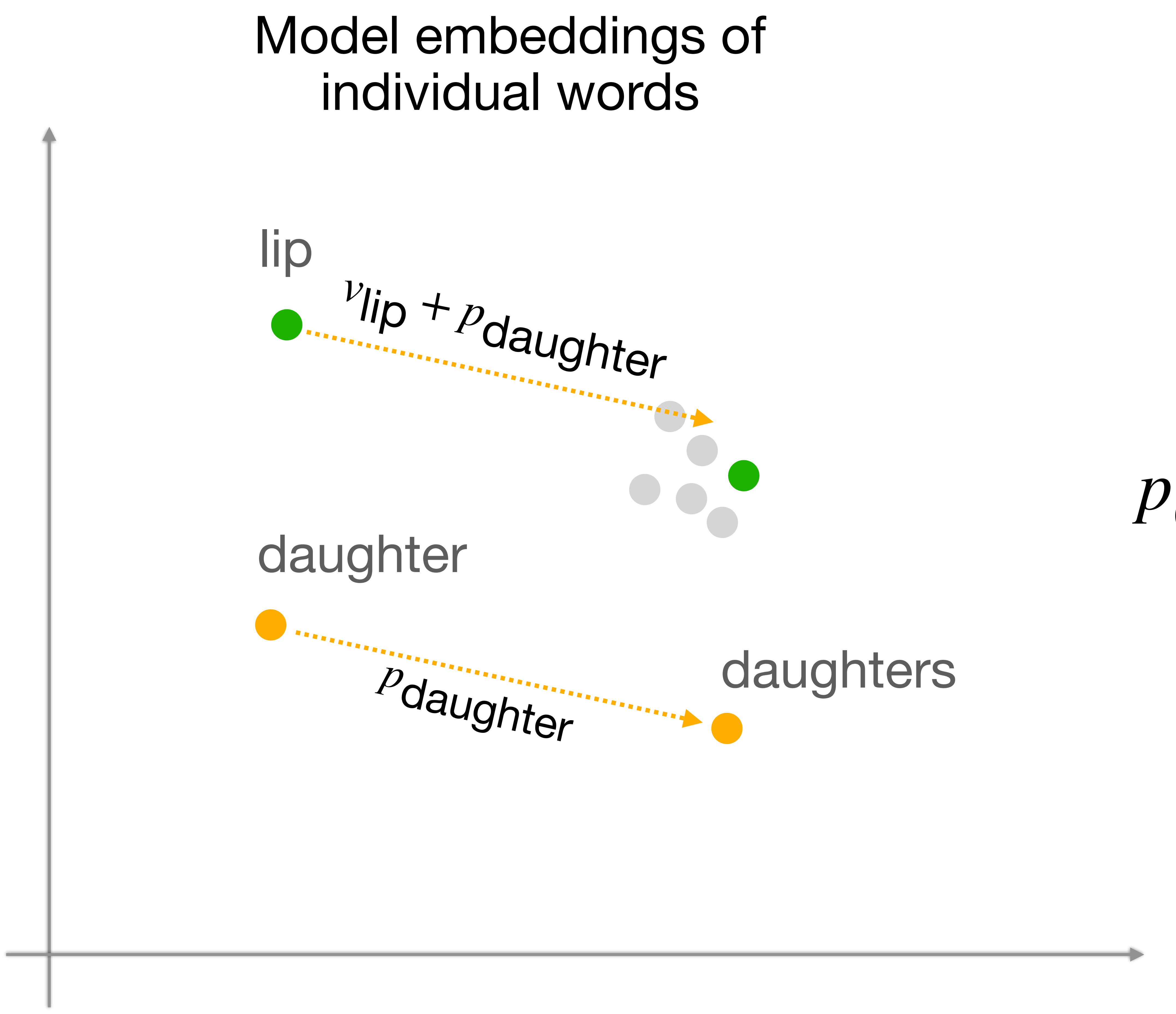
Compute analogy by vector algebra:

$$p_{\text{daughter}} = v_{\text{daughters}} - v_{\text{daughter}}$$

Mikolov et al. (2013)
Ethayarajh, Duvenaud & Hirst (2019)

Prediction

`daughter : daughters :: lip : _____`



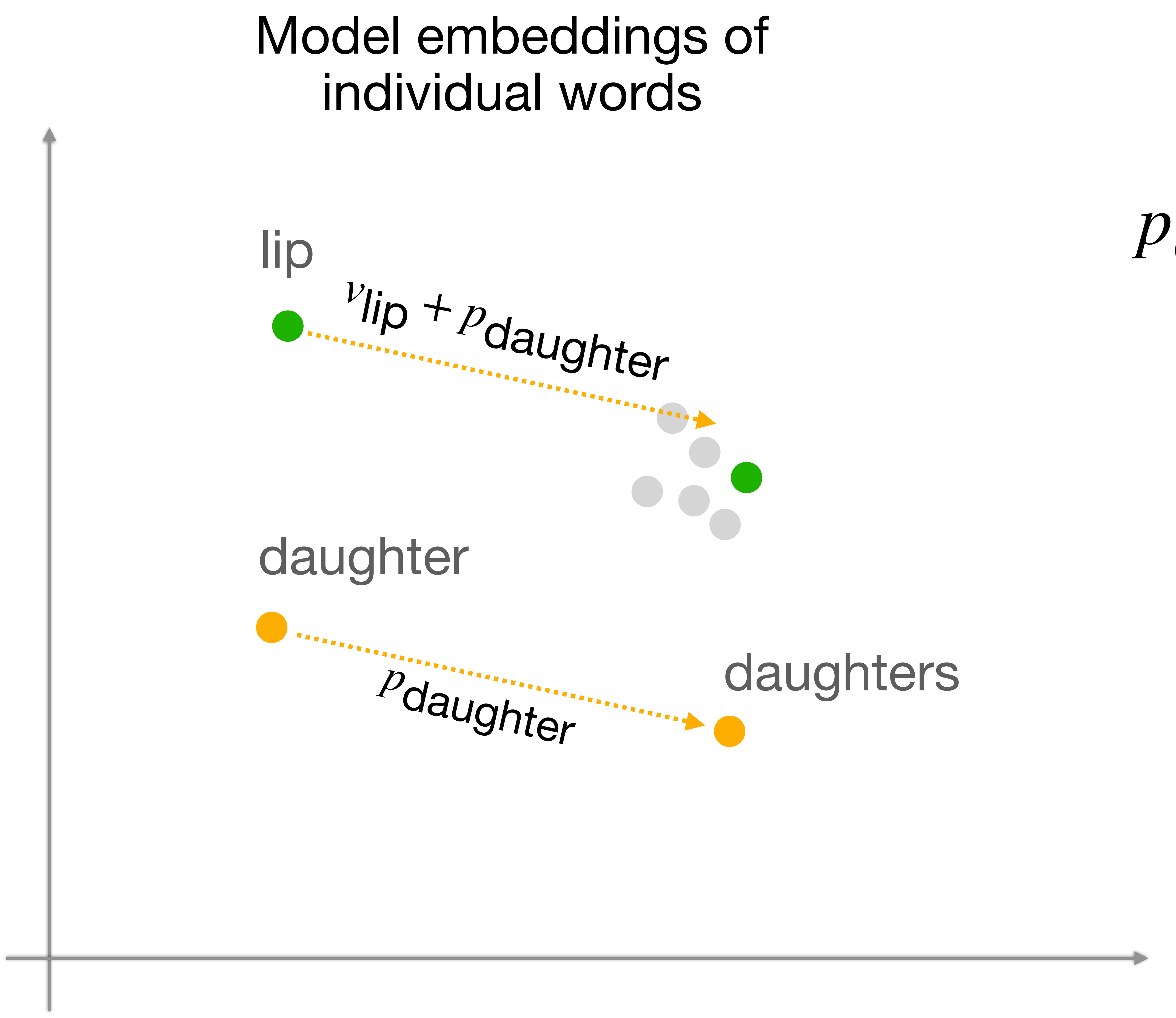
Compute analogy by vector algebra:

$$p_{\text{daughter}} = v_{\text{daughters}} - v_{\text{daughter}}$$

Mikolov et al. (2013)
Ethayarajh, Duvenaud & Hirst (2019)

Prediction

`daughter : daughters :: lip : _____`



Compute analogy by vector algebra:

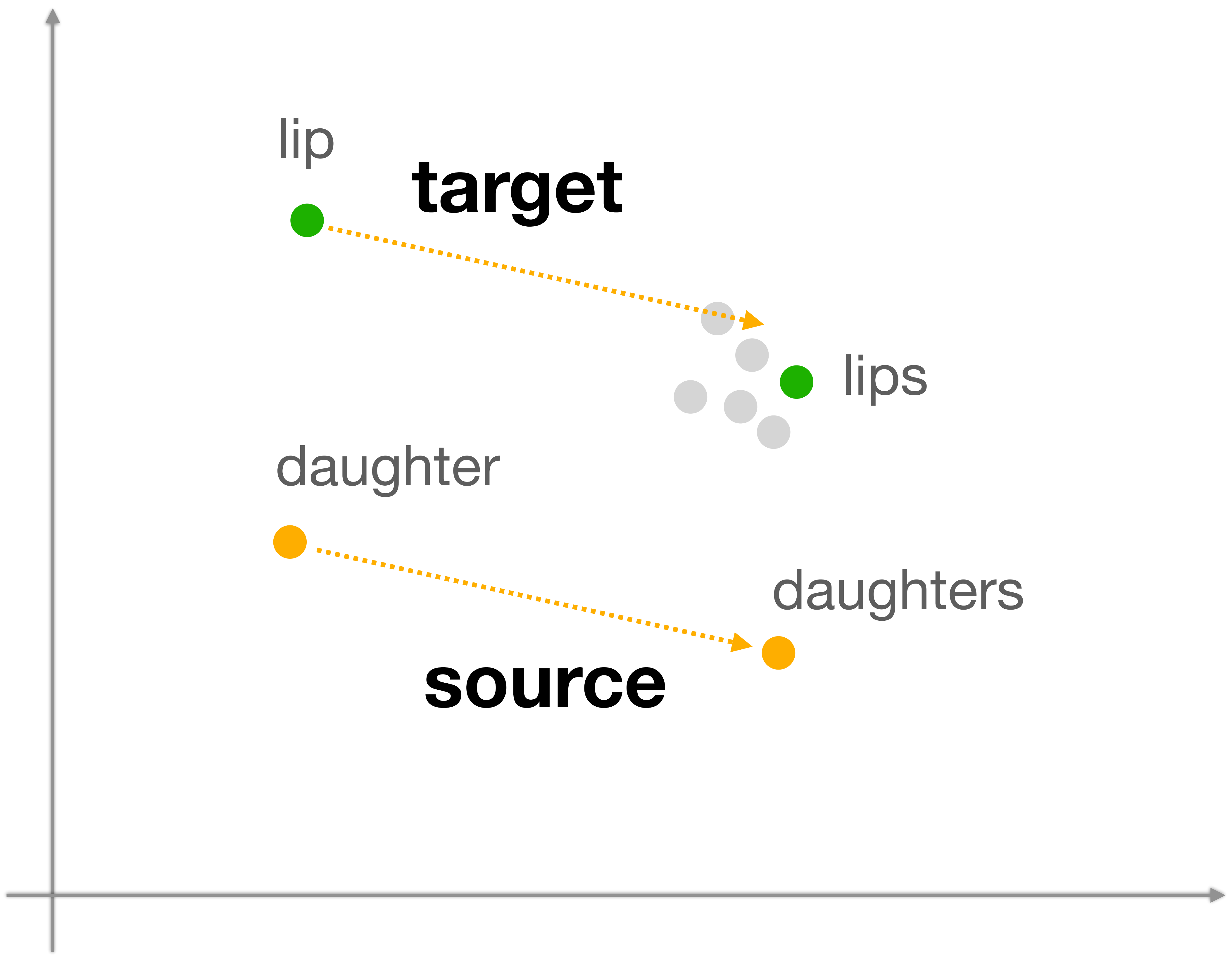
$$p_{\text{daughter}} = v_{\text{daughters}} - v_{\text{daughter}}$$

Rank evaluation:

Rank	Word
0	list
1	less
2	lips
3	lend

Mikolov et al. (2013)
Ethayarajh, Duvenaud & Hirst (2019)

Experimental questions



What is encoded in this translation?

- Is it a **morphological** transformation?
- Is it a **phonological** transformation?
- How does this vary in a model trained for word recognition?

`wav2vec` Audio-contrastive embedding	**Word** Word-contrastive embedding

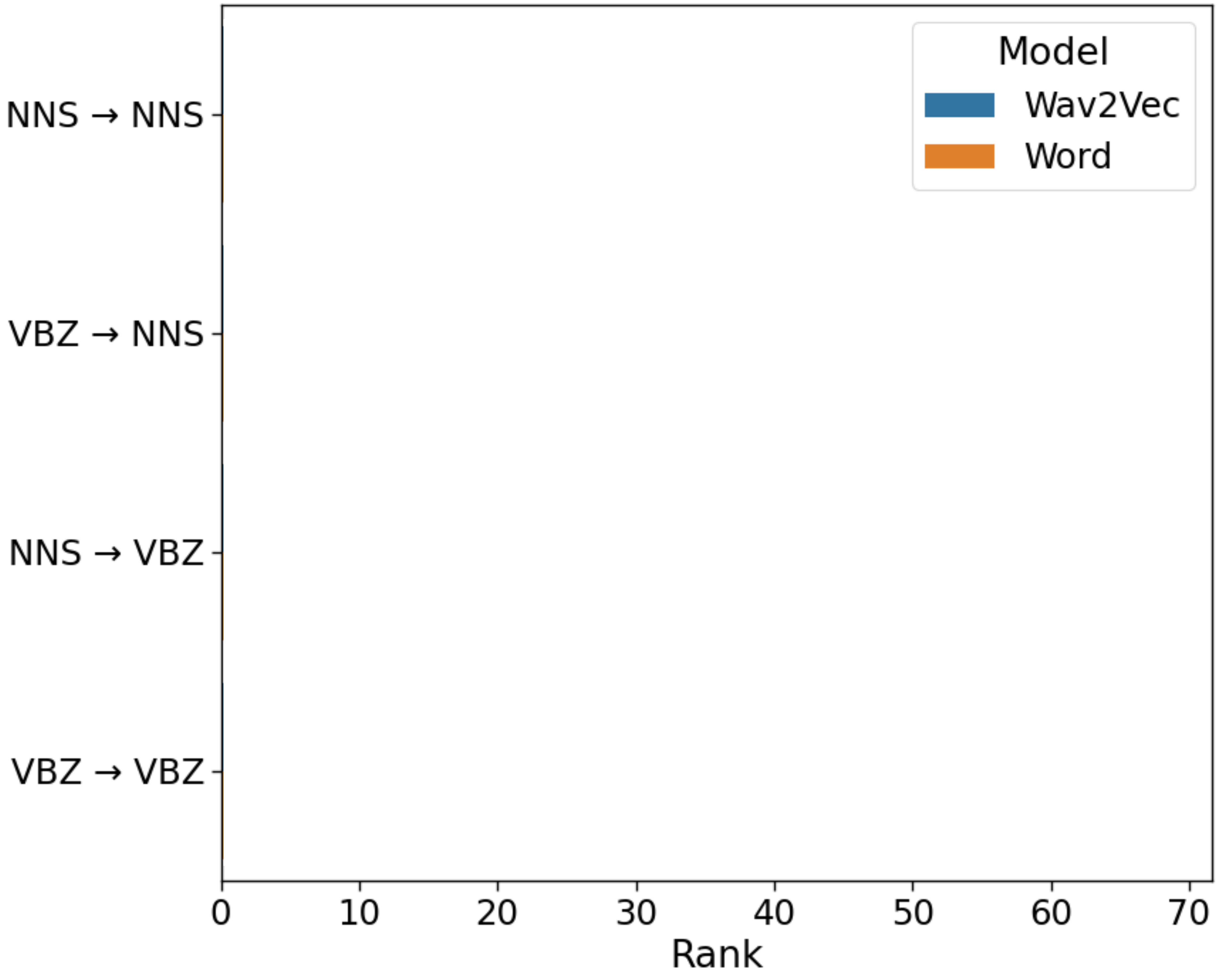
Is this a morphological transformation?

war : wars :: lip : ________
NNS NNS

speak : speaks :: lip : _____
VBZ NNS

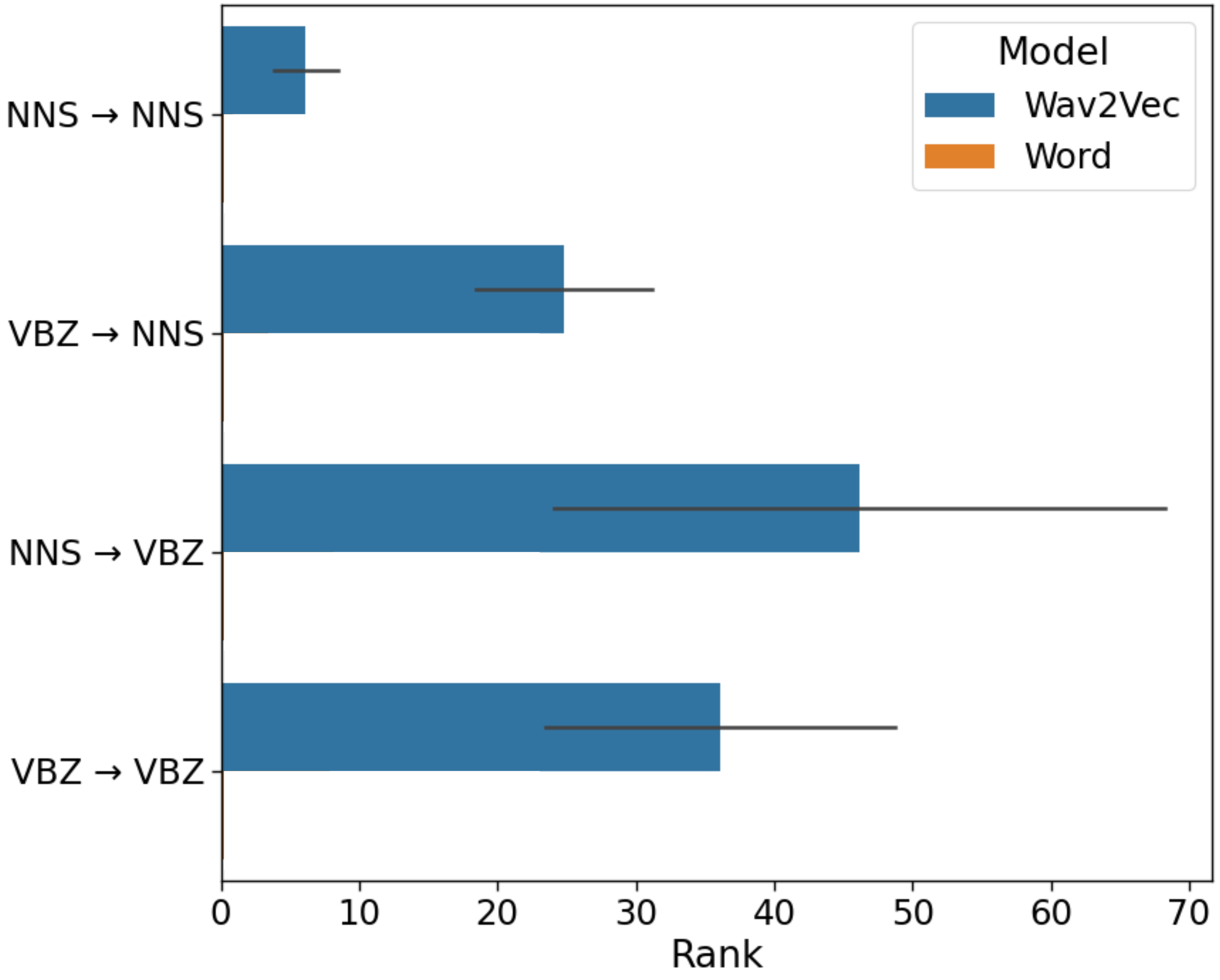
war : wars :: exist : _______
NNS VBZ

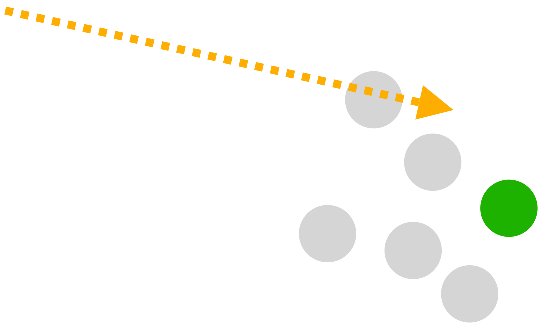
speak : speaks :: exist : ___
VBZ VBZ



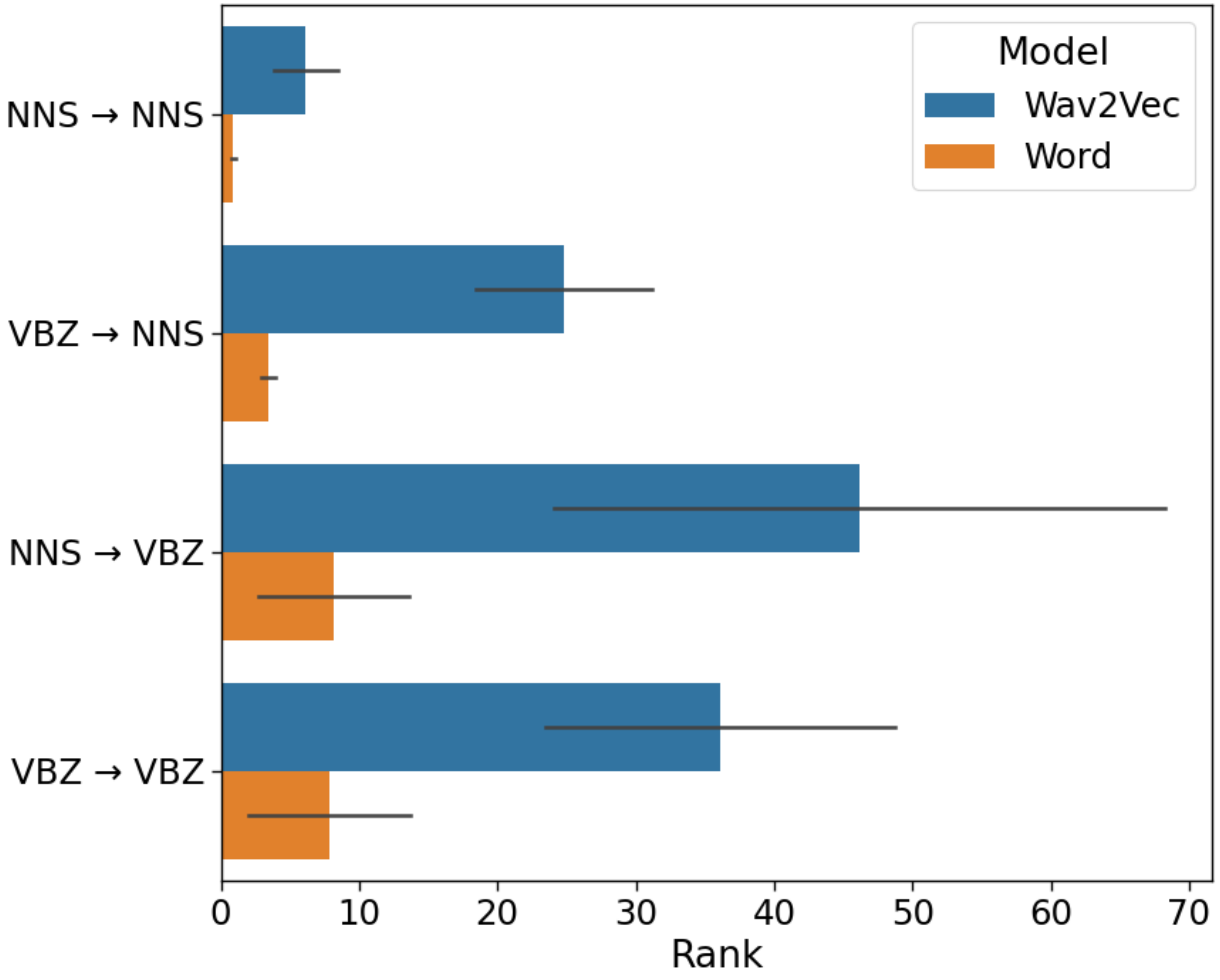
Wav2vec (audio-contrastive) model shows sensitivity to

morphological distinctions



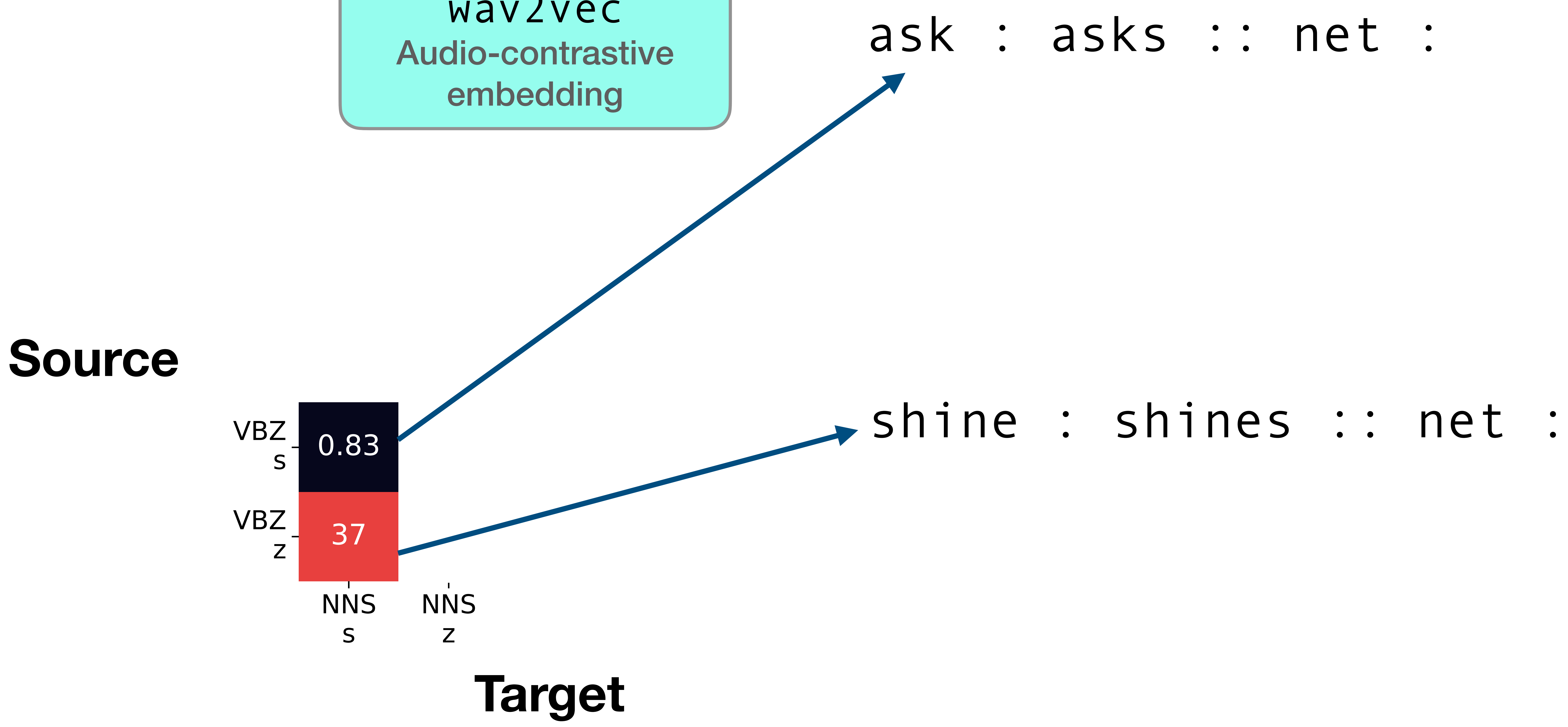


Word-contrastive model shows reduced sensitivity to **morphological** distinctions



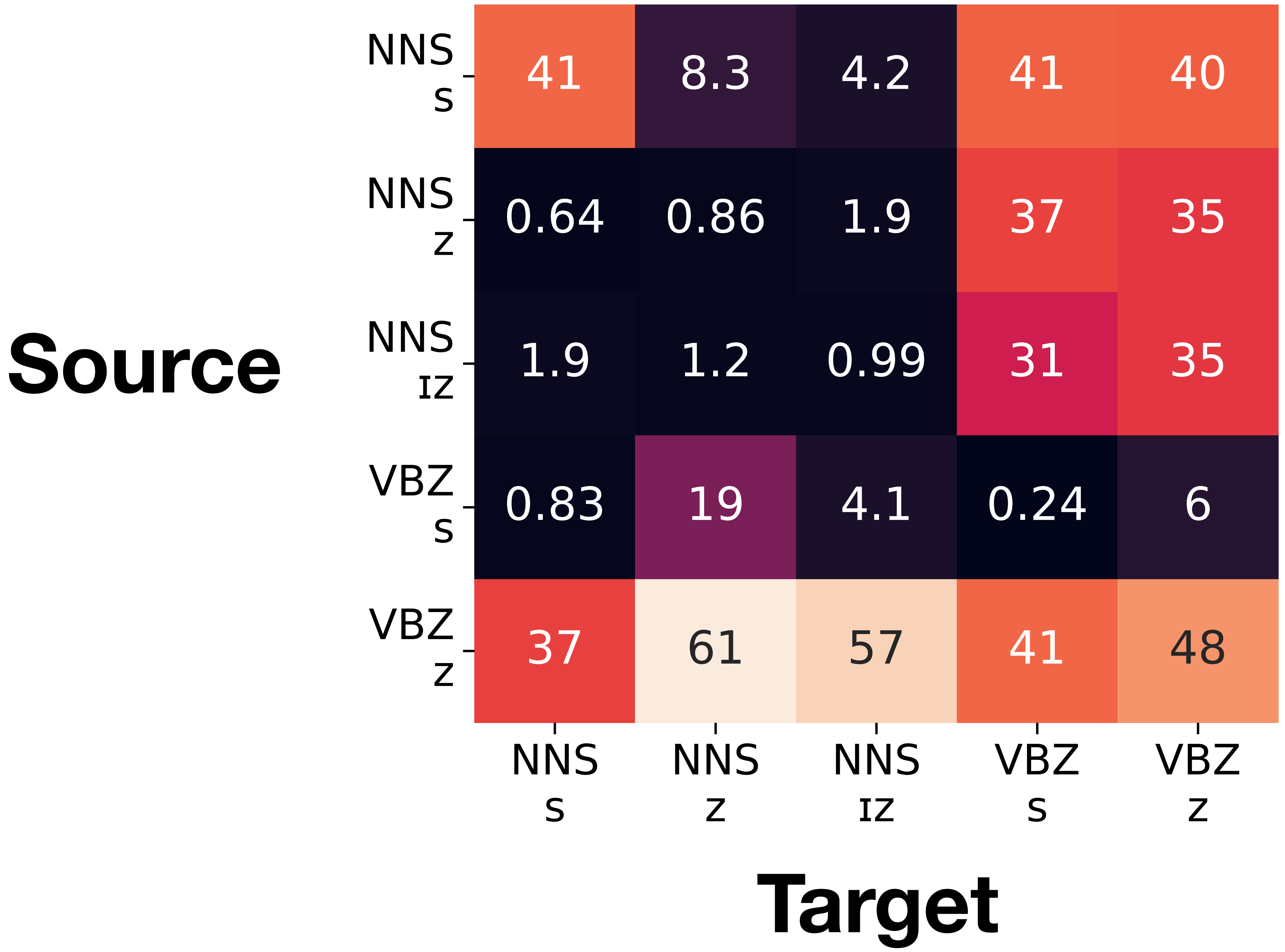
Is this a phonological transformation?

`wav2vec`
Audio-contrastive embedding

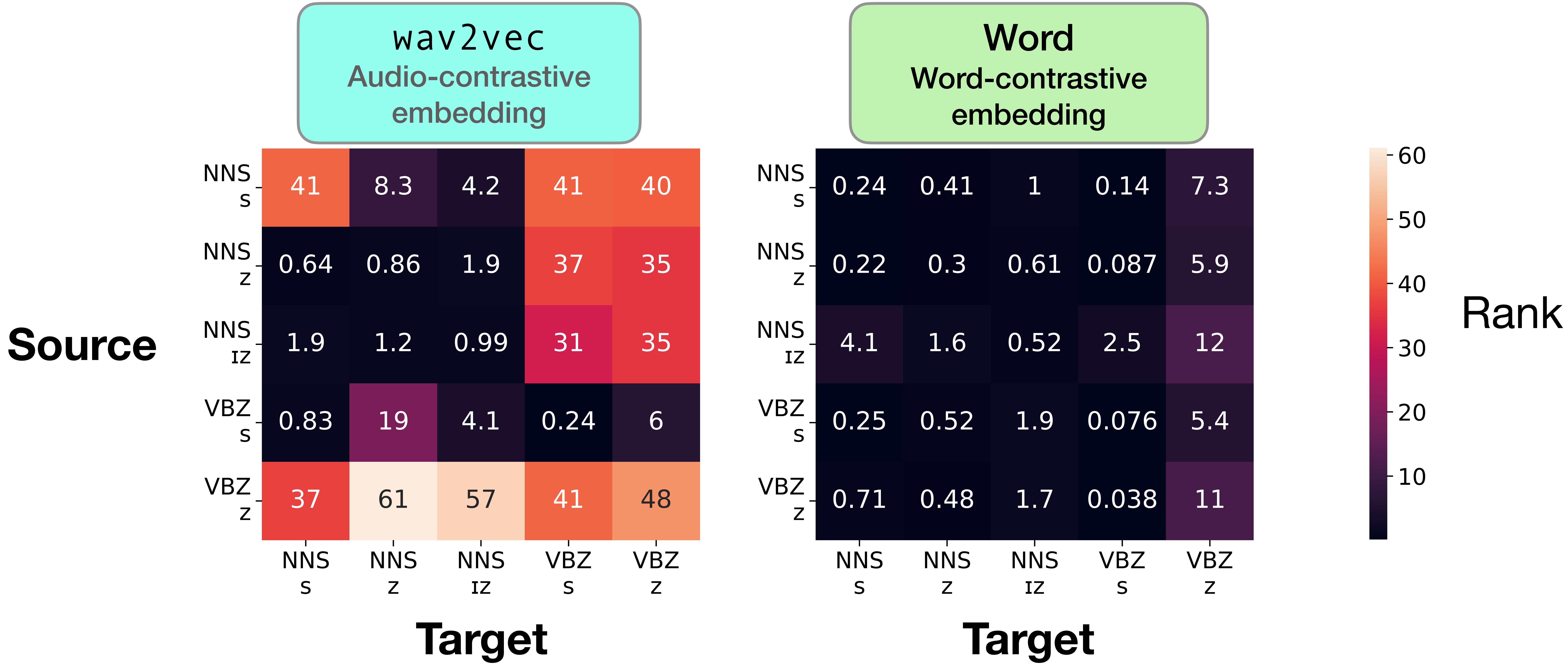


Is this a phonological transformation?

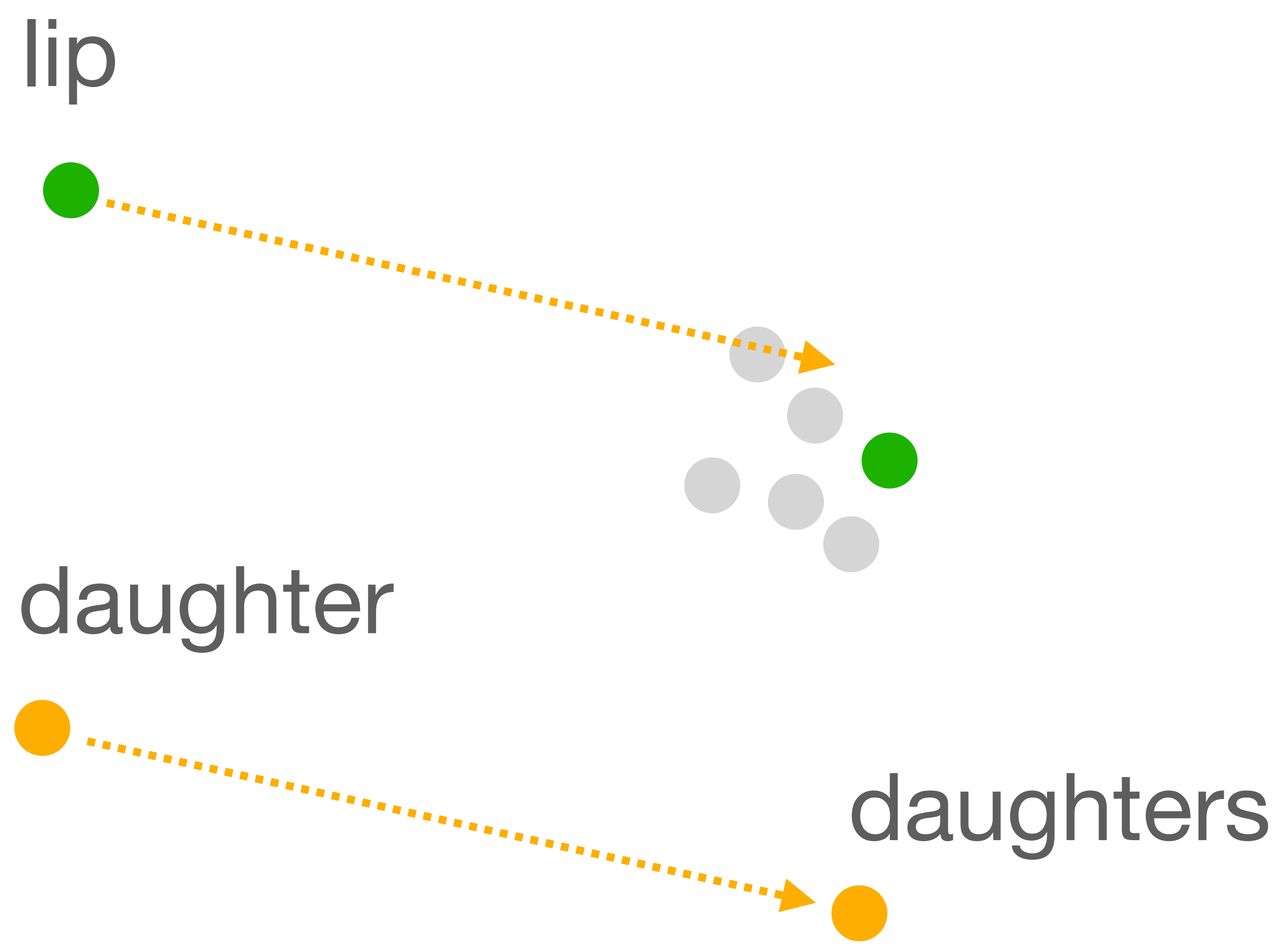
wav2vec
Audio-contrastive embedding



Is this a phonological transformation?



Interim summary



- wav2vec's representations are sensitive to both **morphological** (noun plurals vs. verbs) and **phonological** ([z], [s], [ɪz]) distinctions
- Optimizing for word recognition **minimizes** these distinctions
- What about cases where phonological distinctions matter?

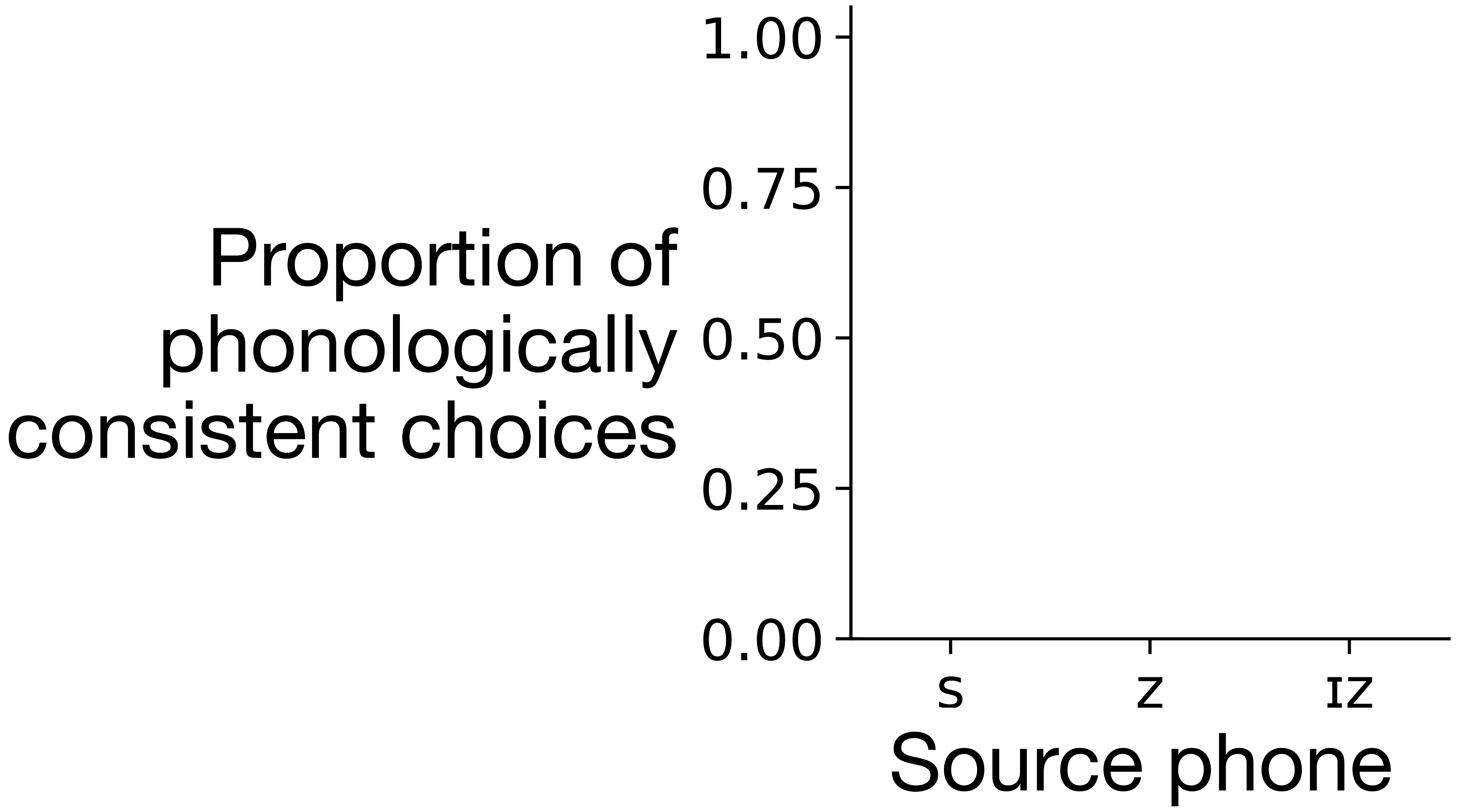
bay — **bays** — **base**

- Hypothesis: analogy maps to the **phonologically consistent** item

[z]
`own : owns` :: `bay` : { `bays` **(consistent)**, `base` **(inconsistent)** }

[s]
`lip : lips` :: `bay` : { `bays` **(consistent)**, `base` **(inconsistent)** }

Phonological consistency



A direction in model space

encodes a **phonological rule**:

Add the phonologically consistent choice of [z], [s], [ɪz], as in noun plurals and verb inflections

[z]

`own : owns :: bay :` { `bays` **(consistent)** / `base` **(inconsistent)** }

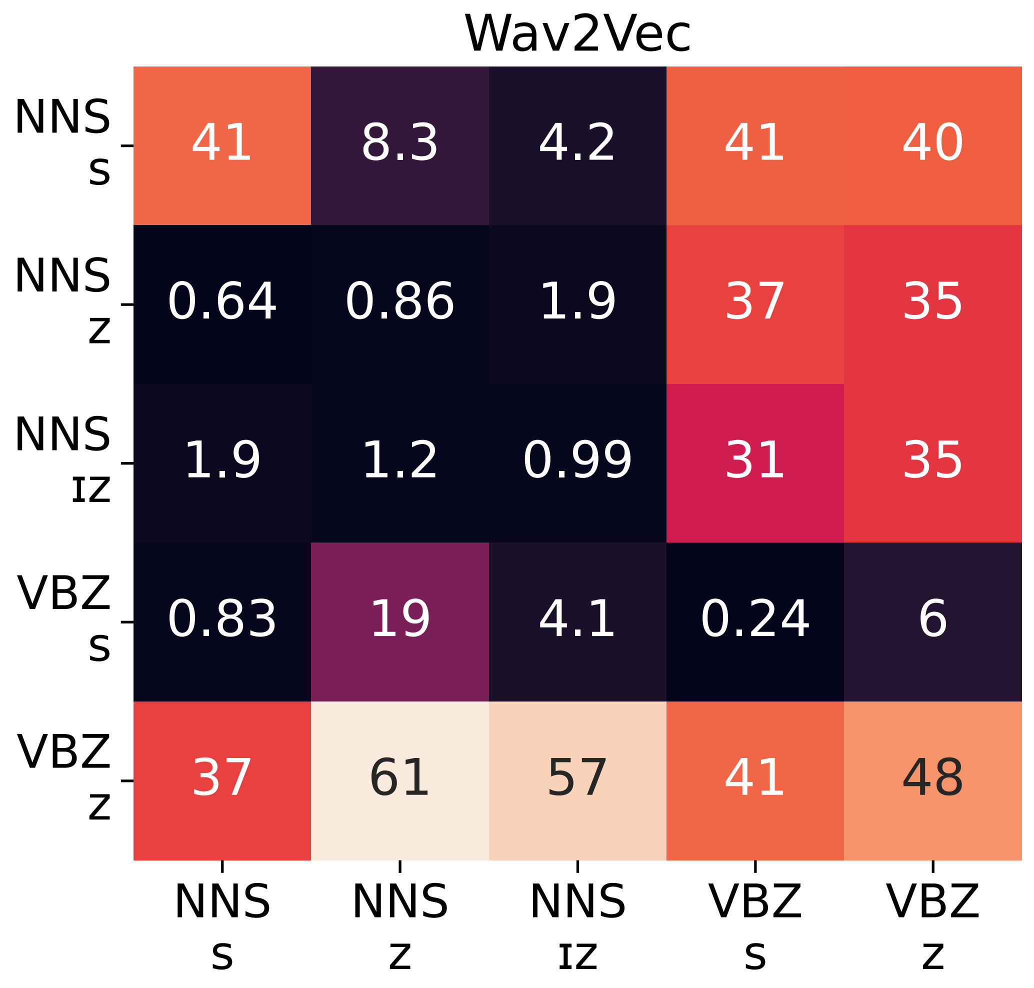
Conclusion: for modelers

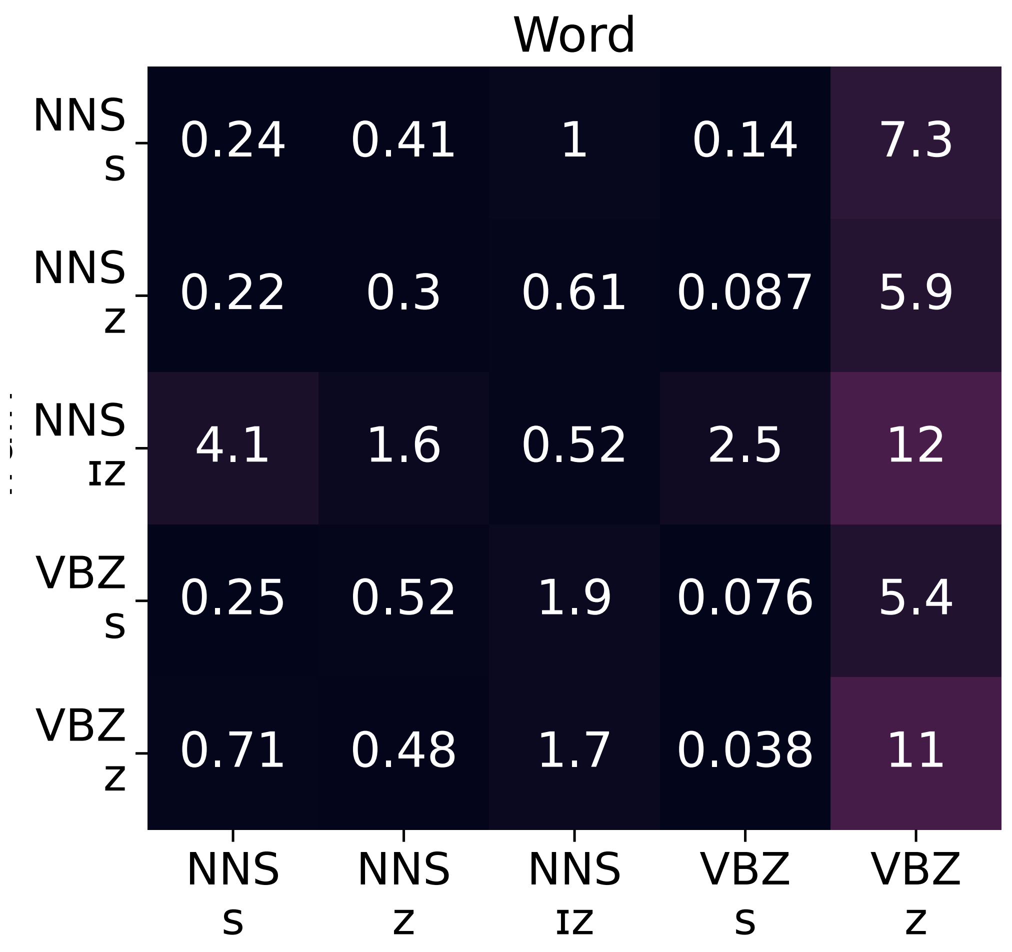
Unconstrained task

“What representations do speech models use?”

Specific task

“What representations do speech models use for spoken word recognition?”

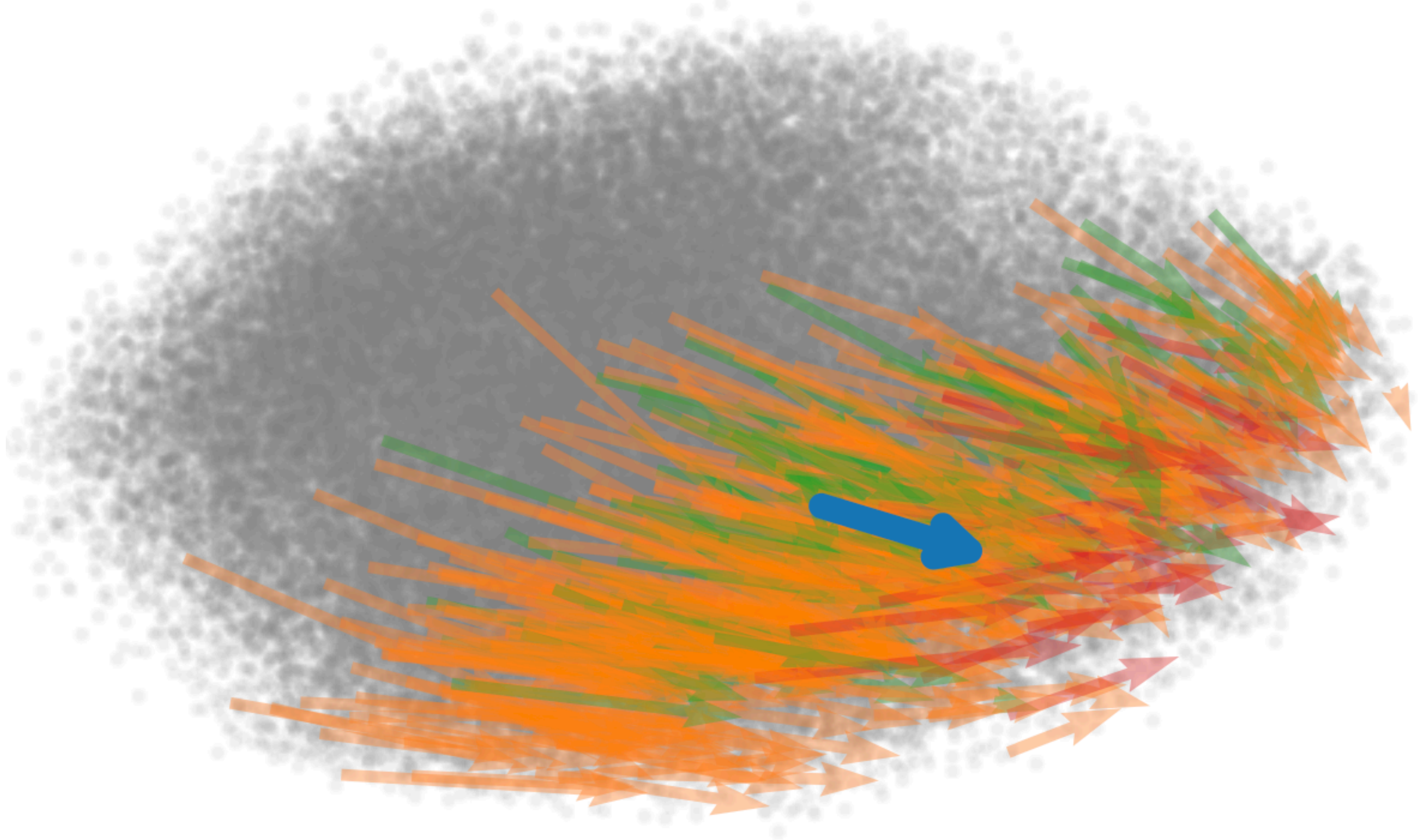




- An optimal word recognition model tracks the **phonological rules** involved in noun and verb inflections using a **simple geometric relationship**

- This is an abstract computation, bridging phonology and morphology

- Next: use these findings to design predictions about the neural implementation of speech comprehension



Canaan Breiss

Matt Leonard

Edward Chang

USC UCSF