Emergent morpho-phonological representations in self-supervised speech models

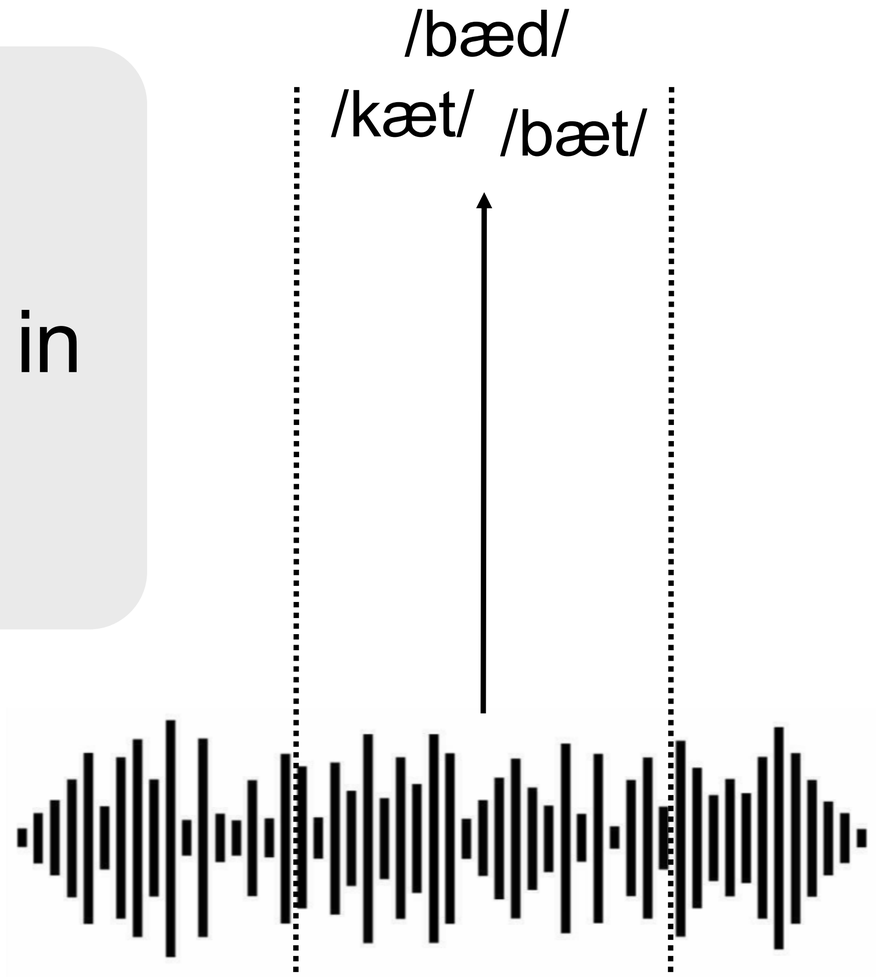
Jon Gauthier[1] Canaan Breiss[2]
Matthew Leonard[1] Edward Chang[1]

[1] UCSF [2] USC

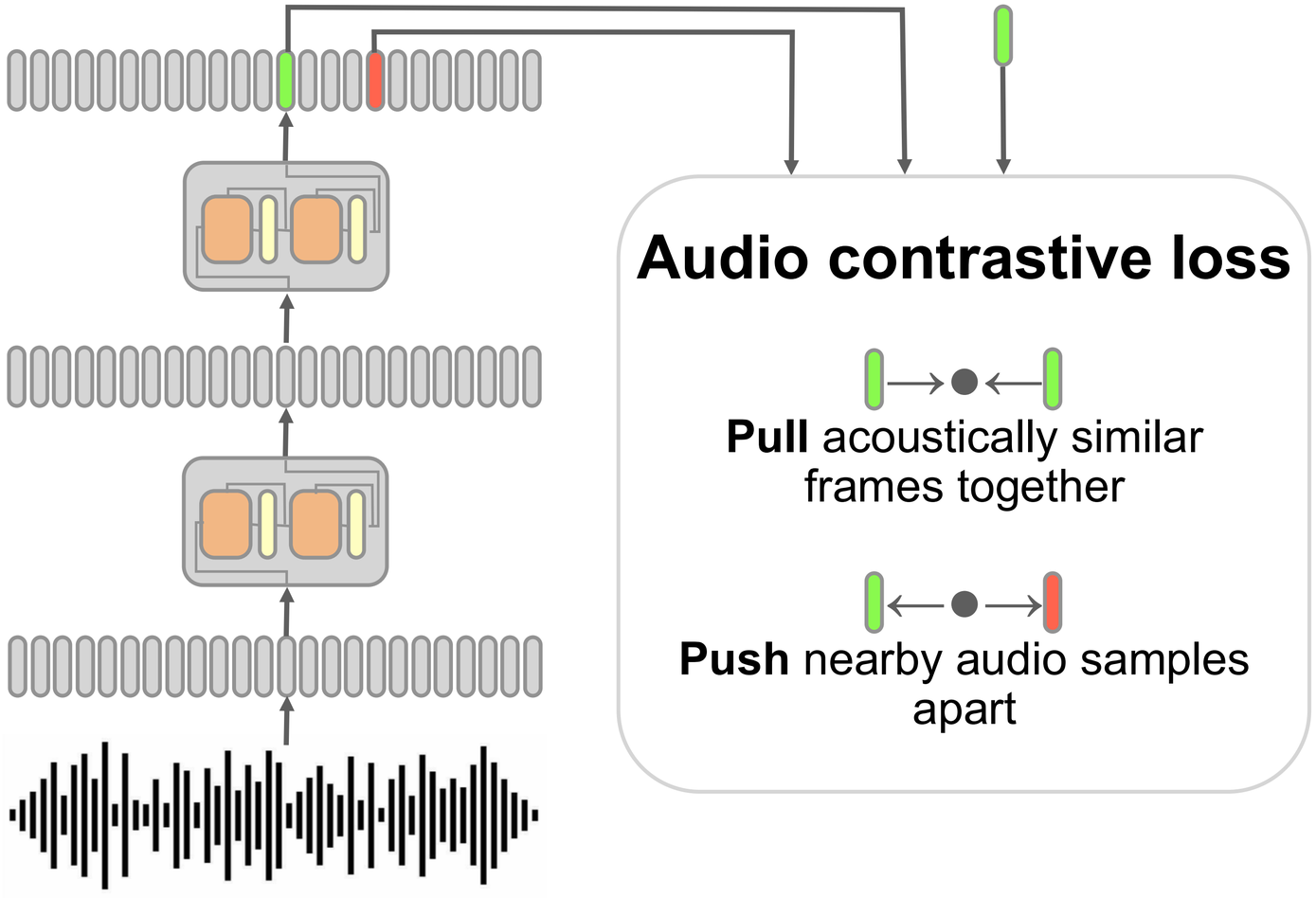
EMNLP 2025
Suzhou, China

Spoken word recognition

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

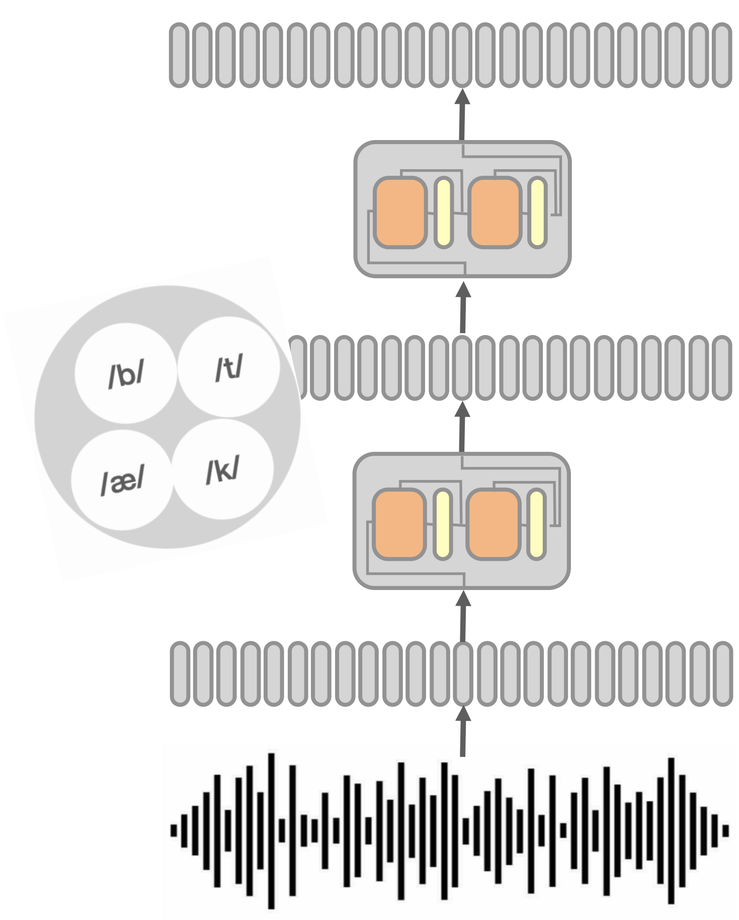


A self-supervised model: wav2vec2



Baevski et al. (2020)

A self-supervised model: wav2vec2

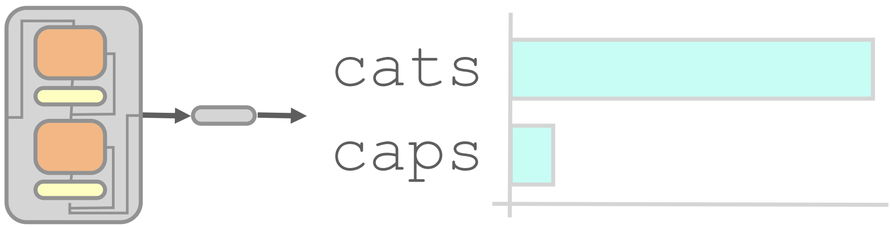


- Self-supervised models encode basic phonological categories
- **… but these may serve many functions beyond word recognition**

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

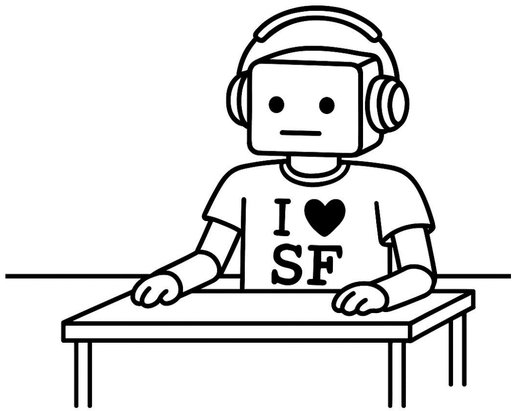
Plan

Model



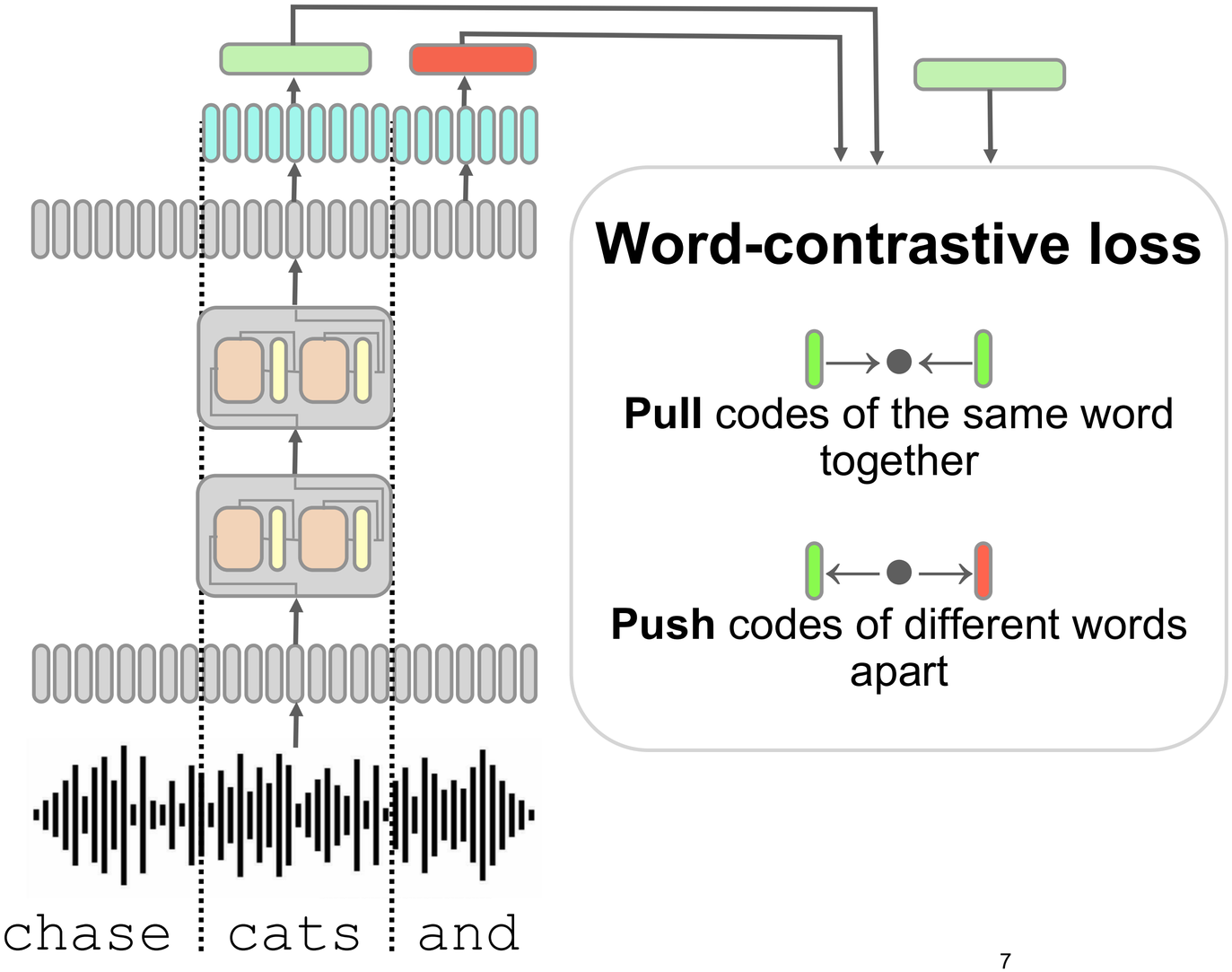
Derive a **word recognition** model from a self-supervised model

Experiment



Dissect its computations by treating it as an experimental subject

Word recognition model



Word-contrastive loss

Pull codes of the same word together

Push codes of different words apart

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

wav2vec
Audio-contrastive embedding

Word
Word-contrastive embedding

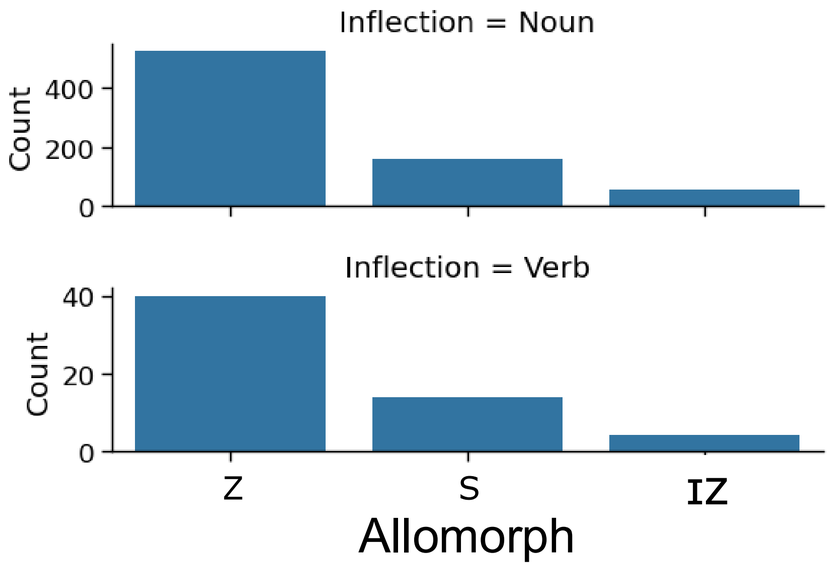
Phenomenon

- 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

	Allomorph	Base	Inflected
Noun plural (NNS)	/z/	daughter	daughters
	/s/	lip	lips
	/ɪz/	age	ages
Verb 3rd-person singular (VBZ)	/z/	bring	brings
	/s/	speak	speaks
	/ɪz/	please	pleases

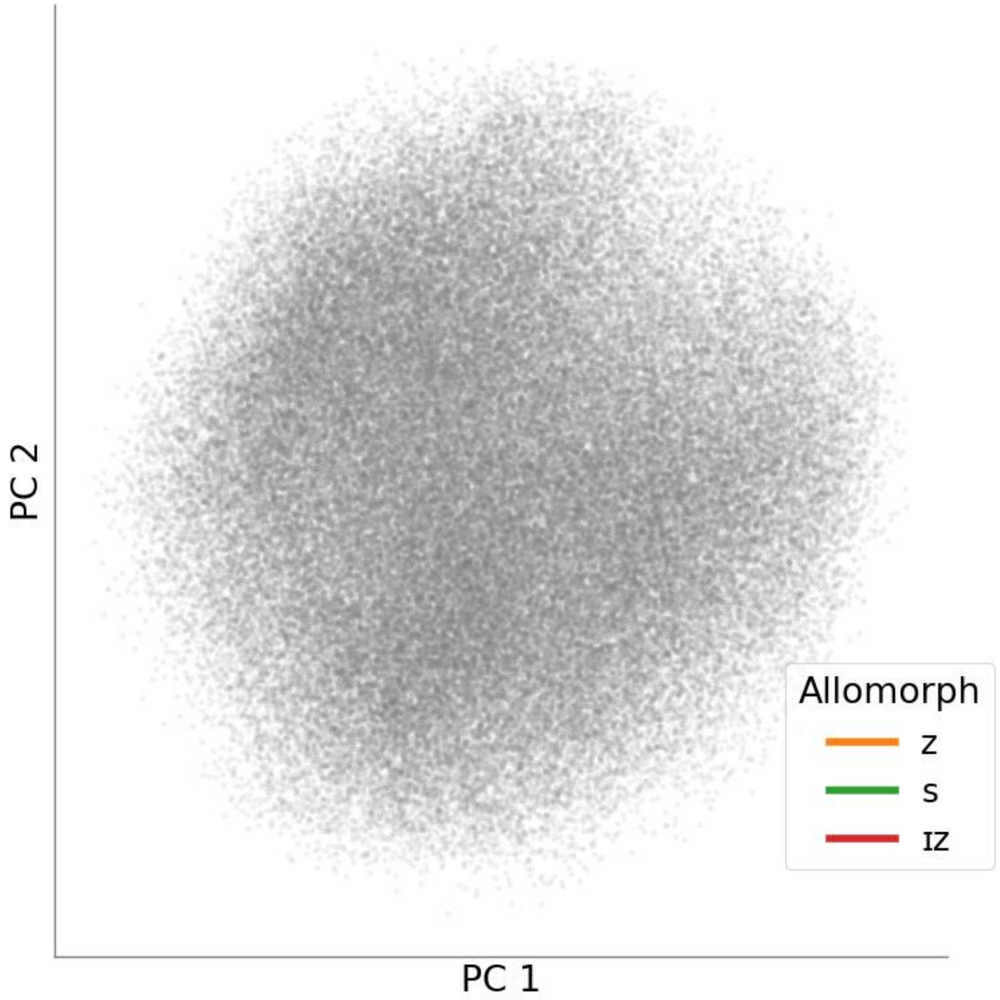
Corpus

- 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



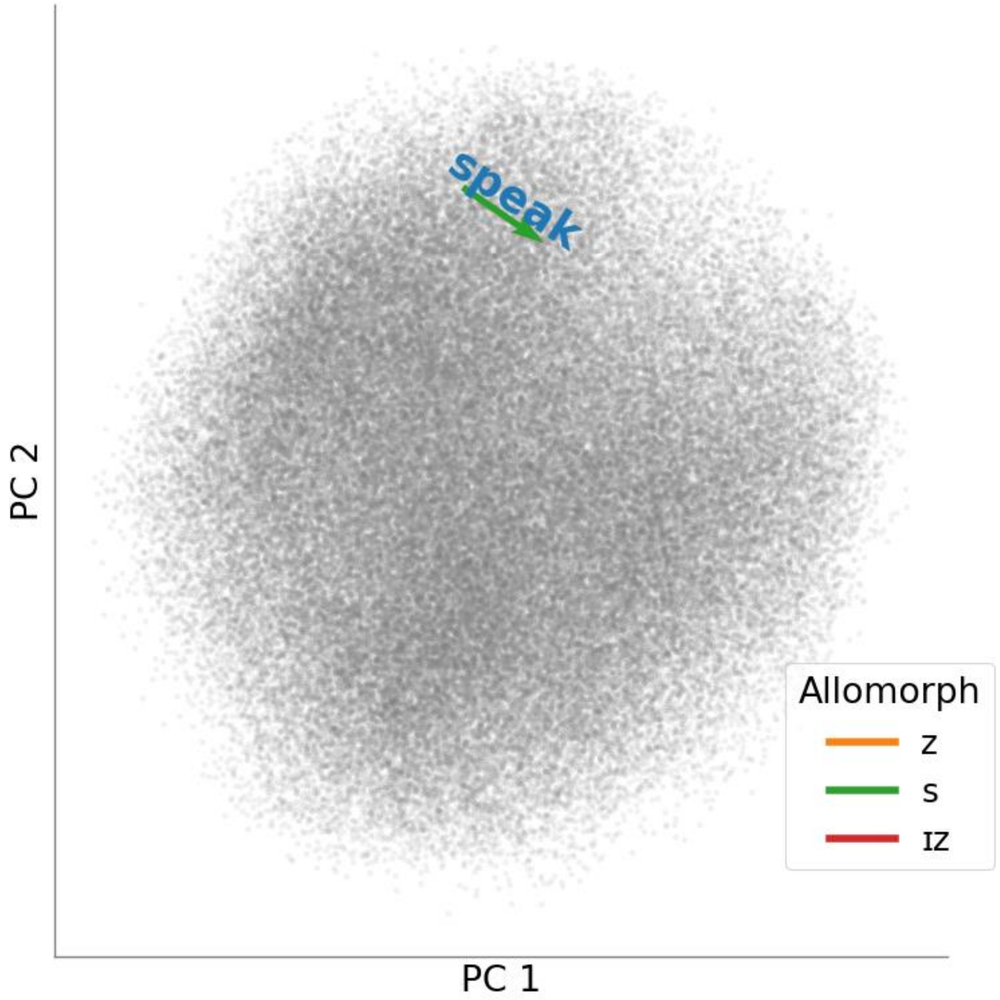
Global linear geometry

Word
Word-contrastive embedding



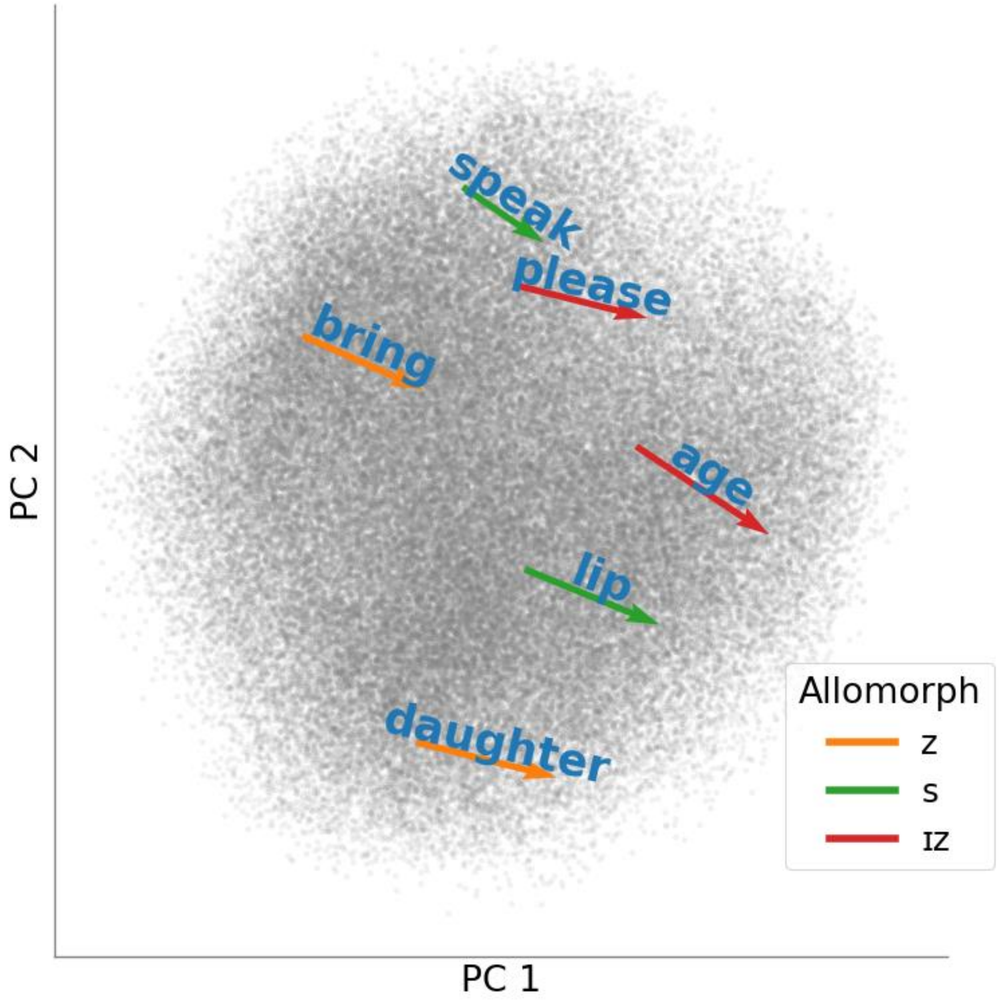
Global linear geometry

Word
Word-contrastive embedding



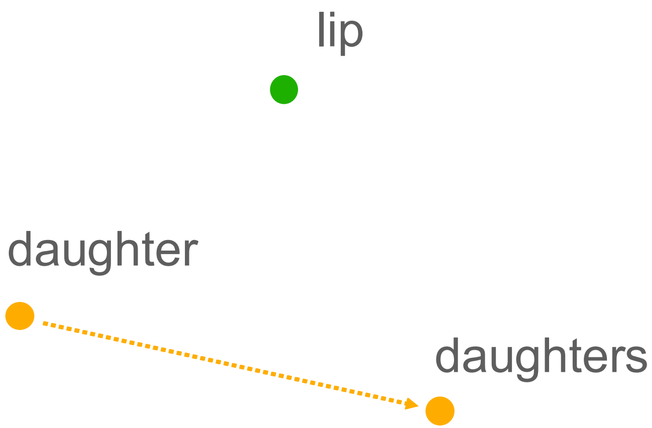
Global linear geometry

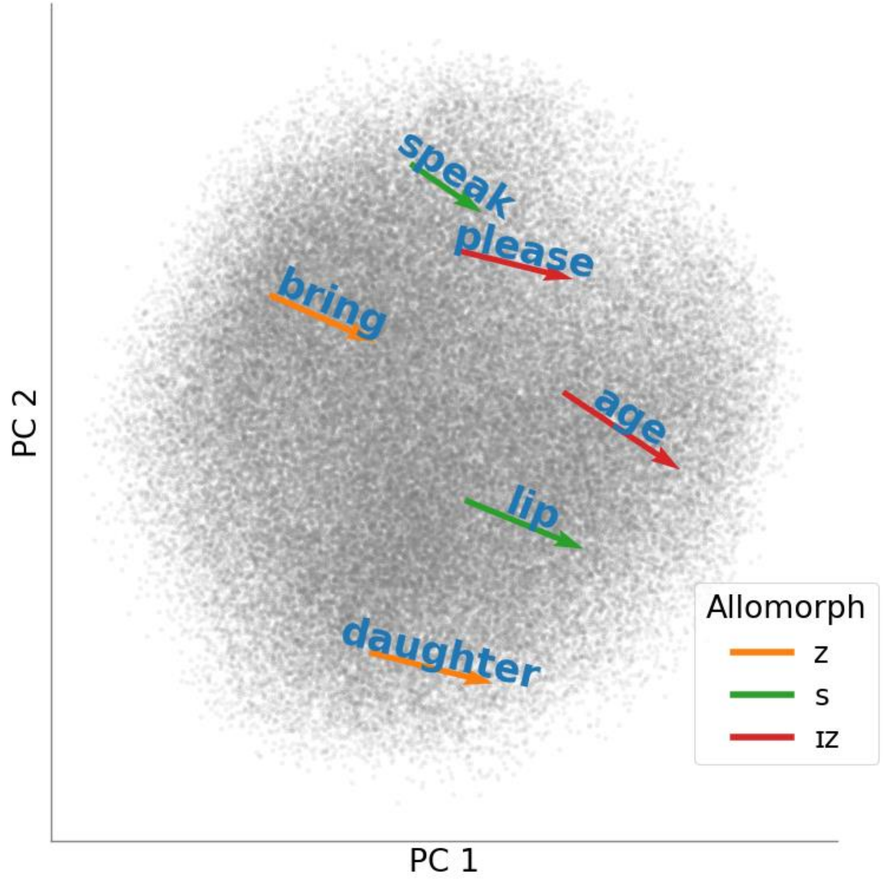
Word
Word-contrastive embedding



Hypothesis

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

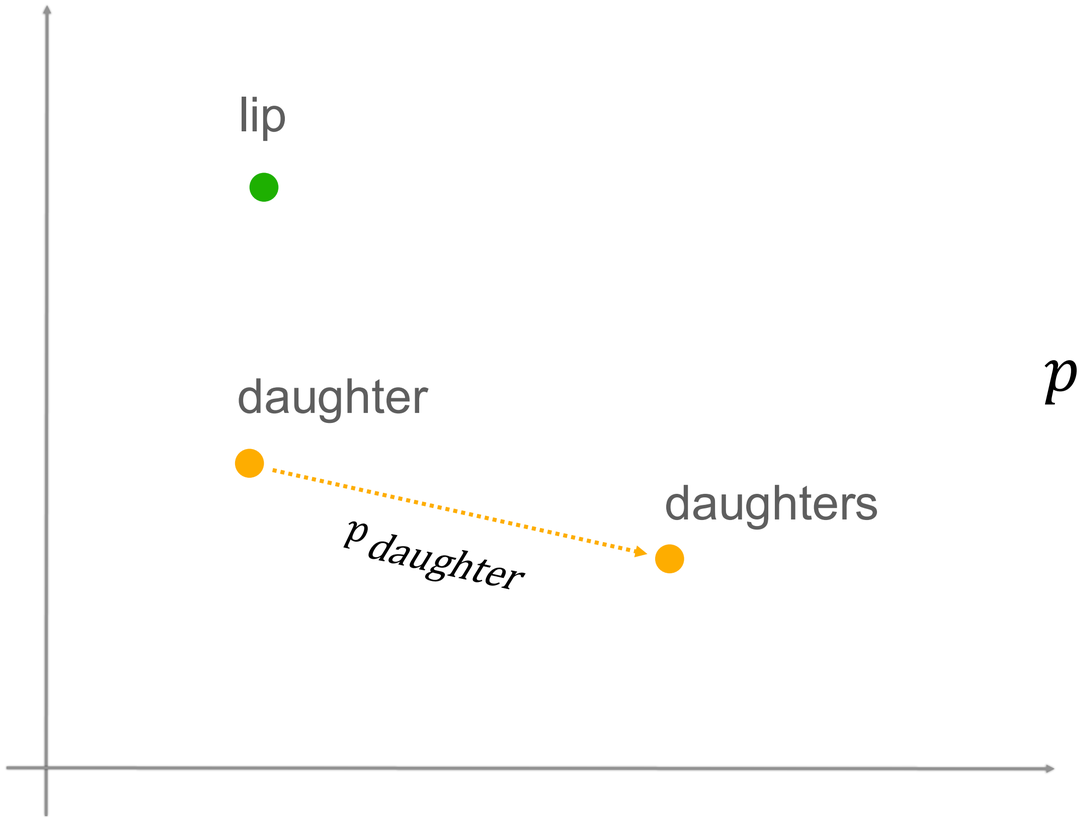




Prediction

daughter : daughters :: lip : _____

Model embeddings of individual words



Compute analogy by vector algebra:

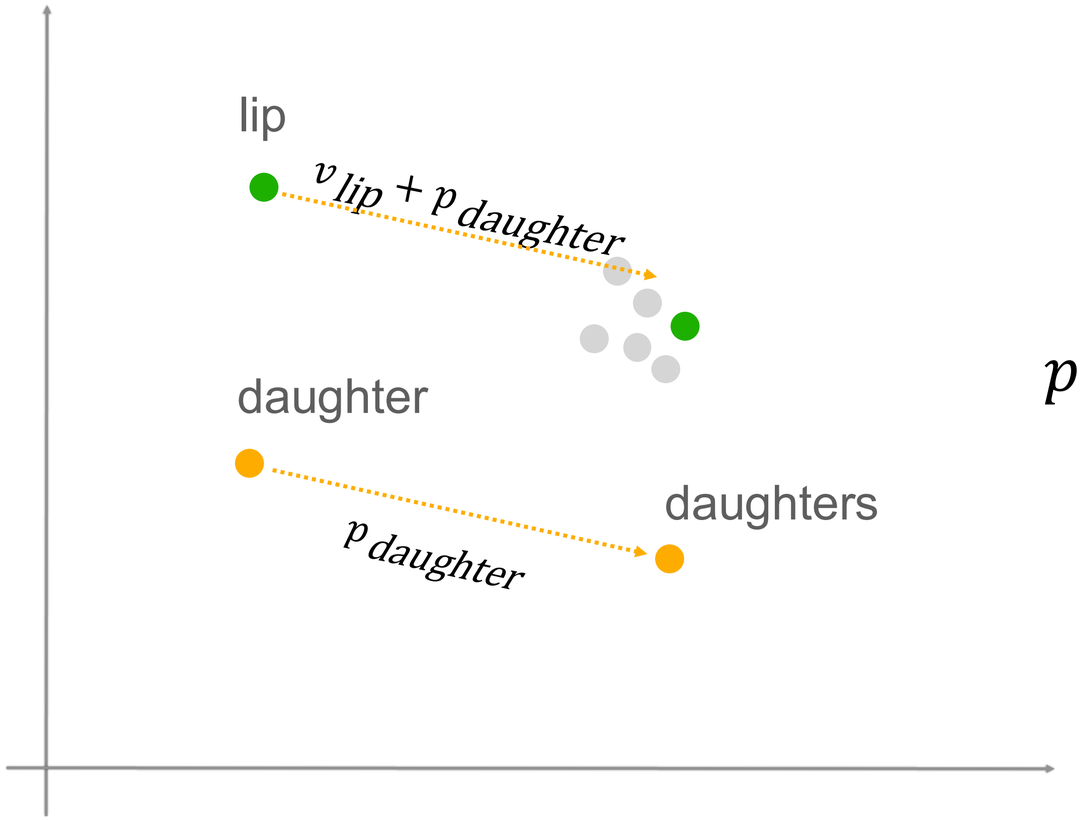
$$p_{daughter} = v_{daughters} - v_{daughter}$$

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

Prediction

daughter : daughters :: lip : _____

Model embeddings of individual words



Compute analogy by vector algebra:

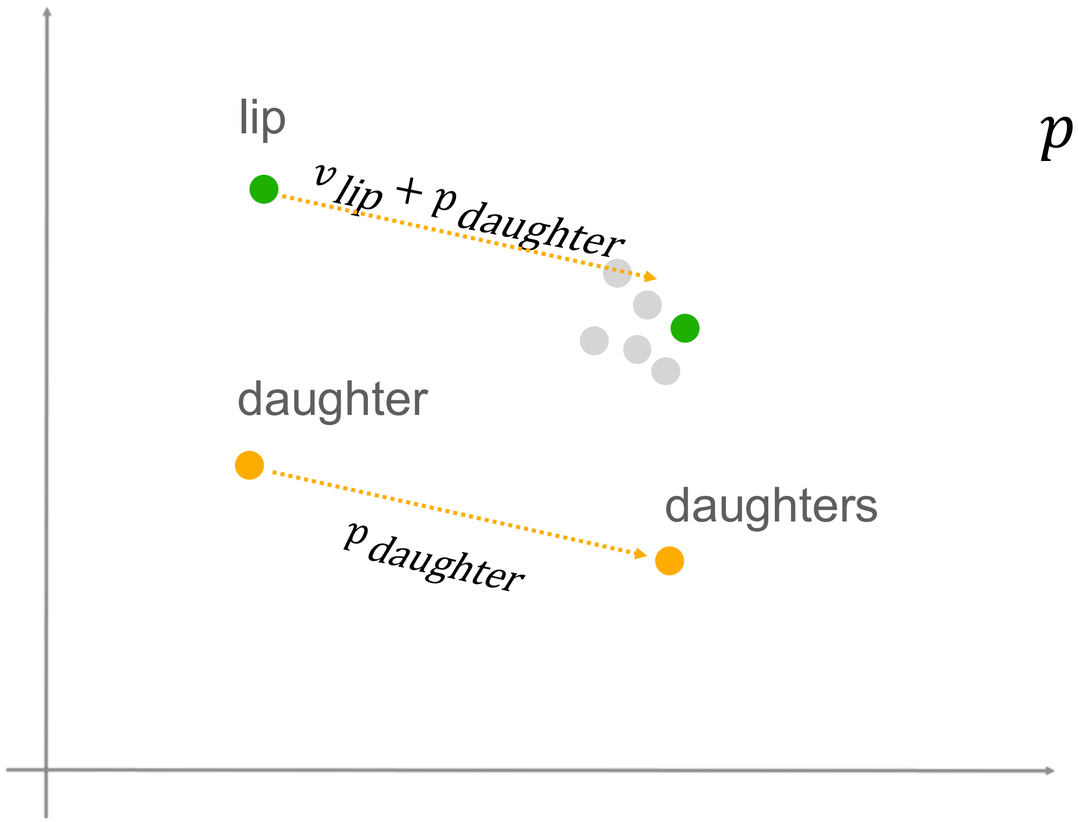
$$p_{daughter} = v_{daughters} - v_{daughter}$$

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

Prediction

daughter : daughters :: lip : _____

Model embeddings of individual words



Compute analogy by vector algebra:

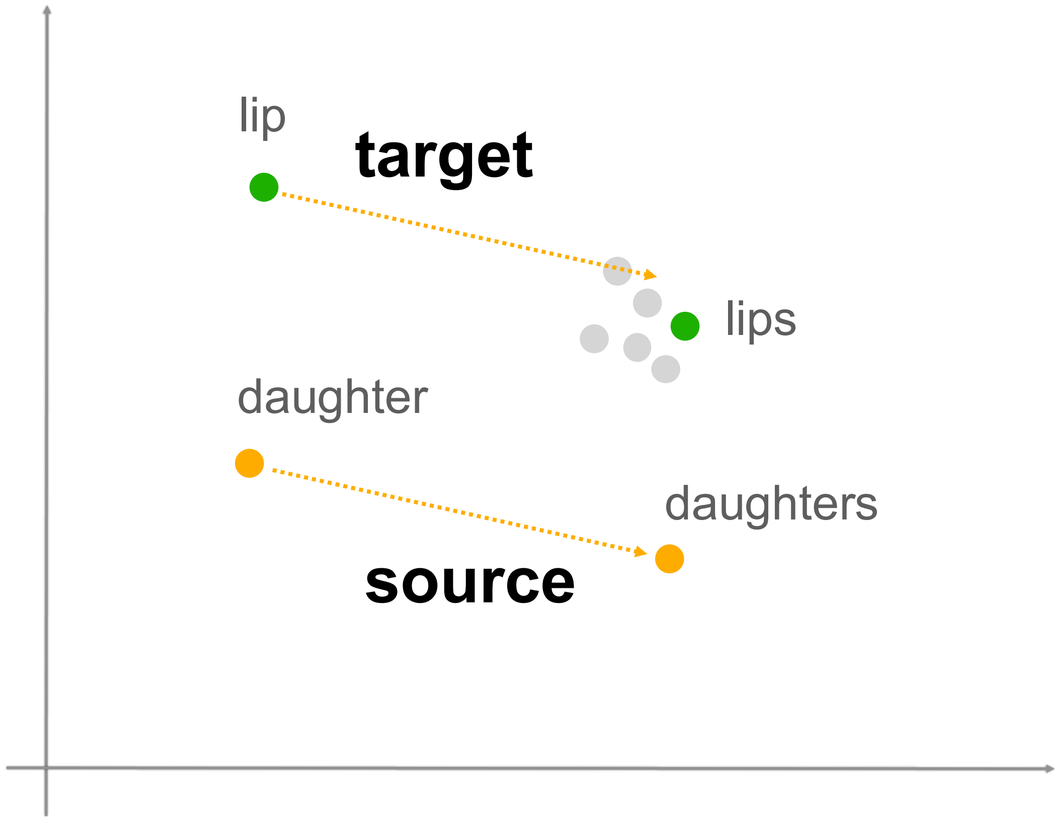
$$p_{daughter} = v_{daughters} - v_{daughter}$$

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	Word
Audio-contrastive embedding	Word-contrastive embedding

Is this a morphological transformation?

war : wars :: lip : _____

Noun Noun

speak : speaks :: lip : _____

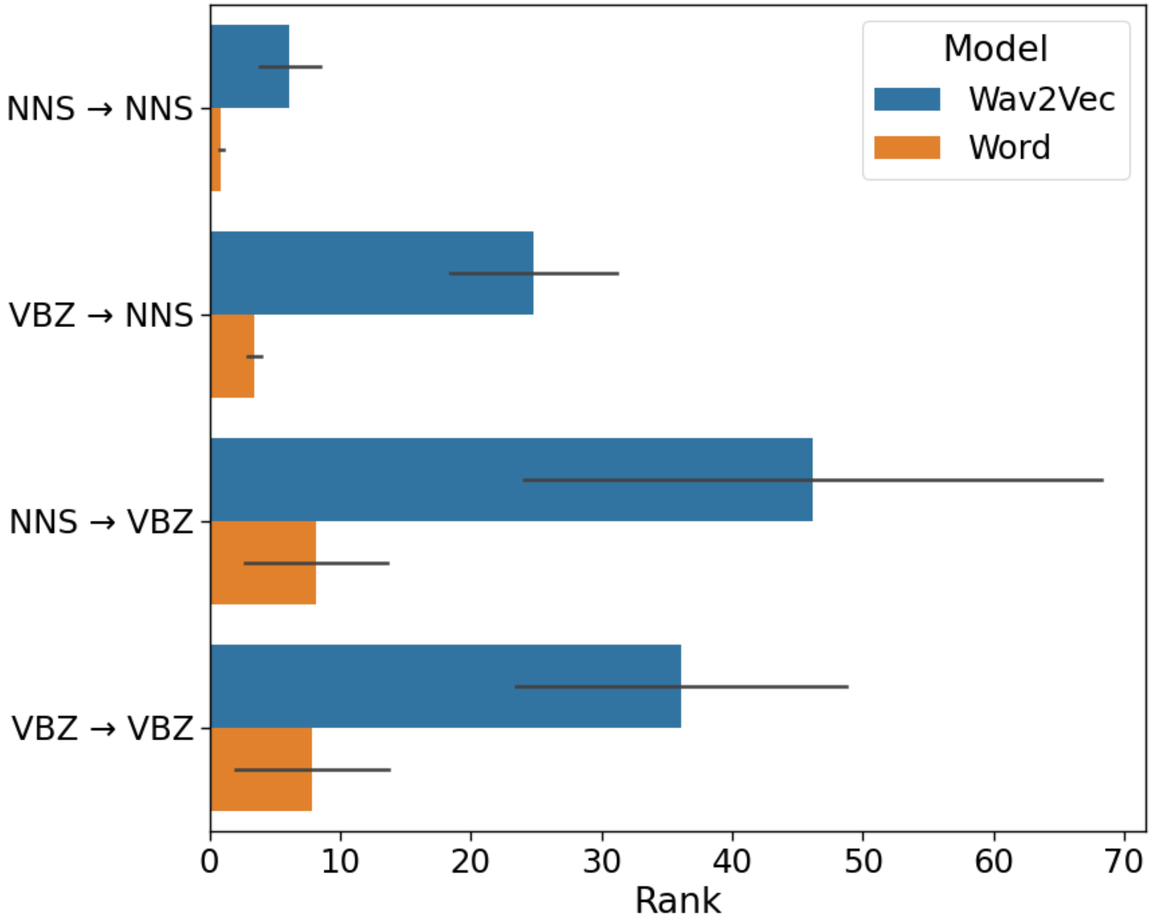
Verb Noun

war : wars :: exist : _____

Noun Verb

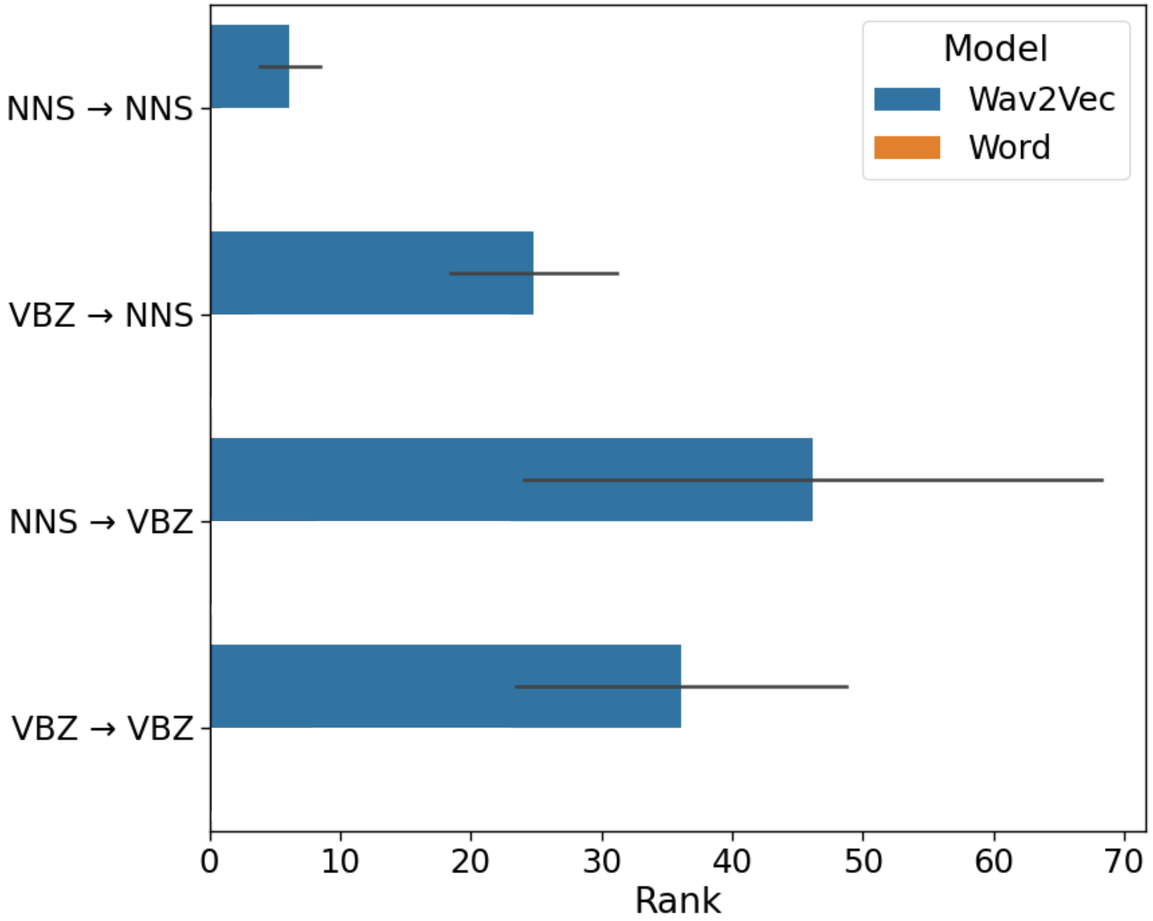
speak : speaks :: exist : ____

Verb Verb



Wav2vec (audio-contrastive) model shows sensitivity to

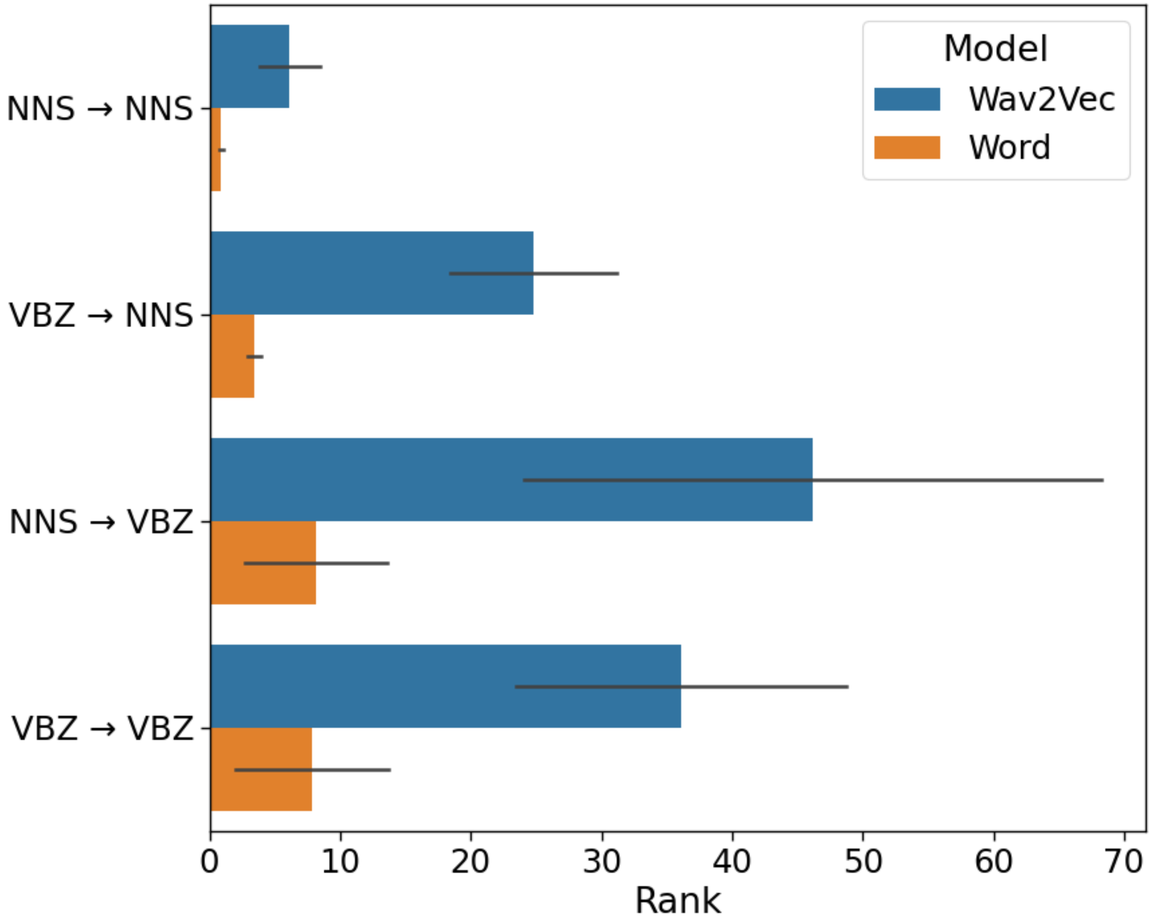
morphological distinctions



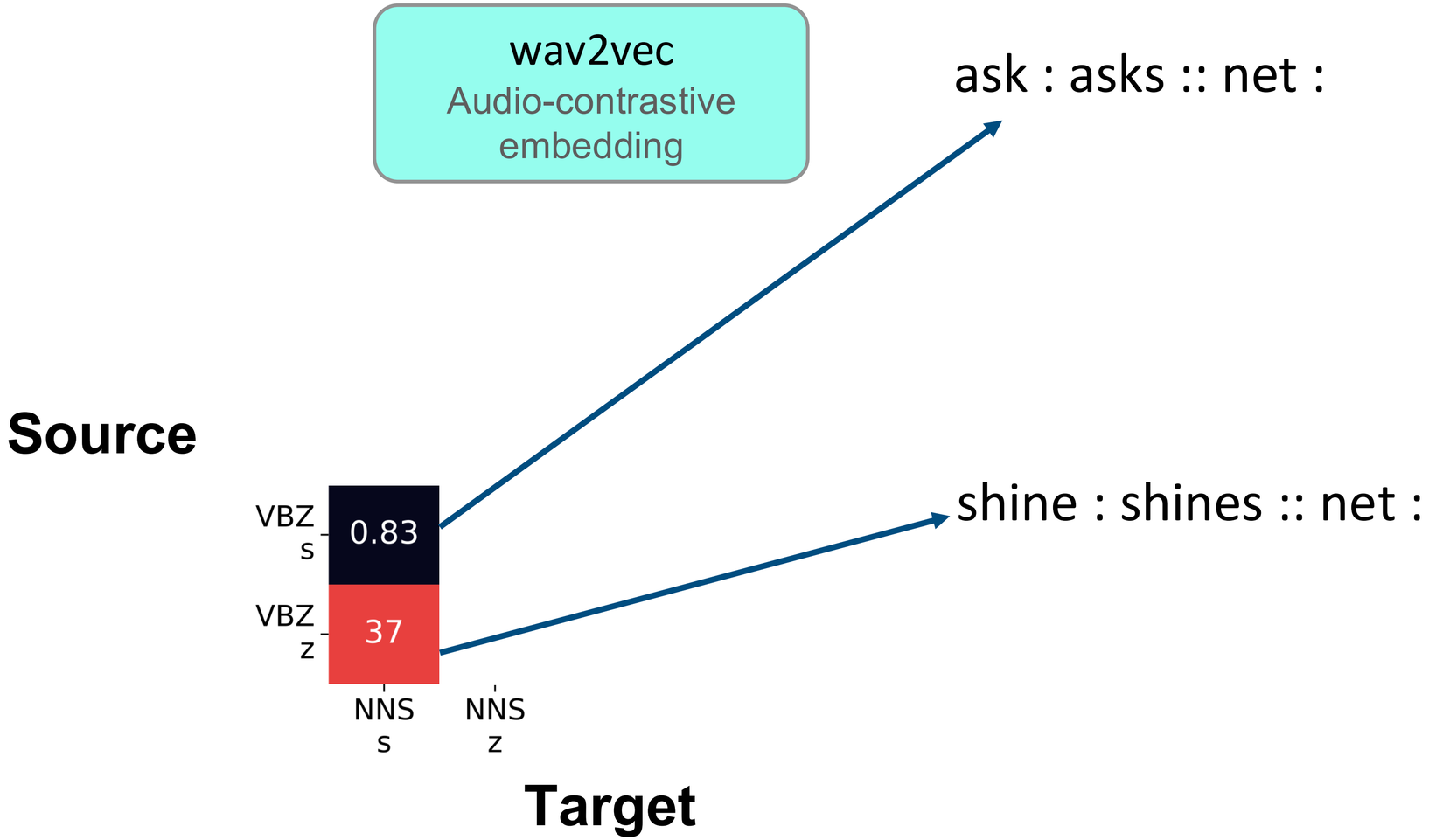
Word-contrastive model shows

reduced sensitivity to

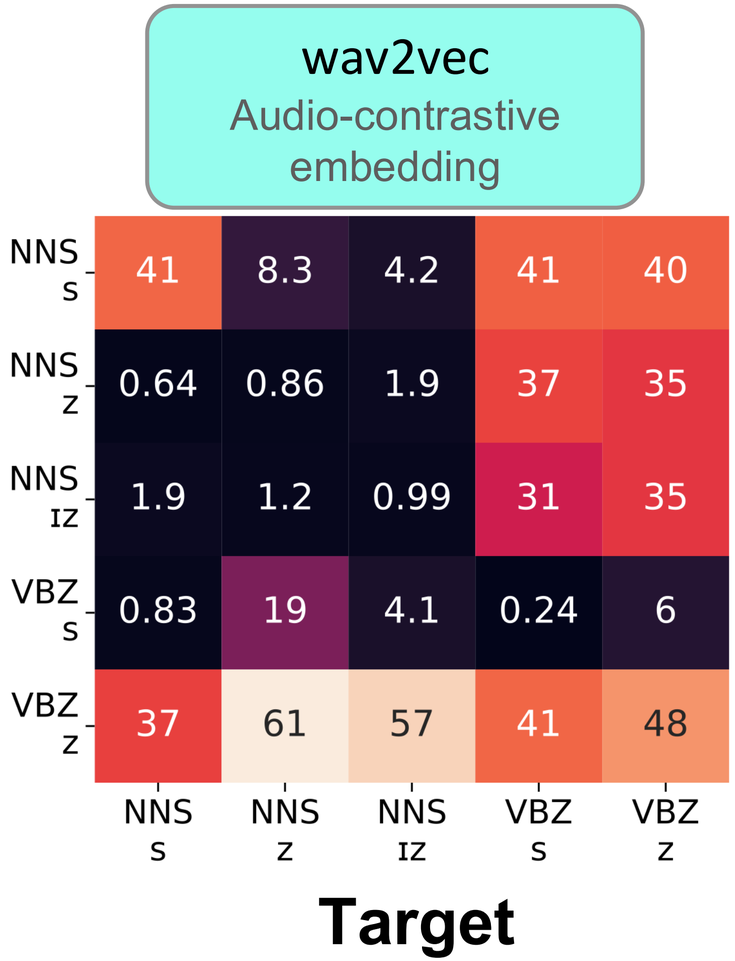
morphological distinctions



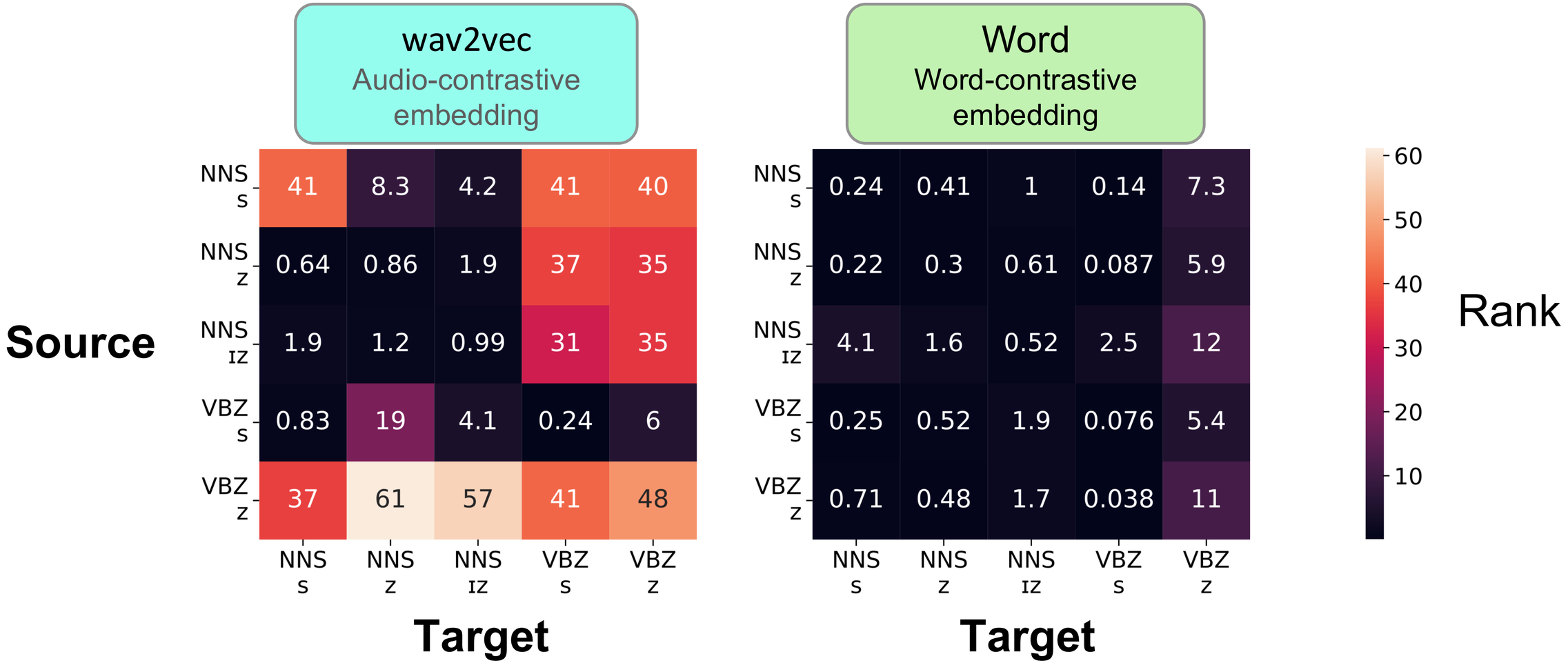
Is this a phonological transformation?



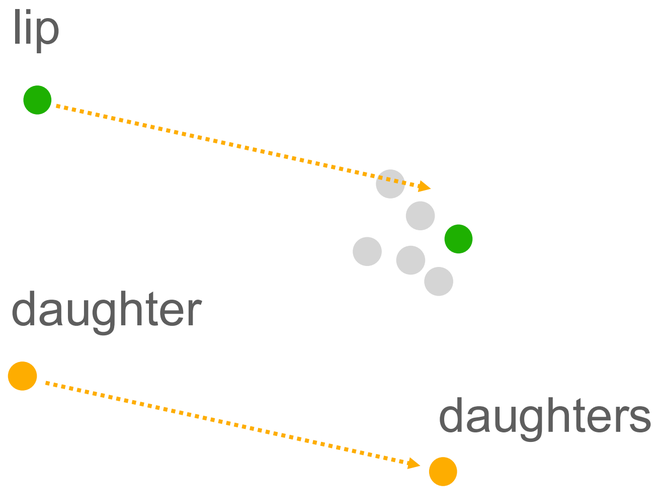
Is this a phonological transformation?



Is this a phonological transformation?

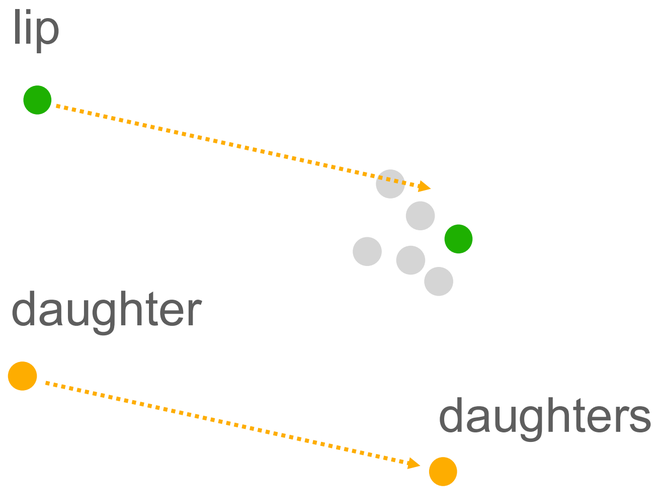


Interim summary



- 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

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

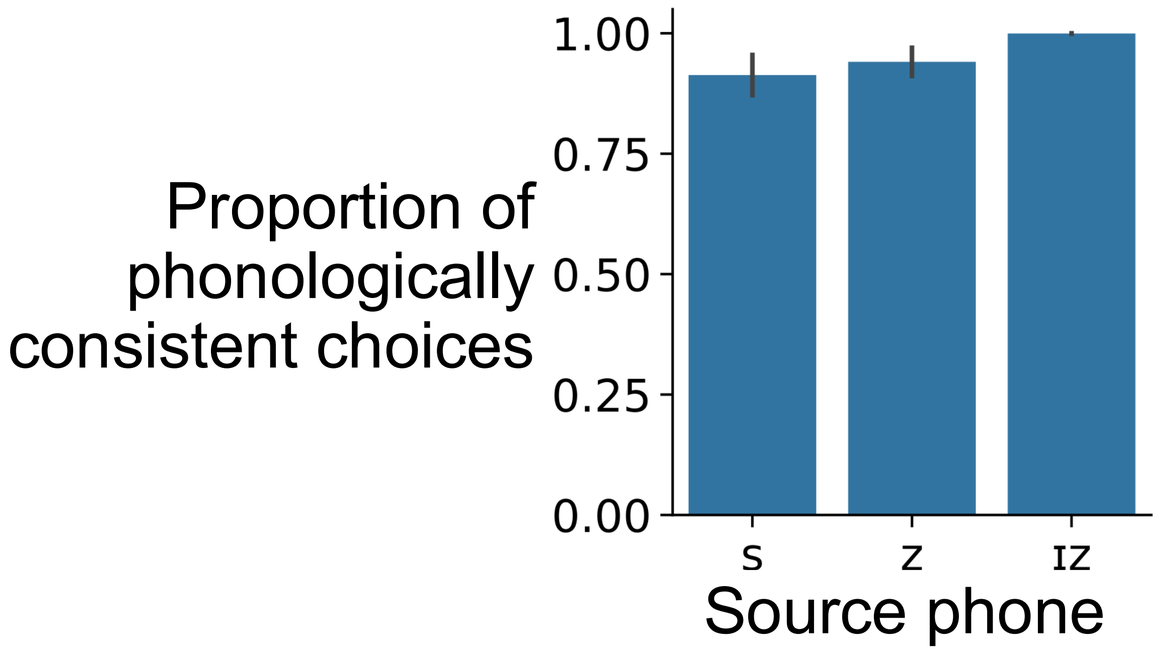
[z]
own : owns :: bay :

{ 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

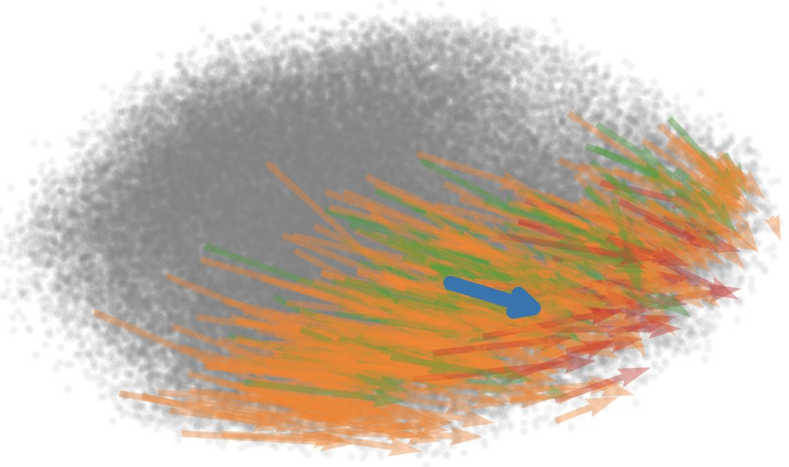
[z]
own : owns :: bay :

{ bays **(consistent)**
base **(inconsistent)** }

- 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 new models of human spoken word recognition



Canaan Breiss

Matt Leonard

Edward Chang