
# Grounded Language Learning, from Sounds & Images to Meaning (and everything in between)

Afra Alishahi

Tilburg University

Can we simulate language learning
In a naturalistic setting?

# Language Acquisition in the Wild

# Grounded Language Learning

- Acquire knowledge of language from utterances grounded in perceptual context
  - What linguistic representations are learned?
- Historically, linguistic formalisms were given, and models were adapted to these formalisms.
  - Example: how to induce a Context Free Grammar from a corpus?
- How do we know that humans use similar representations?

# An Alternative Approach

- Neural models of language allow for applying general-purpose architectures to performing different language tasks

- What type(s) of linguistic knowledge are necessary for performing a given task?

# Focus of this Talk

How to simulate
grounded language learning?

How to analyze emerging knowledge
in neural models of language?

# A Naturalistic Scenario

*Look, they are washing the elephants!*

- Where to get the input data from?

# Using Images of Natural Scenes

- a woman is playing a frisbee with a dog.
- a woman is playing frisbee with her large dog.
- a girl holding a frisbee with a dog coming at her.
- a woman kneeling down holding a frisbee in front of a white dog.
- a young lady is playing frisbee with her dog.

- Datasets such as Flickr30K and MS-Coco provide images of natural scenes and their textual description
  - Image-caption pairs as proxy for linguistic & perceptual context

# Imaginet: Learning by Processing

$$L^T(\theta) = -\frac{1}{\tau}\sum_{t=1}^{\tau}\log p(S_t|S_{1:t})$$

linguistic pathway → $w_{t+1}$

$w_1$ $w_2$ $w_3$ ....

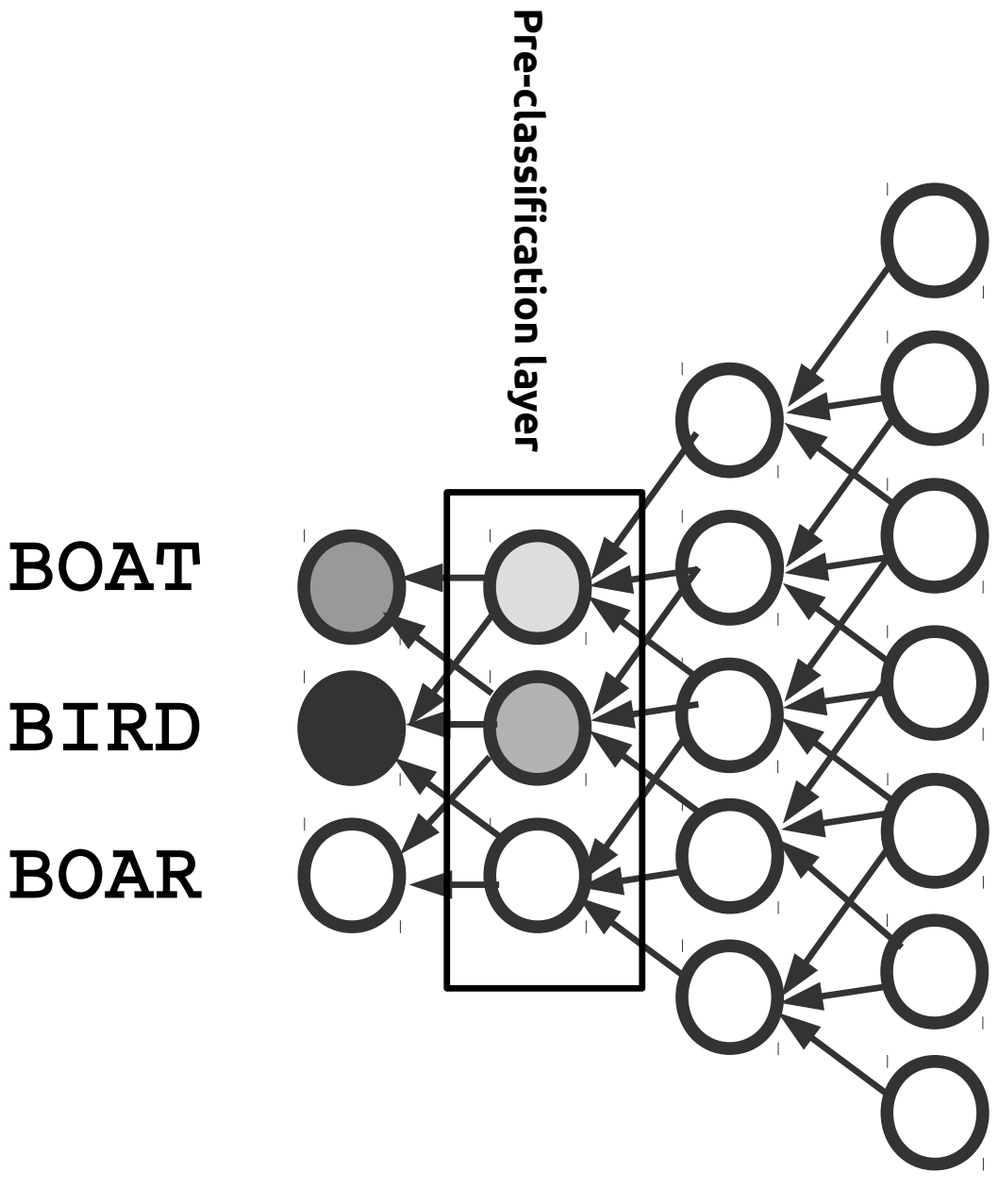
visual pathway →

$$L^V(\theta) = 1 - \frac{\hat{\mathbf{i}} \cdot \mathbf{i}}{\|\hat{\mathbf{i}}\|\|\mathbf{i}\|}$$

shared word representations

$$L = \alpha L^T + (1-\alpha)L^V$$

# Image Representations




VGG-16: Simonyan & Zisserman (2014)

# What Does the Model Learn?

- Evaluate the model through a set of comprehension and production tasks
  - Production: retrieve the closest words or phrases to an image
  - Comprehension: retrieve the closest images to a word or multi-word phrase
- Estimate word pair similarity, and correlate with human judgments

# Word Meaning

**Word:** parachute

**Original label:** parachute

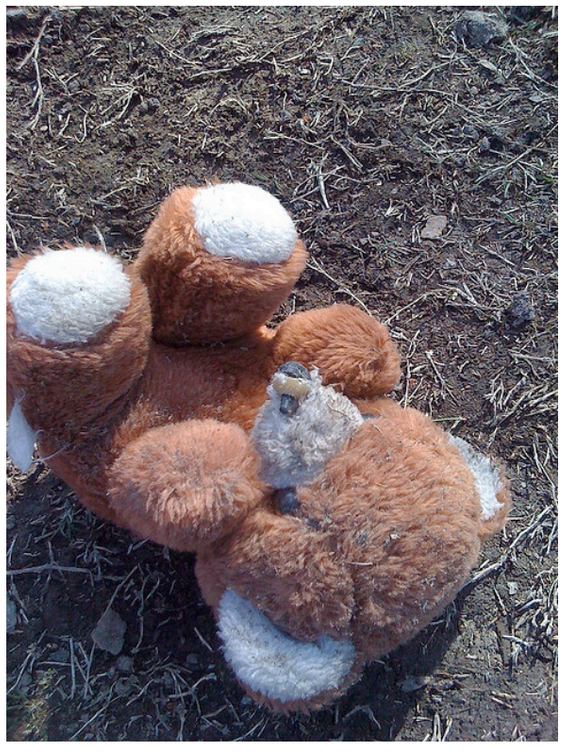
**Word:** bicycle

**Original label:** bicycle-built-for-two

# Utterance Meaning

**Input sentence:**

*"a brown teddy bear lying on top of a dry grass covered ground ."*



- Does the model learn and use any knowledge about sentence structure?

# Utterance Meaning

**Input sentence:**

*"a brown teddy bear lying on top of a dry grass covered ground ."*

**Scrambled input sentence:**

*"a a of covered laying bear on brown grass top teddy ground . dry"*

# Another Example

**Original:** "a variety of kitchen utensils hanging from a UNK board ."

**Scrambled:** "kitchen of from hanging UNK variety a board utensils ."


# What Kind of Structure is Learned?

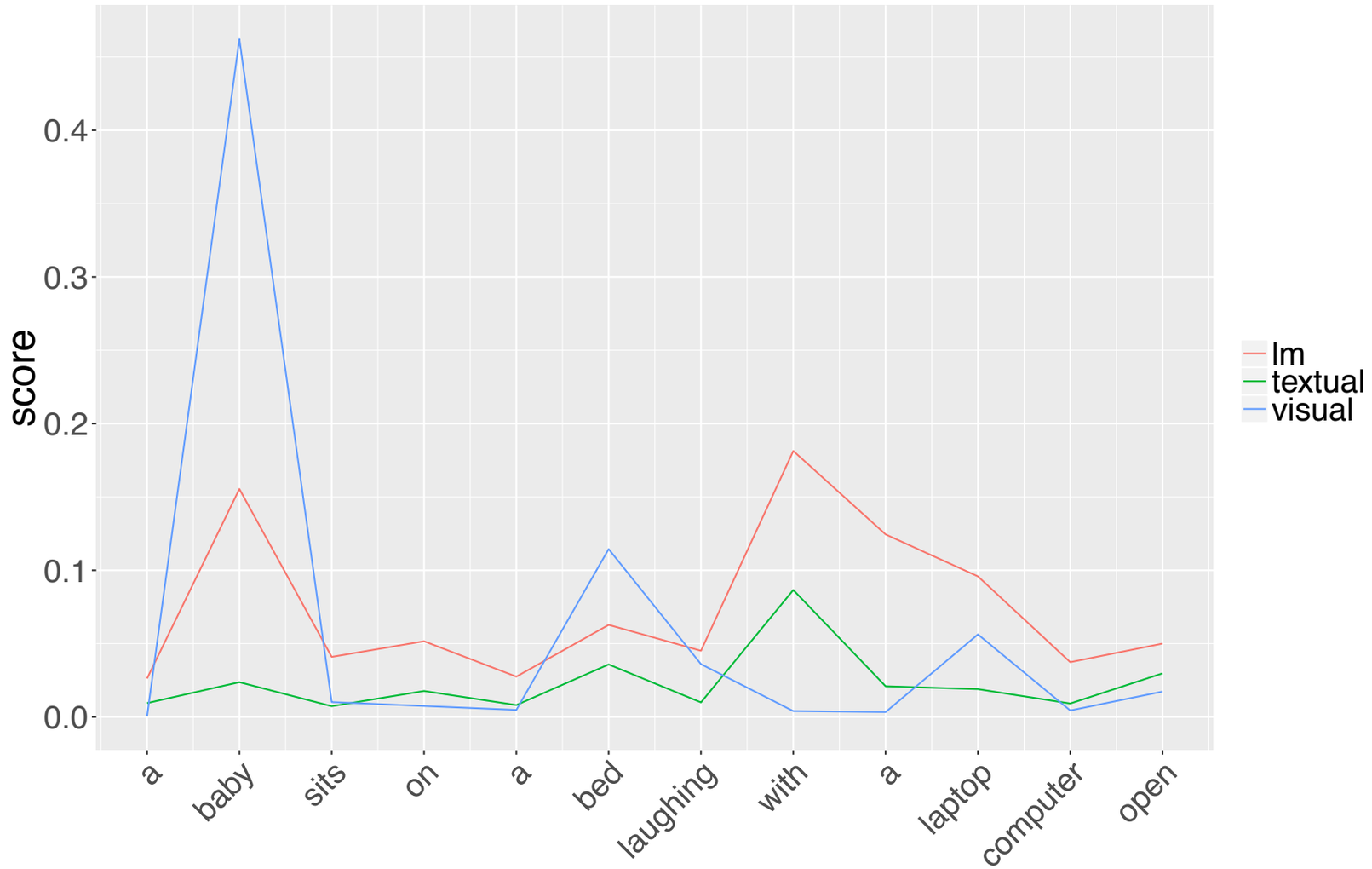
- Some structural patterns:
    - Topics appear in sentence-initial position
    - Same words are more important as heads than modifiers
    - Periods terminate a sentence
    - Sentences tend to not start with conjunctions
- Can we pin down the acquired structure more systematically?

# Analysis Technique: Omission Score

- Perturbation study: how would the omission of a word from an input sentence affect the model?

$$\text{omission}(i, S) = 1 - \text{cosine}(\mathbf{h}_{\text{end}}(S), \mathbf{h}_{\text{end}}(S_{\setminus i}))$$

# Contribution of each Word




*a baby sits on a bed laughing with a laptop computer open*

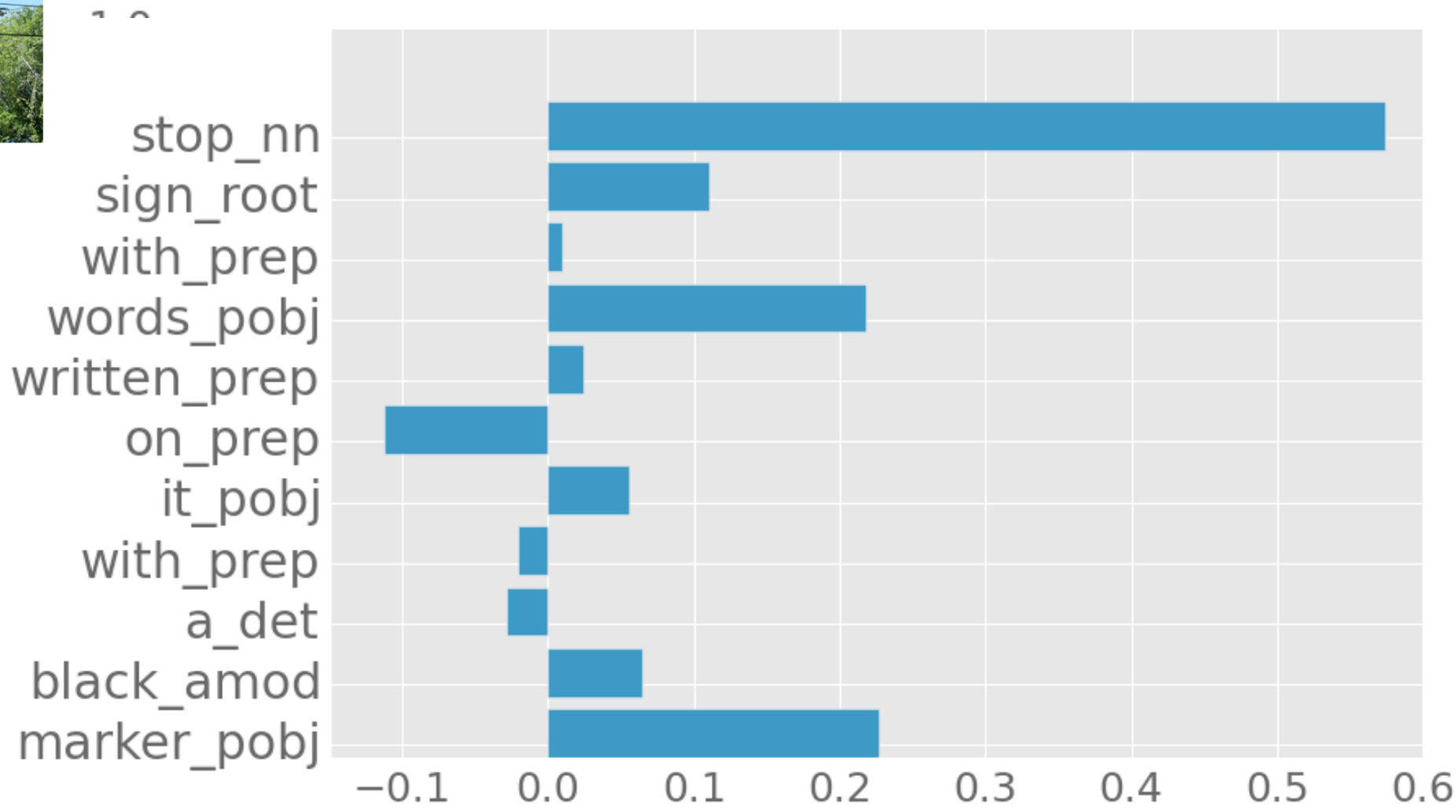
# Contribution of each Word

*a baby sits on a bed laughing with a laptop computer open*

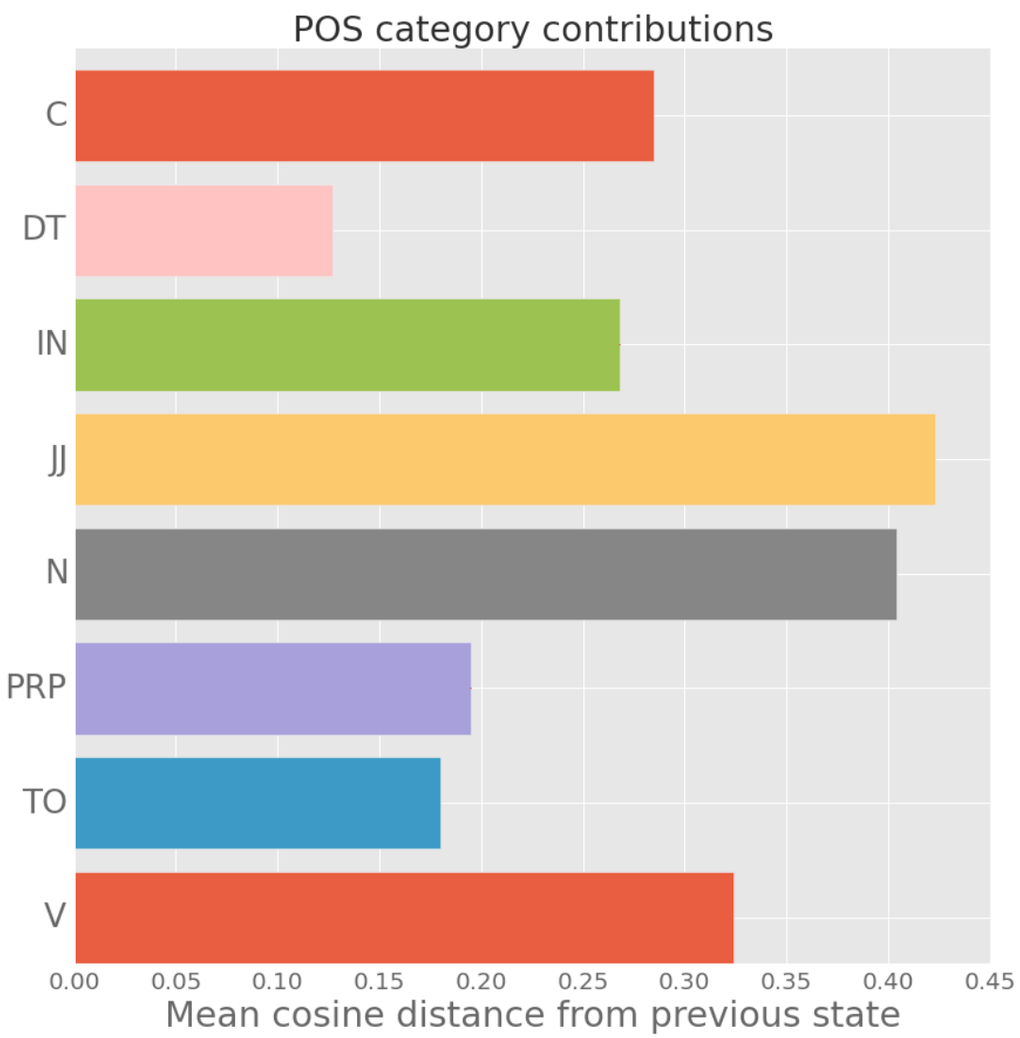
# An Example

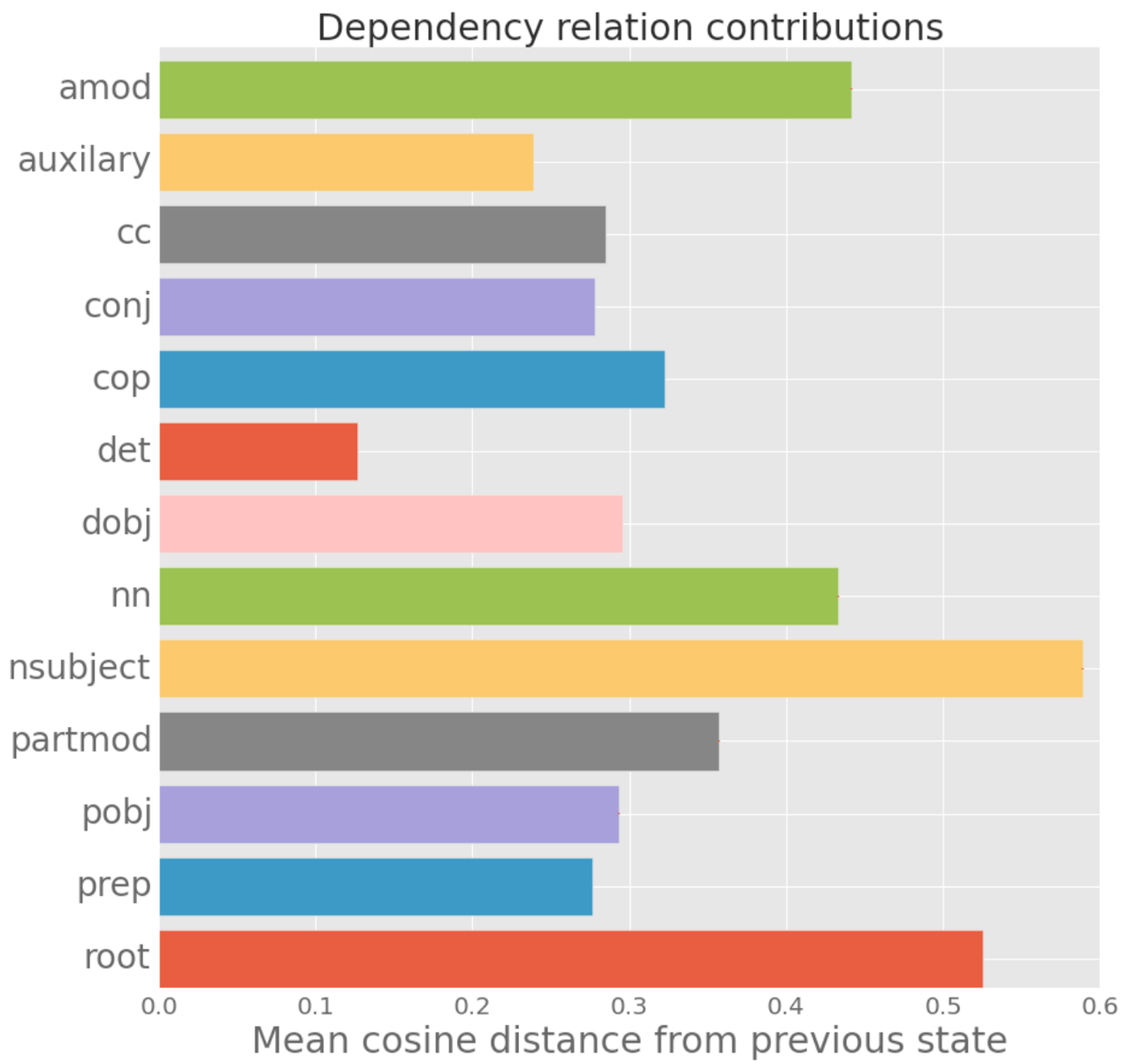

**Input sentence:**

*"Stop sign with words written on it with a black marker ."*



# Impact of POS & Dependency Categories



# What We Have Learned

- Perceptual and linguistic context provide complementary sources of information for learning form-meaning associations.
- Analysis of the model's hidden states gives us insight into useful structural knowledge for language comprehension and production, but more sophisticated techniques are needed.

# A Typical Language Learning Scenario

# A Typical Language Learning Scenario

# Modeling Grounded Speech

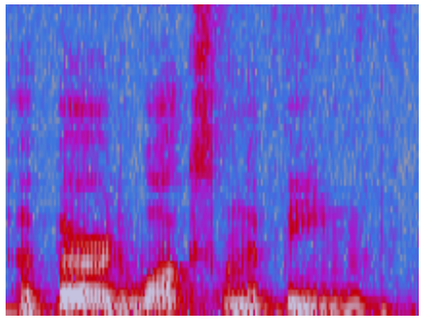


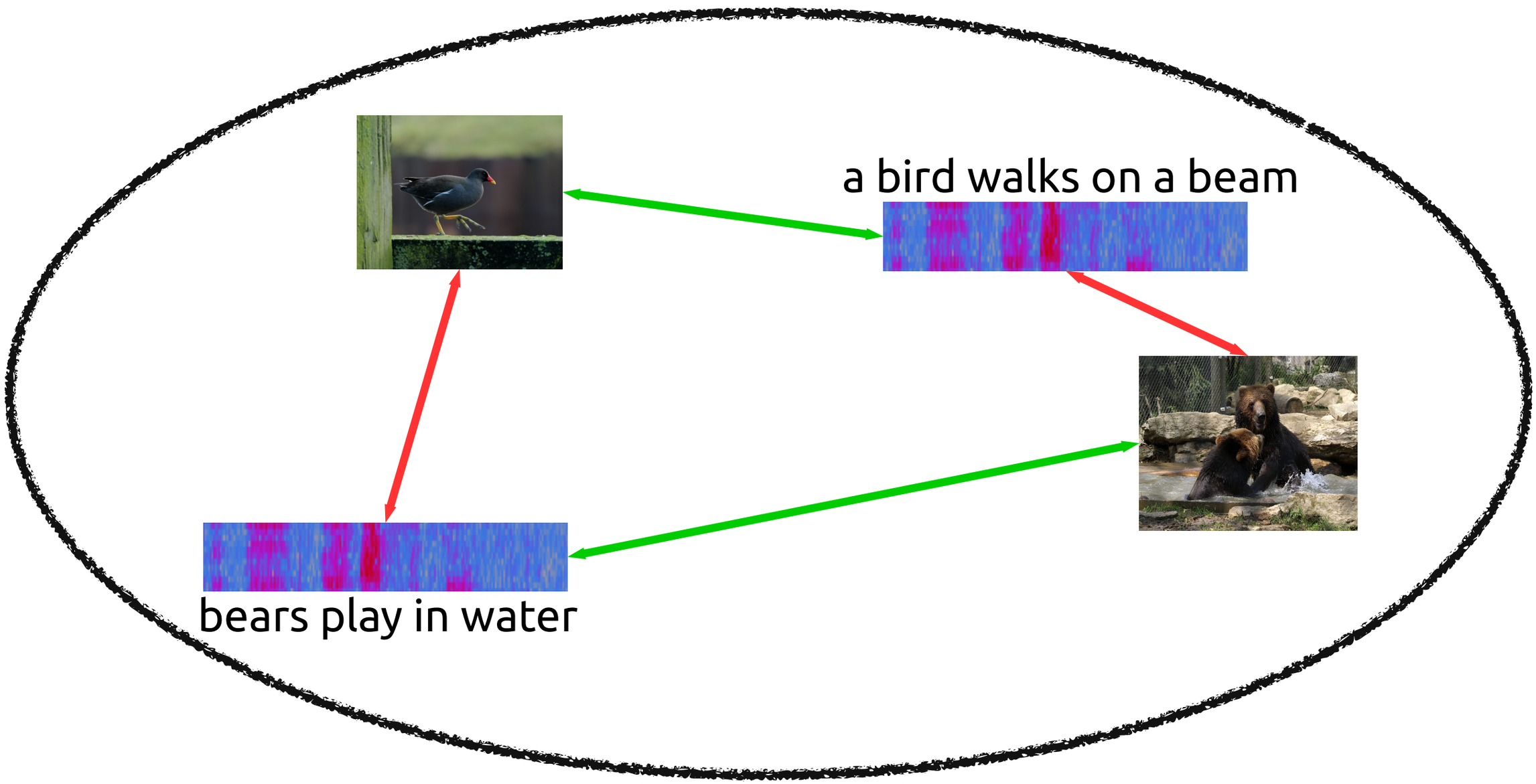
Image Model

Speech Model

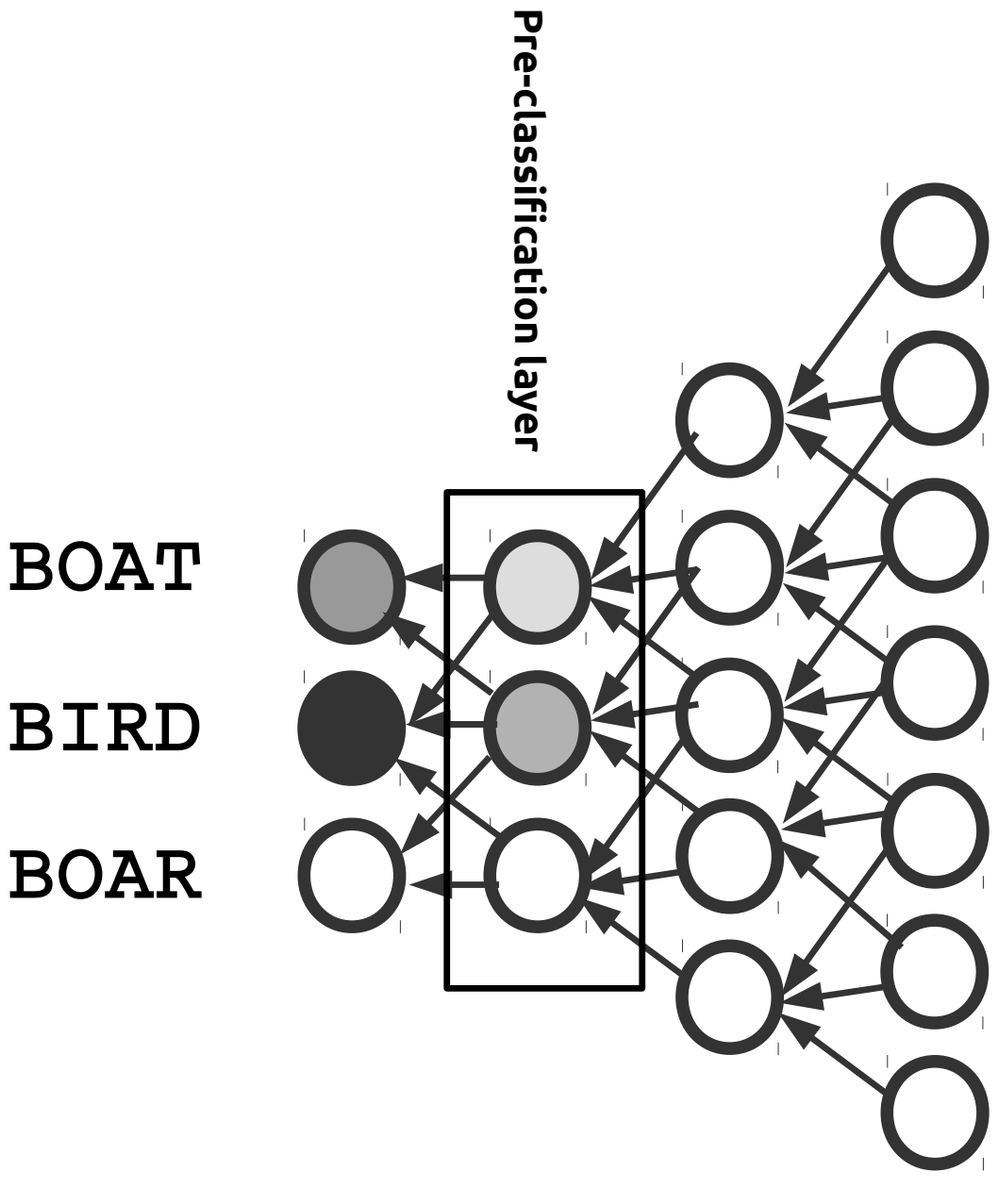
Joint Semantic Space

# Joint Semantic Space

$$\sum_{u,i}\left(\sum_{u'}\max[0,\alpha+d(u,i)-d(u',i)]+\sum_{i'}\max[0,\alpha+d(u,i)-d(u,i')]\right)$$

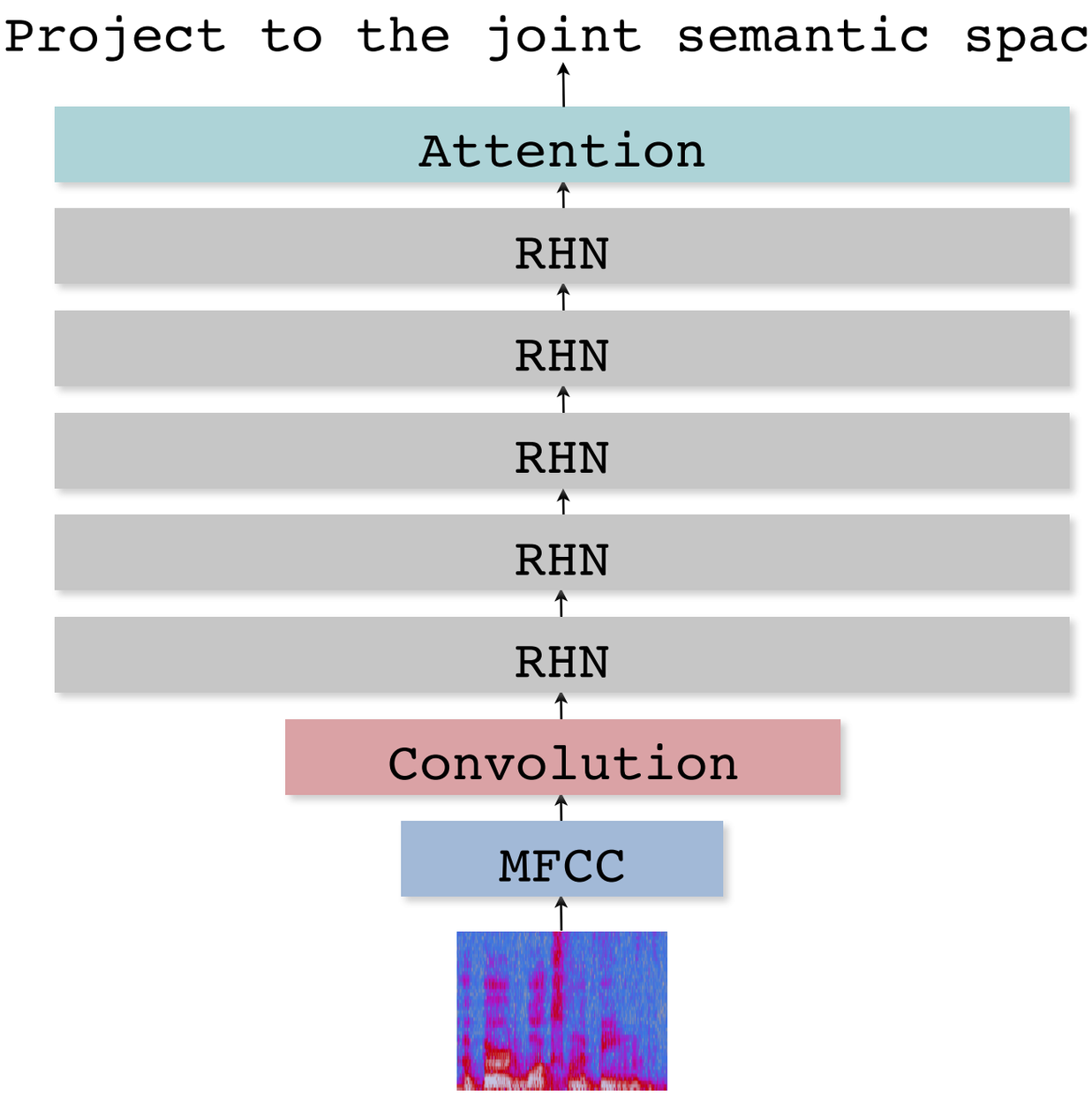


# Image Model




VGG-16: Simonyan & Zisserman (2014)

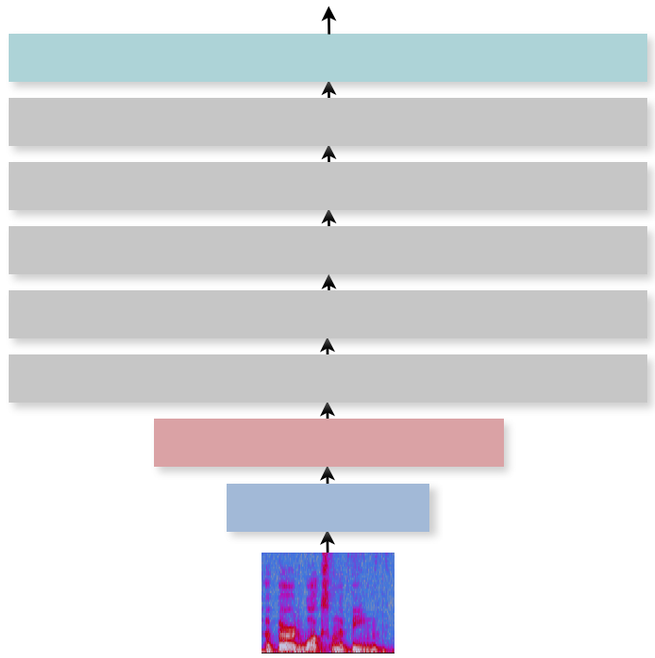
# Speech Model




- Attention: weighted sum of last RHN layer units
- RHN: Recurrent Highway Networks (Zilly et al., 2016)
- Convolution: subsampling MFCC vector

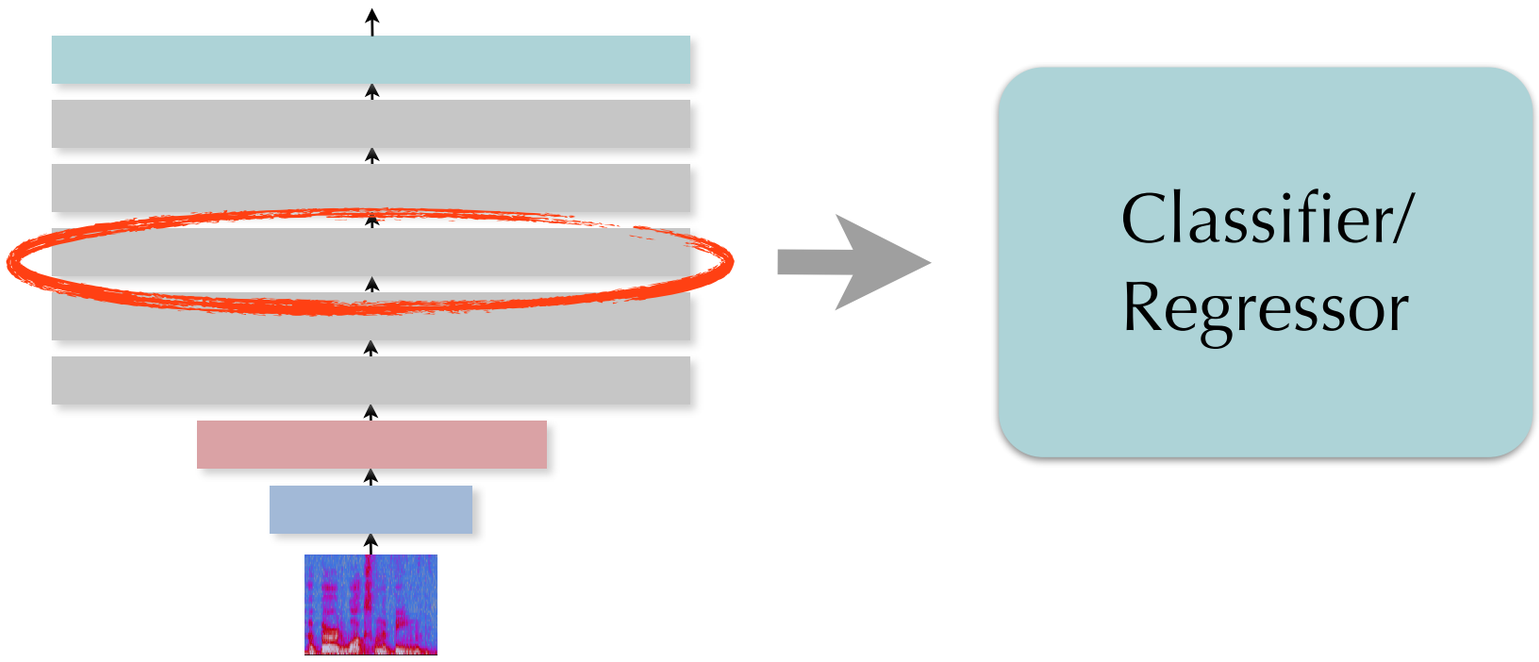
# What does the Model Learn?

- What aspects of language does the model encode?
  - Does the model encode linguistic form, and on which layers?
  - Does it encode meaning, and on which layers?



- Auxiliary tasks using layer activations as feature

# “Auxiliary Tasks” or “Probing Classifiers”




Trained to perform Task X

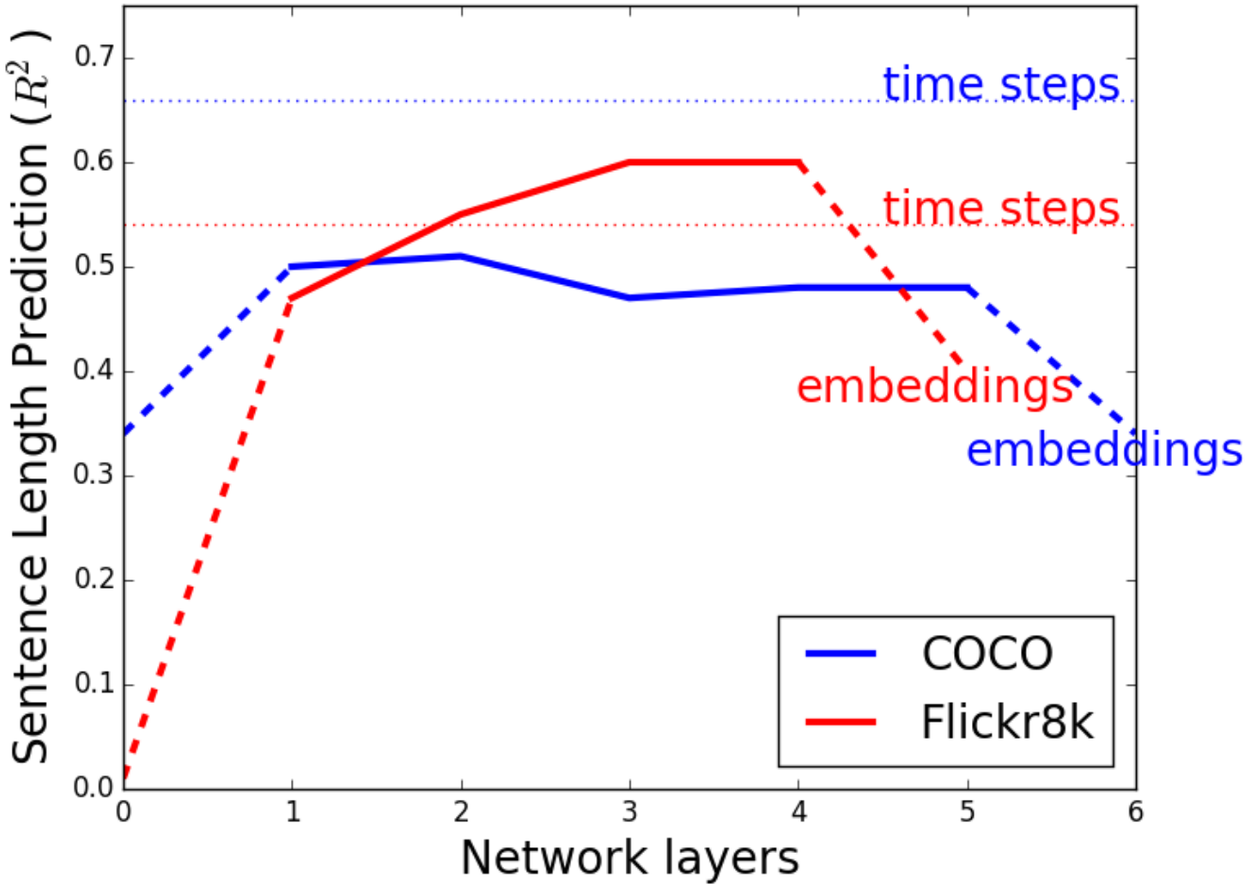
Trained to predict linguistic knowledge Y

# Analysis via Auxiliary Tasks

- Predicting utterance length
- Predicting the presence of specific words
- Representation of similarity
- Homonym detection
- Synonym detection

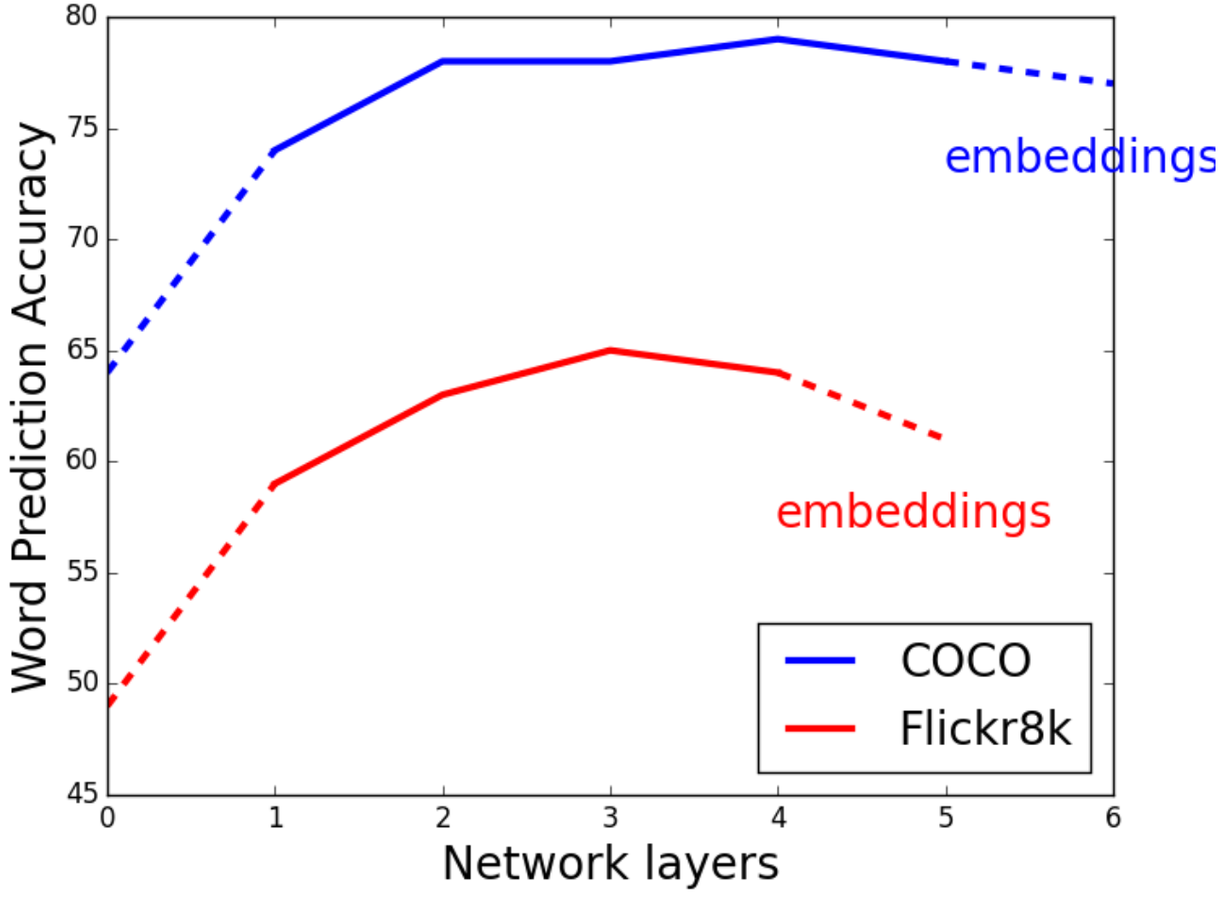
# Predicting Utterance Length

- Input: MFCC features from speech signal (layer 0), layer activations (1--4/5), or utterance embeddings (5--6)
- Prediction model: Linear Regression



# Presence of Individual Words

- Input: MFCC features for words + layer activations
- Prediction model: Multilayer Perceptron



# Synonym Discrimination

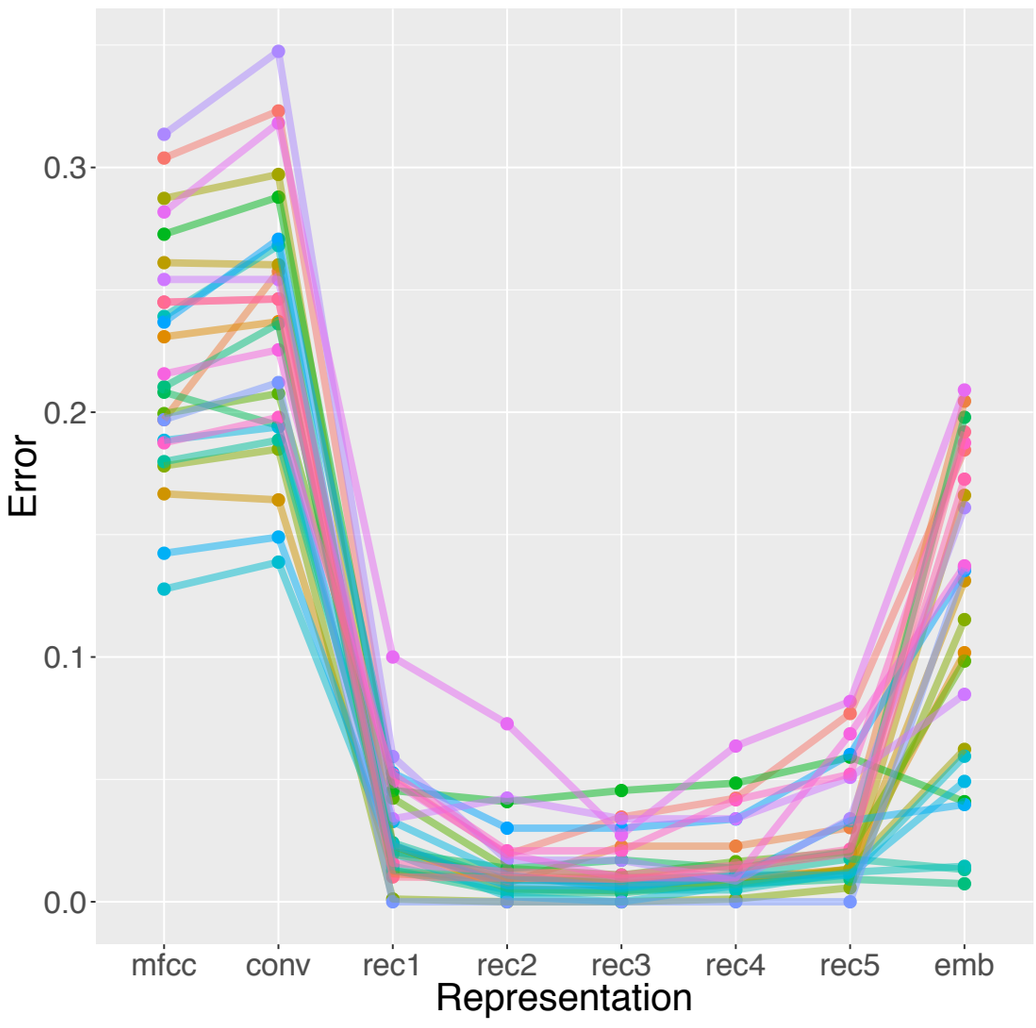
- Distinguishing between synonym pairs in the same context:

  *A girl looking at a **photo***

  *A girl looking at a **picture***

- Synonyms were selected using WordNet synsets:
  - The pair have the same POS tag and are interchangeable
  - The pair clearly differ in form (not *donut/doughnut*)
  - The more frequent token in a pair constitutes less than 95% of the occurrences.

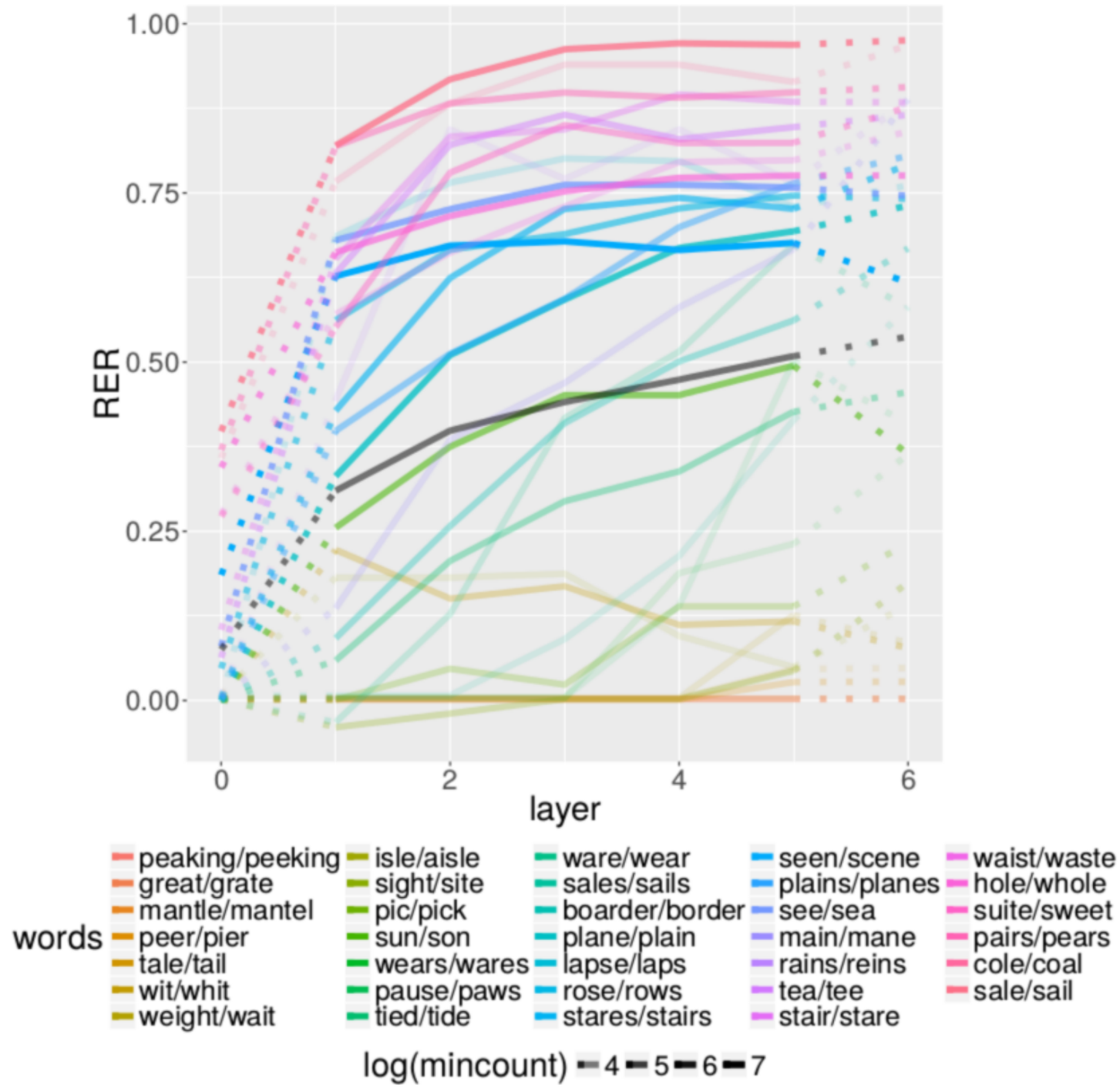
# Synonym Discrimination

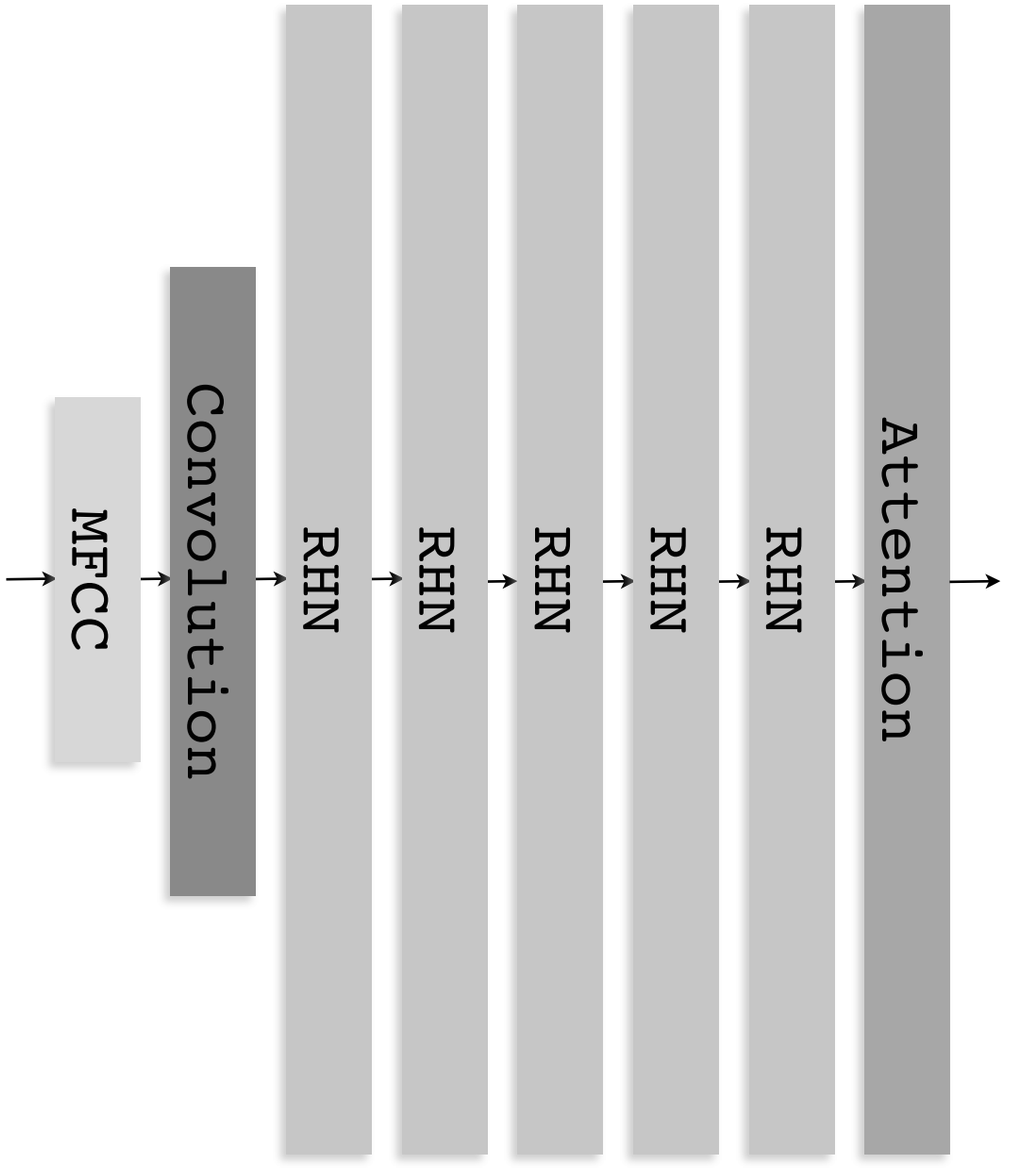



# Homonym Disambiguation

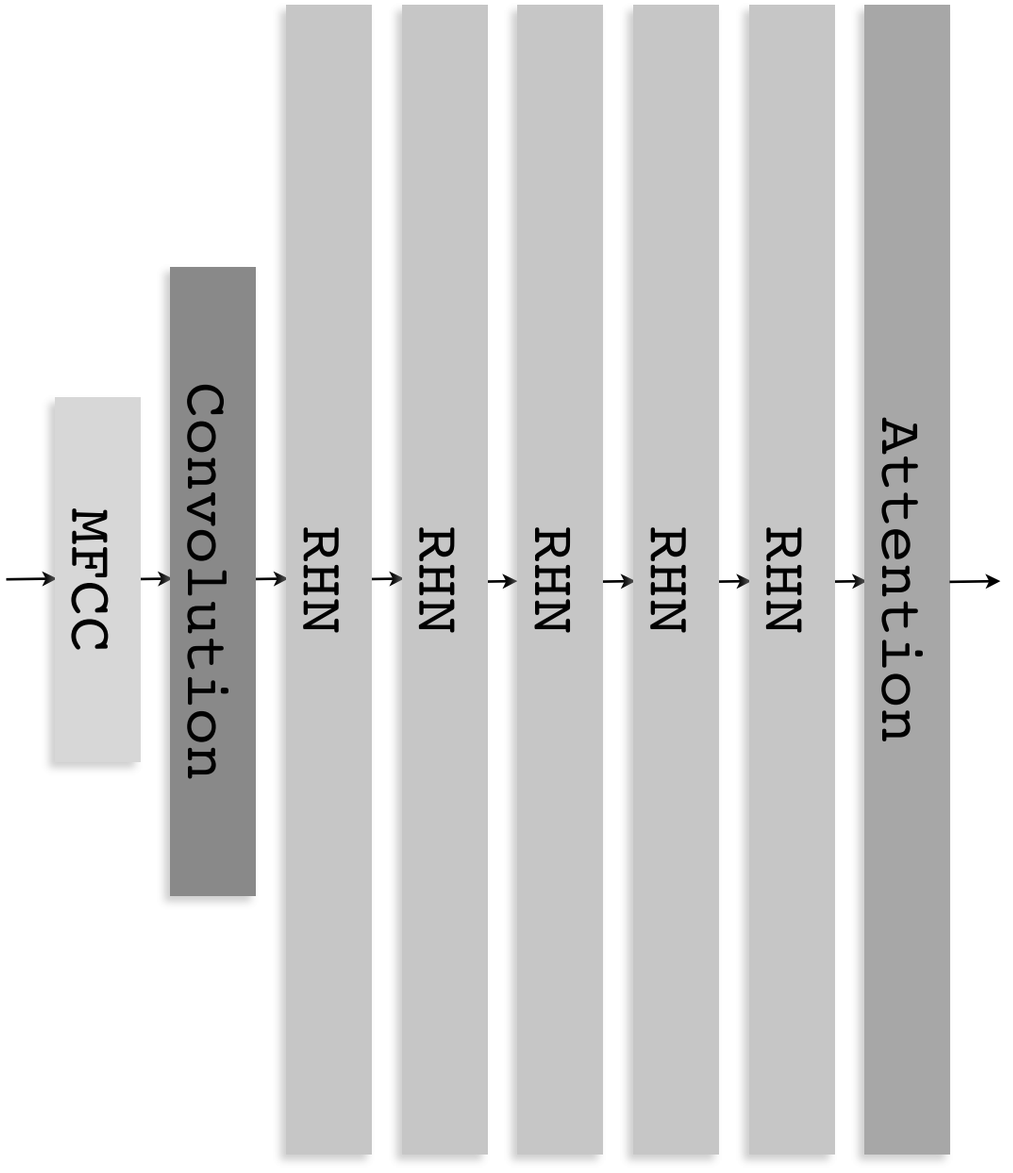
- Distinguishing between homonym pairs in different contexts, e.g. ***suite*** versus ***sweet*** or ***sale*** versus ***sail***

- We selected homonym pairs in our dataset where
  - both forms appear more than 20 times
  - the two forms have different meanings (not *theatre/theater*)
  - neither form is a function word
  - the more frequent form constitutes less than 95% of the occurrences.


1.00
0.75
0.50
0.25
0.00
RER
0
2
4
6
layer
words
peaking/peeking
great/grate
mantle/mantel
peer/pier
tale/tail
wit/whit
weight/wait
isle/aisle
sight/site
pic/pick
sun/son
wears/wares
pause/paws
tied/tide
ware/wear
sales/sails
boarder/border
plane/plain
lapse/laps
rose/rows
stares/stairs
seen/scene
plains/planes
see/sea
main/mane
rains/reins
tea/tee
stair/stare
waist/waste
hole/whole
suite/sweet
pairs/pears
cole/coal
sale/sail
log(mincount)
4
5
6
7




| | | | |
|---|---|---|---|
| * | * | | Utterance Length |
| | | * | Word Presence |
| * | | | Edit Similarity |
| | | * | Meaning Similarity |
| | | * | * Homonym Detection |
| * | | | Synonym Detection |




| | | | |
|---|---|---|---|
| * | * | | Utterance Length |
| | | * | Word Presence |
| | * | | Edit Similarity |
| | | * | Meaning Similarity |
| | | * * | Homonym Detection |
| | * | | Synonym Detection |

# Encoding of Phonology

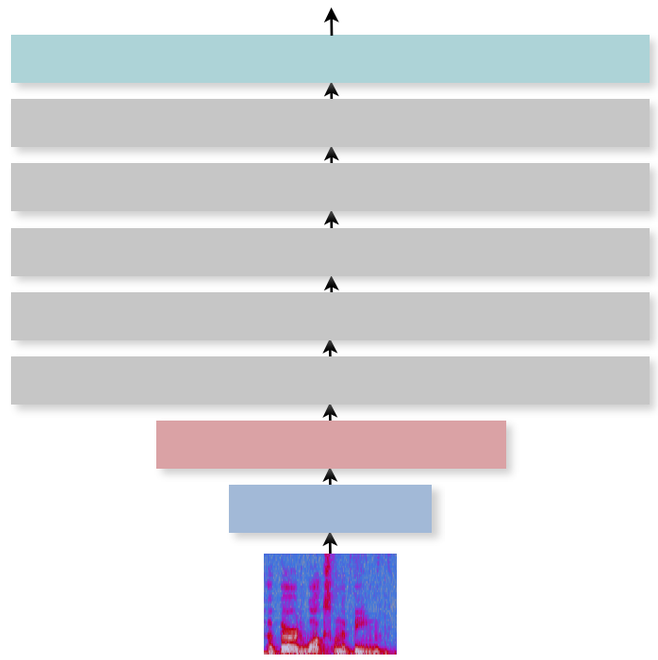
- Questions: how is phonology encoded in
  - MFCC features extracted from speech signal?
  - activations of the layers of the model?

# Encoding of Phonology

- Questions: how is phonology encoded in
  - MFCC features extracted from speech signal?
  - activations of the layers of the model?

- Data: Synthetically Spoken COCO dataset

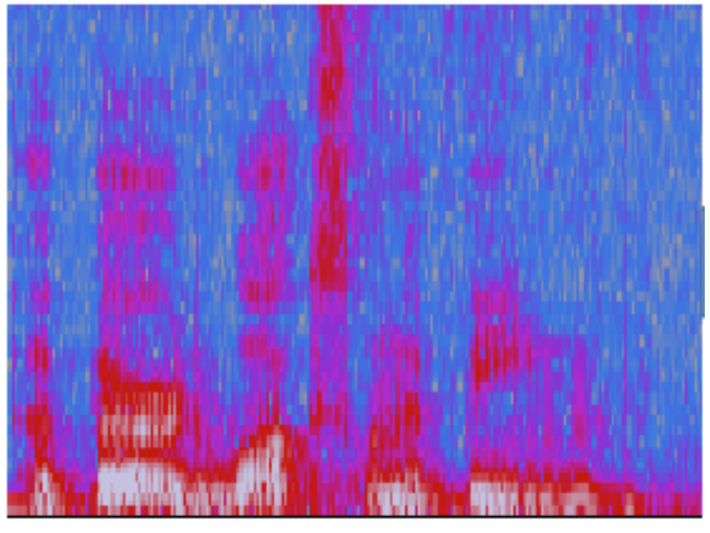
- Experiments:
  - Phoneme decoding and clustering
  - Phoneme discrimination



# Phoneme Decoding

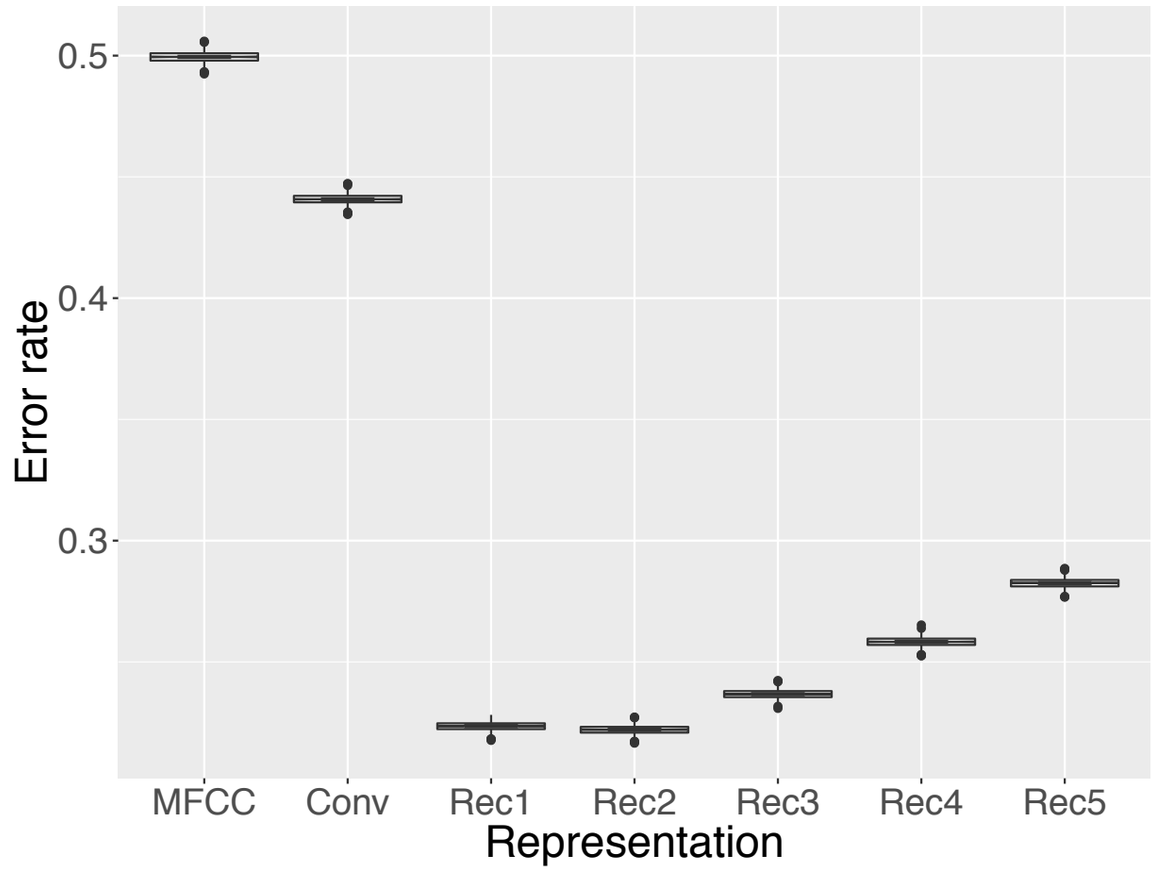
- Identifying phonemes from speech signal/activation patterns: supervised classification of aligned phonemes

- Speech signal was aligned with phonemic transcription using Gentle toolkit (based on Kaldi, Povey et al., 2011)



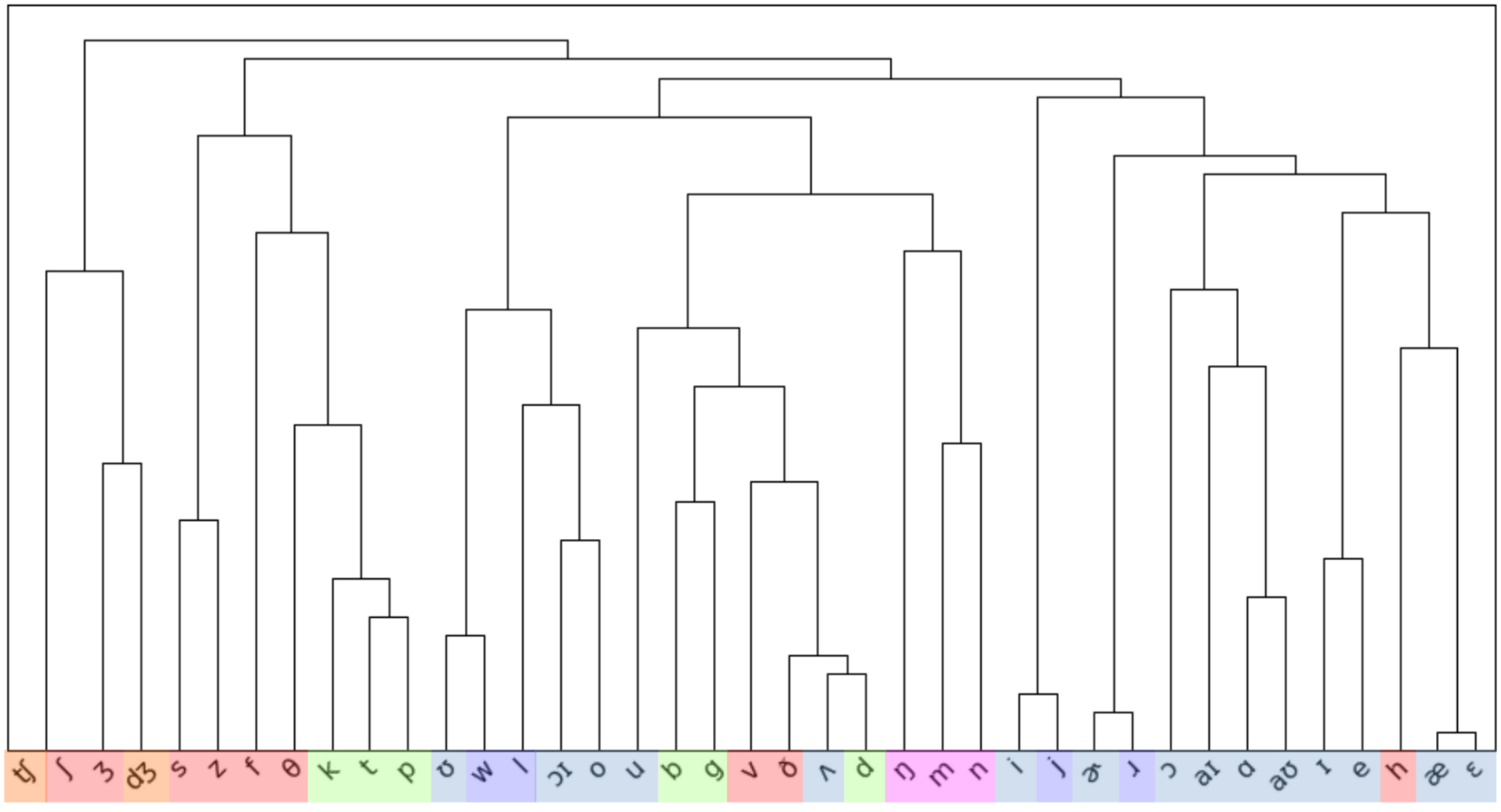
# Phoneme Decoding

- Identifying phonemes from speech signal/activation patterns: supervised classification of aligned phonemes



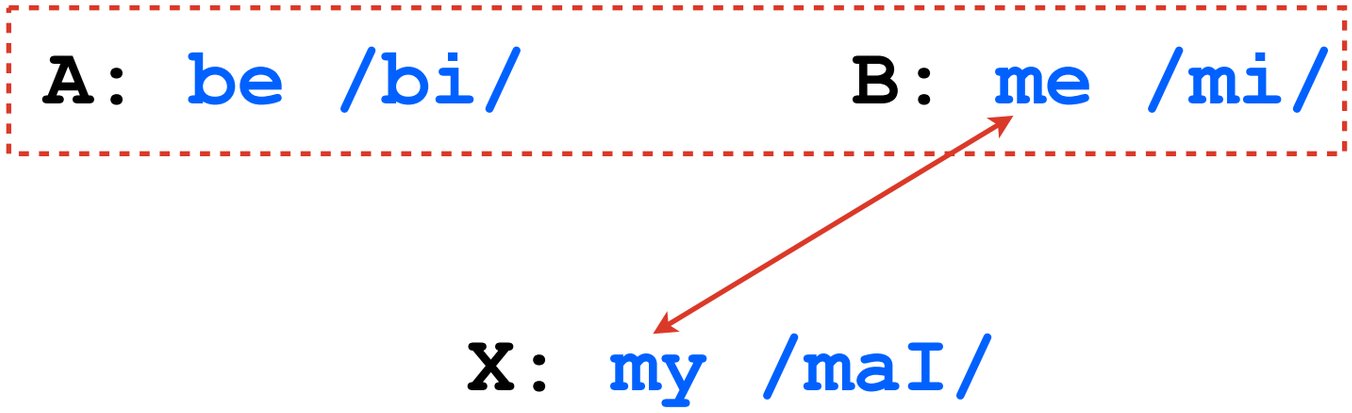
# Organization of Phonemes

- Agglomerative hierarchical clustering of phoneme activation vectors from the first hidden layer:



# Phoneme Discrimination

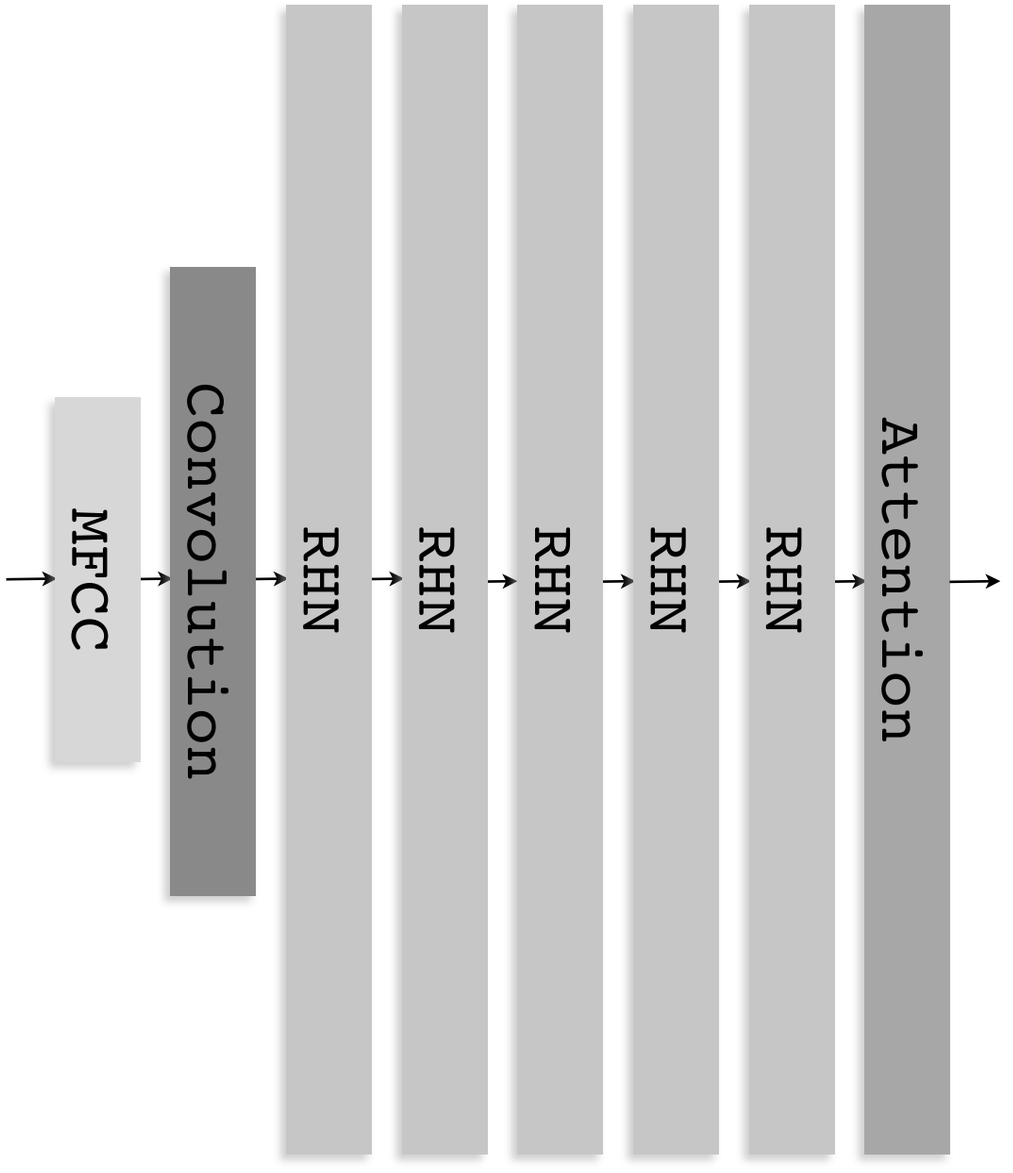
- ABX task (Schatz et al., 2013): discriminate minimal pairs; is X closer to A or to B?




- A, B and X are CV syllables
- (A,B) and (B,X) are minimum pairs, but (A,X) are not (34,288 tuples in total)

# Phoneme Discrimination

| | |
|---|---|
| MFCC | 0.71 |
| Convolutional | 0.73 |
| Recurrent 1 | 0.82 |
| Recurrent 2 | 0.82 |
| Recurrent 3 | 0.80 |
| Recurrent 4 | 0.76 |
| Recurrent 5 | 0.74 |



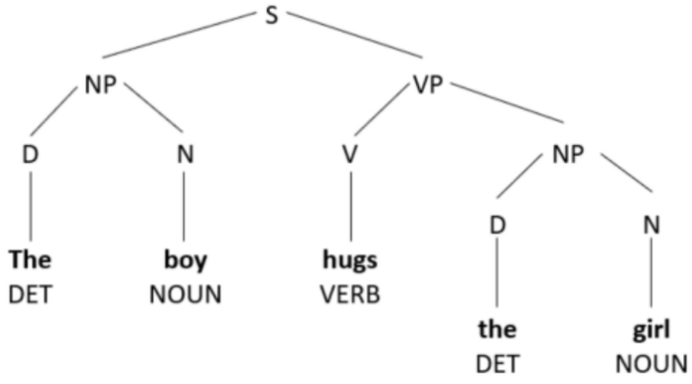

| | | | |
|---|---|---|---|
| * | * | | Utterance Length |
| | | * | Word Presence |
| | * | | Edit Similarity |
| | | * | Meaning Similarity |
| | | * * | Homonym Detection |
| | * | | Synonym Detection |
| * | * | | Phoneme Decoding |
| * | * | | Phoneme Discrimination |

# What We Have Learned

- Encodings of form and meaning emerge and evolve in hidden layers of stacked RNNs processing grounded speech

- Phoneme representations are most salient in lower layers, although large amount of phonological information persists up to the top recurrent layer

# Back to Structure



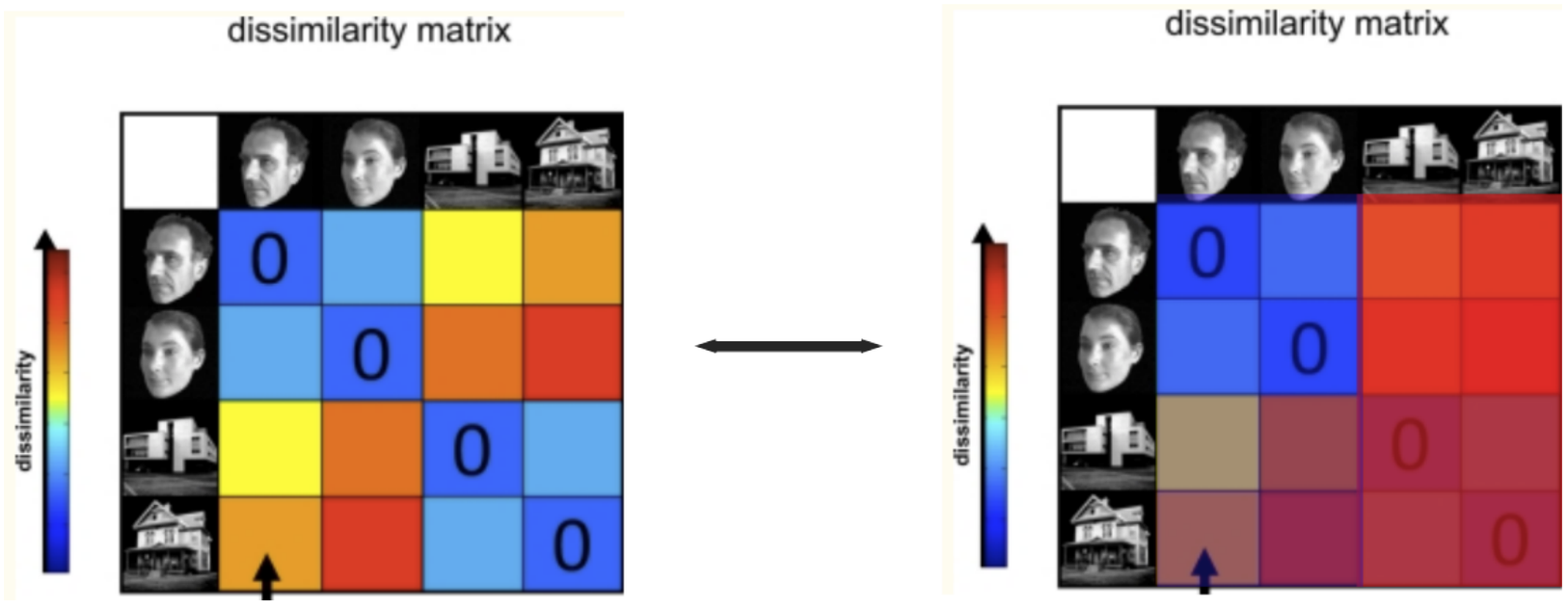


∃ h, b, g:
instance(a, hug-01) ^
instance(b, boy) ^
instance(c, girl) ^
ARG0(a, b) ^
ARG1(a, c)

- Diagnostic classifiers are not useful in probing structured representations

# Representational Similarity Analysis (RSA)

- RSA (Kriegeskorte et al., 2008): Measuring correlations between representations A and B in a similarity space
  - Compute representation (dis)similarity matrices in two spaces
  - Measure correlations between upper triangles



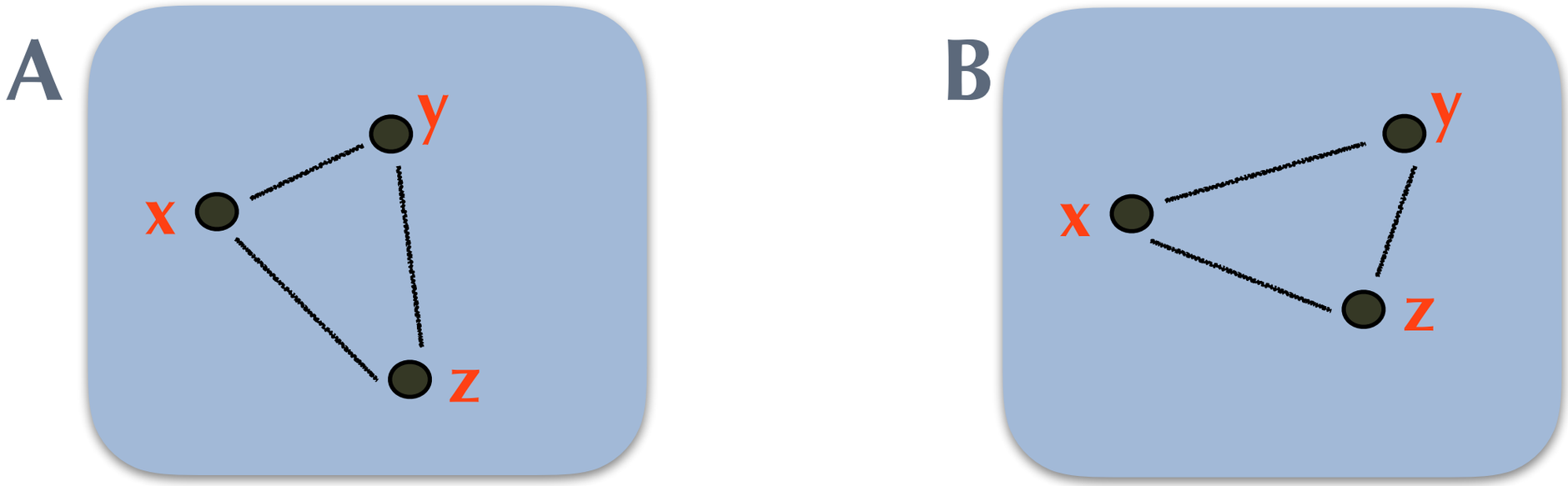
# RSA: An Example

- Sentence similarity according to
  - Sim A: human judgment
  - Sim B: estimated by a model

| Stimulus 1 | Stimulus 2 | Sim A | Sim B |
|---|---|---|---|
| A slice of pizza | A bowl of salad | 7.0 | 6.2 |
| Two dogs run | A kitty running | 8.0 | 9.0 |
| A yellow and white bird | A kitty running | 1.0 | 4.5 |

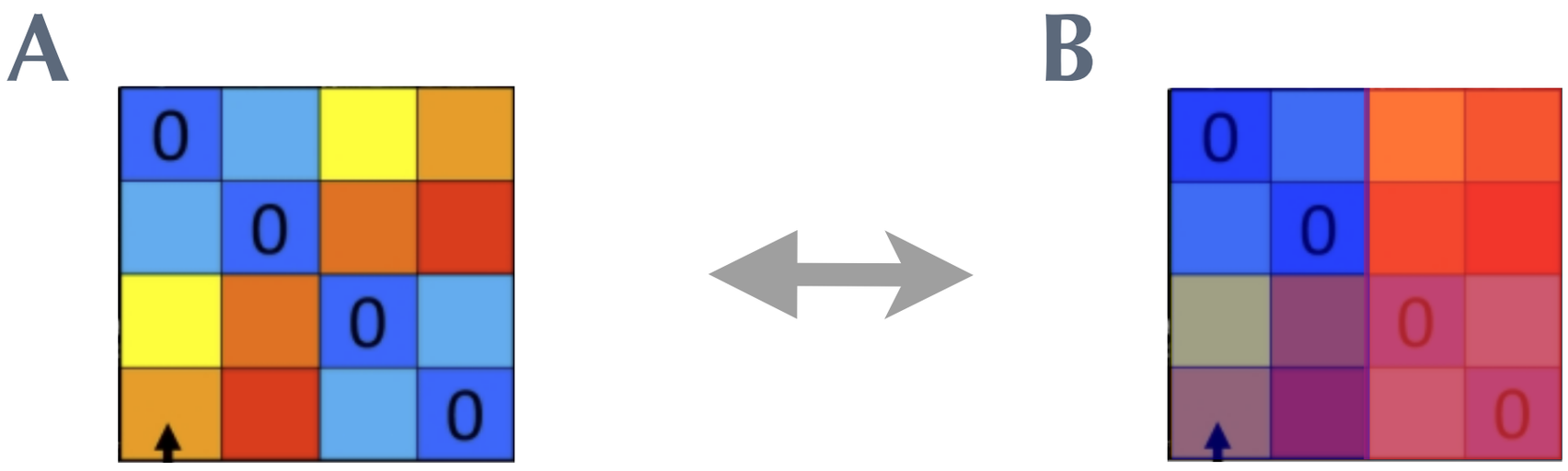
- RSA score: correlation between Sim A and Sim B

# Applying RSA to Language




- What we need: a similarity metric within two spaces A & B
  - Eg., A is a vector space, B is a space of trees or graphs
- What we do not need: a mapping between space A & B

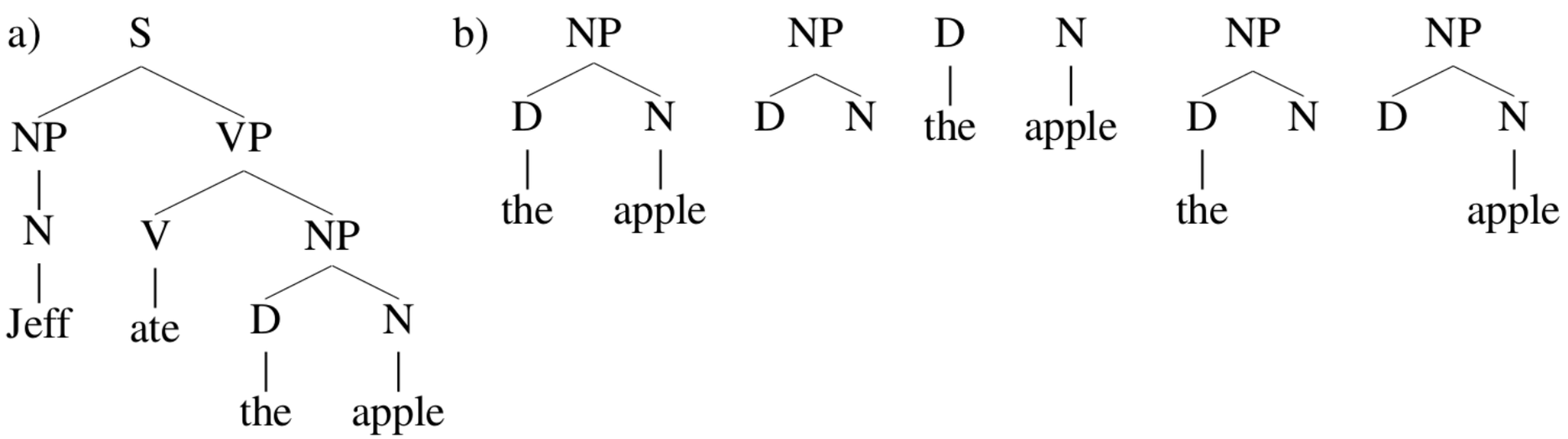
# Applying RSA to Language




- What we need: a similarity metric within two spaces A & B
  - Eg., A is a vector space, B is a space of trees or graphs
- What we do not need: a mapping between space A & B

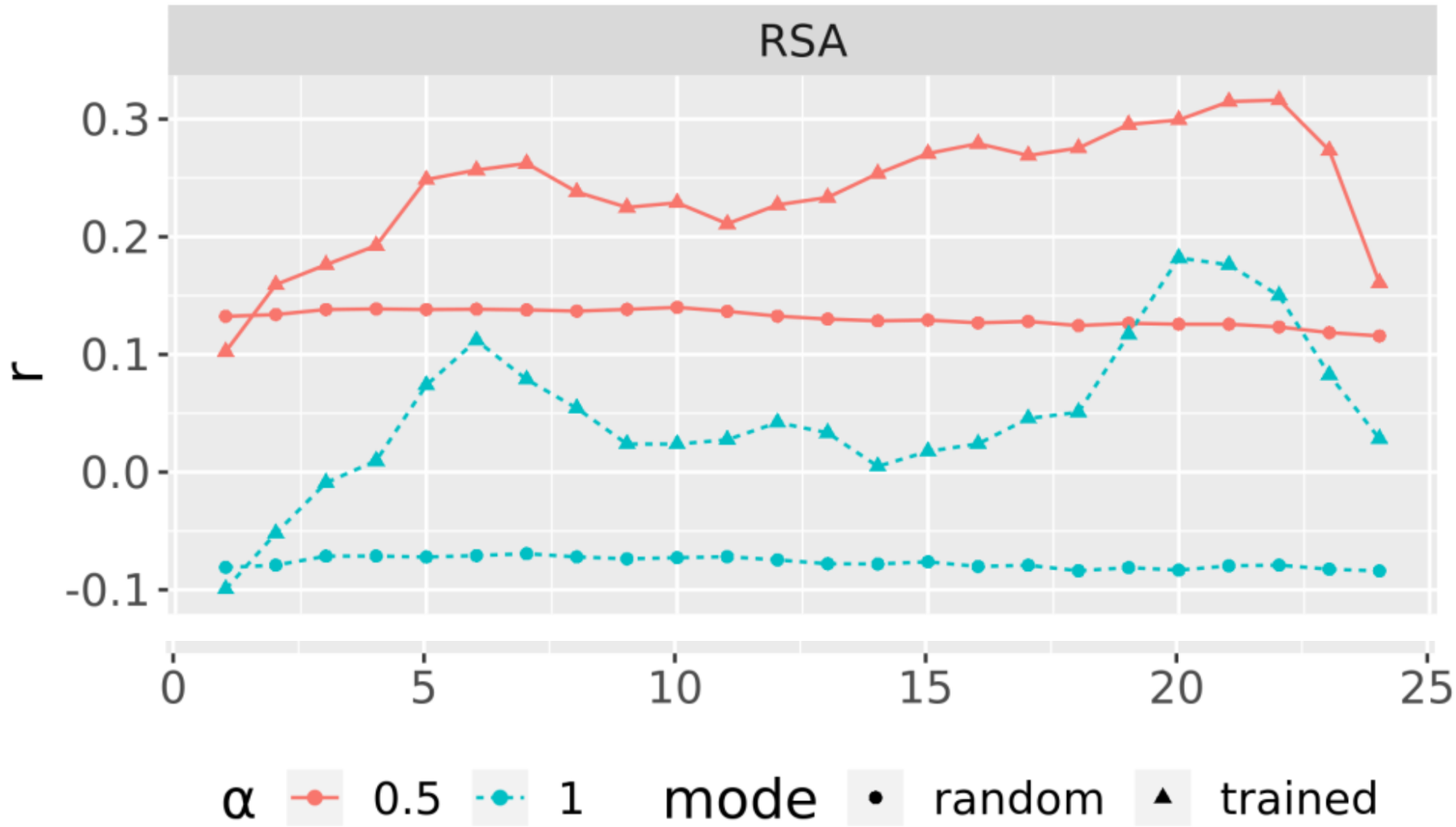
# Tree Kernels

- Measuring the similarity between two syntactic trees: count their overlapping subtrees



# Applying RSA to NN Models of Language

- BERT: a state-of-the-art language model



# Analysis Techniques: Summary

- Interpreting hidden activations in NN models of language
  - Perturbation-based techniques
  - Probing classifiers and auxiliary tasks
  - Representation similarity analysis

BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP

1st edition: EMNLP'2018, Brussels
2nd edition: ACL 2019, Florence

# Modeling Language Learning: What Next?

- Direct comparison with infant behaviour
  - Learning phases, behavioural patterns

- Videos rather than static images
  - Exploiting movement, change and non-verbal cues

- Interaction through dialogue
  - Relying on feedback from communication success

Grzegorz Chrupała

Ákos Kádár

Lieke Gelderloos

Marie Barking


- Chrupała, Kádár & Alishahi (2015). Learning language through pictures. *ACL'2015*.
- Kádár, Chrupała & Alishahi (2017). Representation of linguistic form and function in recurrent neural networks. *Computational Linguistics*.
- Chrupała, Gelderloos & Alishahi (2017). Representation of language in a model of visually grounded speech signal. *ACL'2017*.
- Alishahi, Barking and Chrupała (2017). Encoding of phonology in a recurrent neural model of grounded speech. *CoNLL'2017*.
- Kádár, Elliott, Côté, Chrupała & Alishahi (2018). Lessons learned in multilingual grounded language learning. CoNLL.
- Chrupała and Alishahi (2019). Correlating neural and symbolic representations of language. *ACL'2019*.
- Alishahi, Chrupała and Linzen (2019). Analyzing and Interpreting Neural Networks for NLP. *Natural Language Engineering*.