C.Psyd

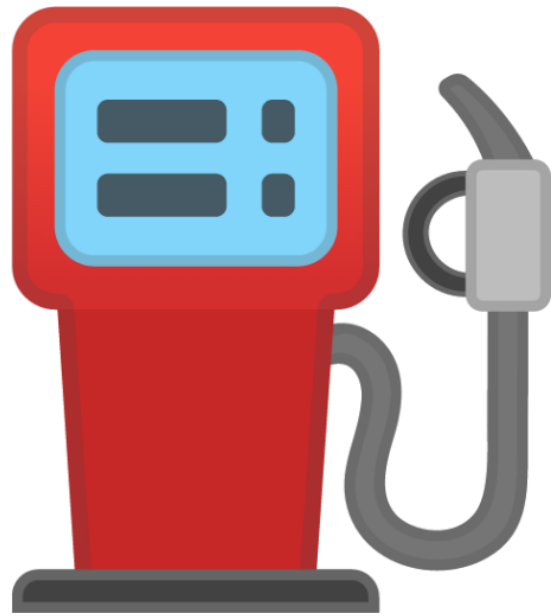
Bust or Robust Representations

By John R. Starr

Columbia NLP Seminar
November 14th, 2024

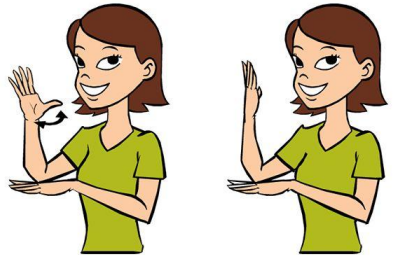
Roadmap

1. What do we want from our representations?
2. How do we interpret representations?
3. How do we test for “robust” representations?
4. How do humans determine representational similarity?

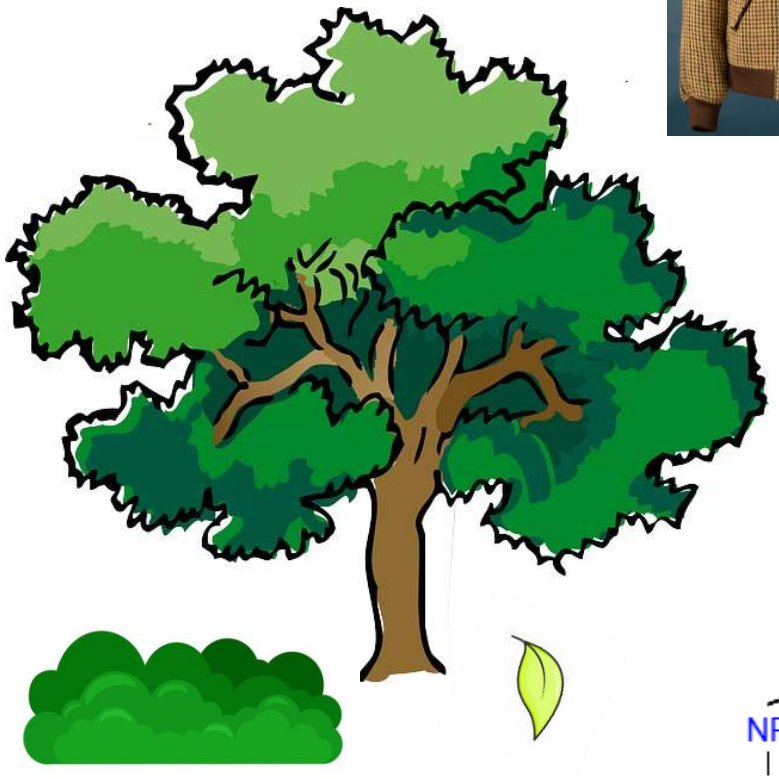


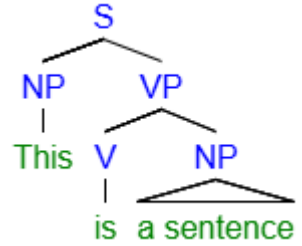
rent	cream	terrier	red	blue
green	hair	landlord	asparagus	marathon

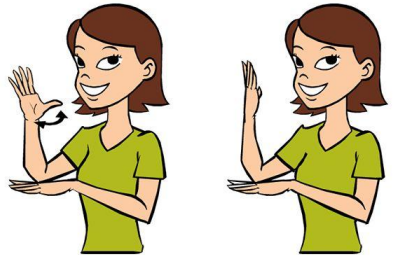
1. What do we want from our representations?



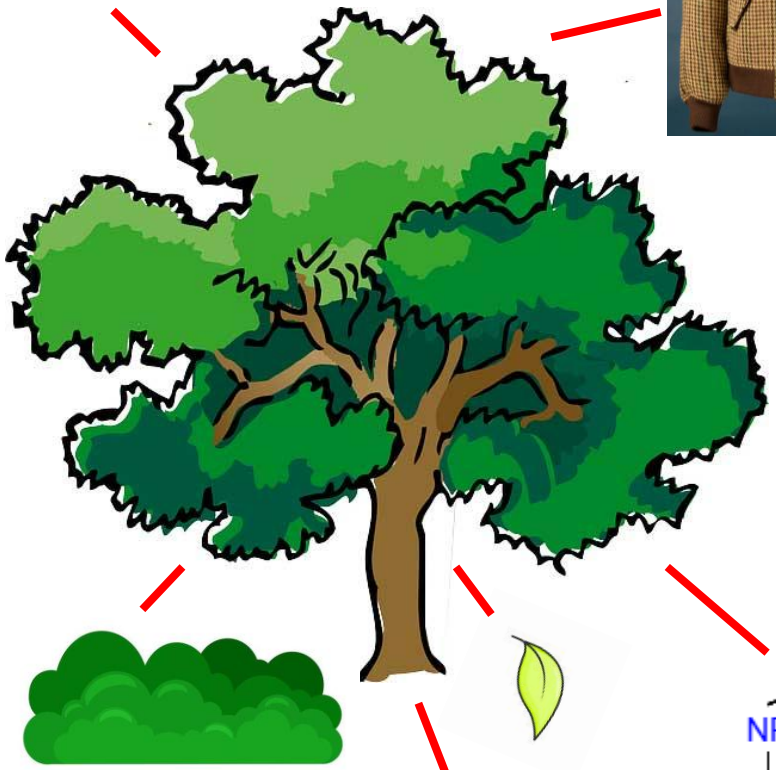
"tree"

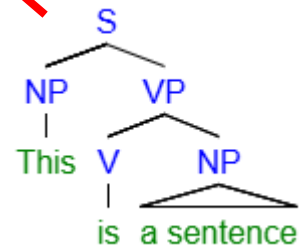




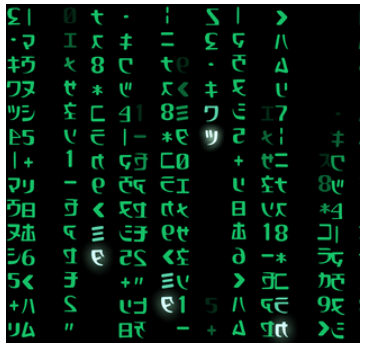


"tree"

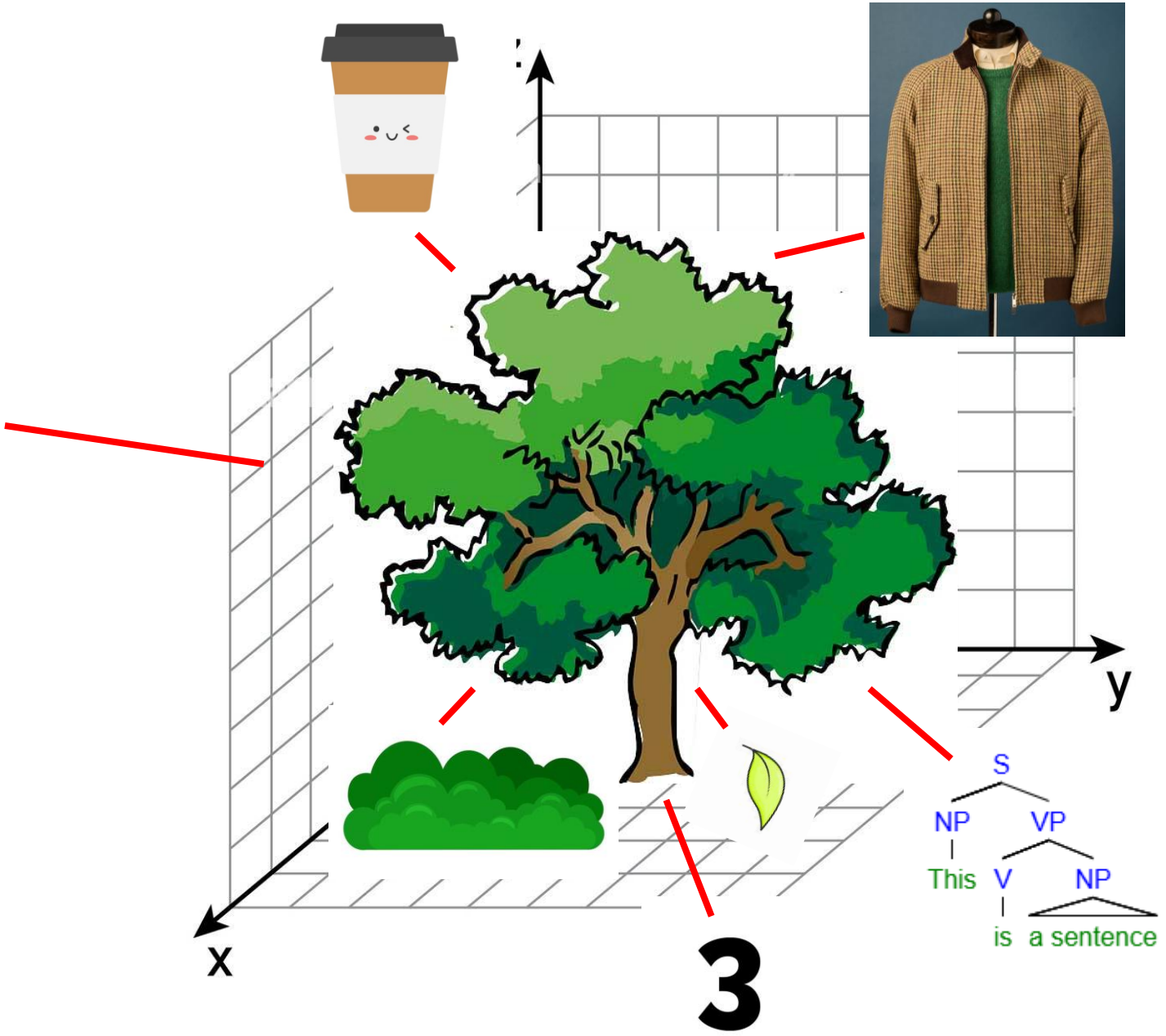




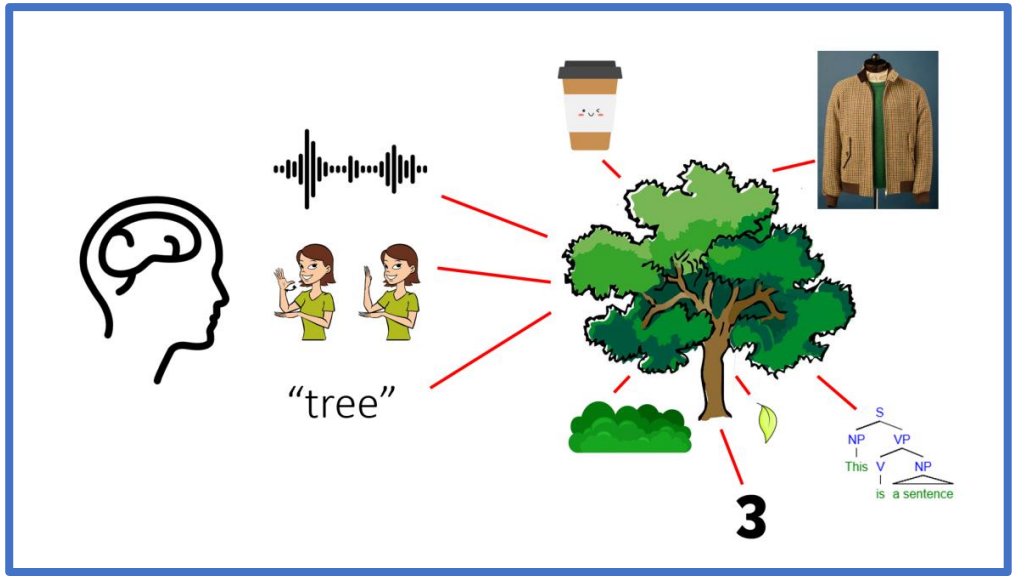
3



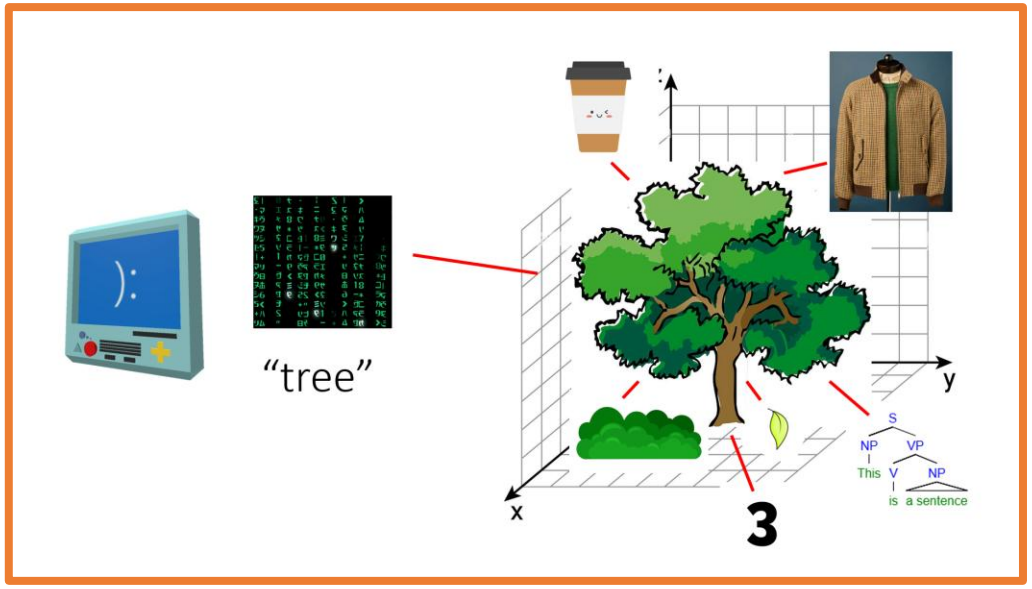
“tree”



Human

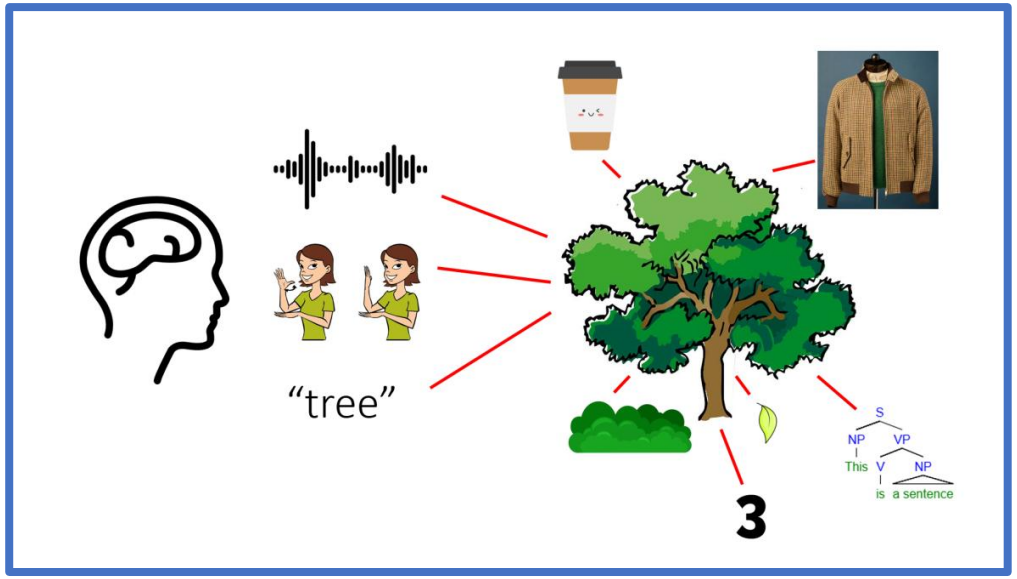


Model

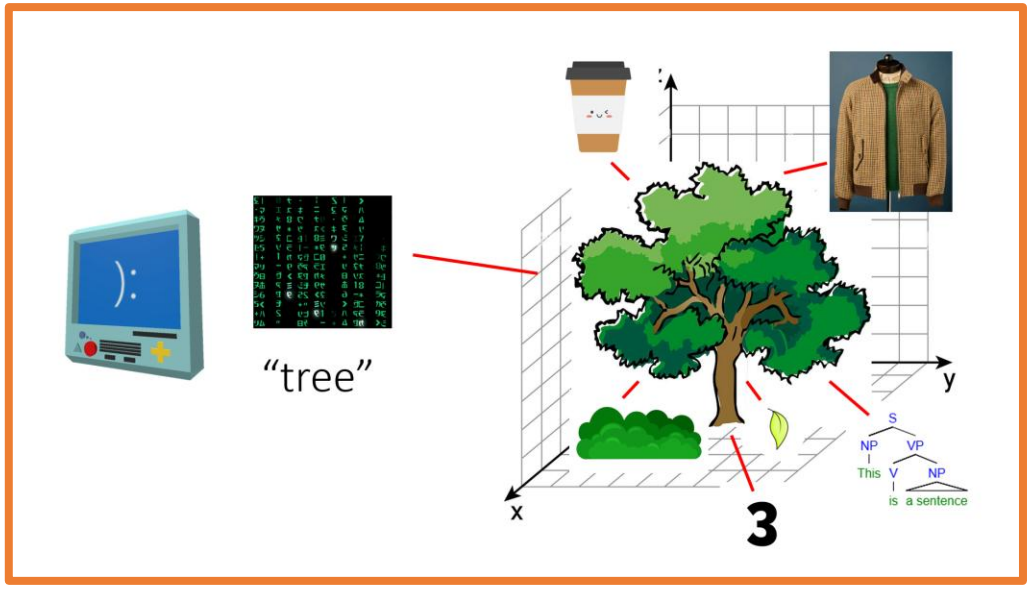


1. Interpretable
2. Robust

Human



Model



1. Interpretable
2. Robust

2. How do we interpret representations?

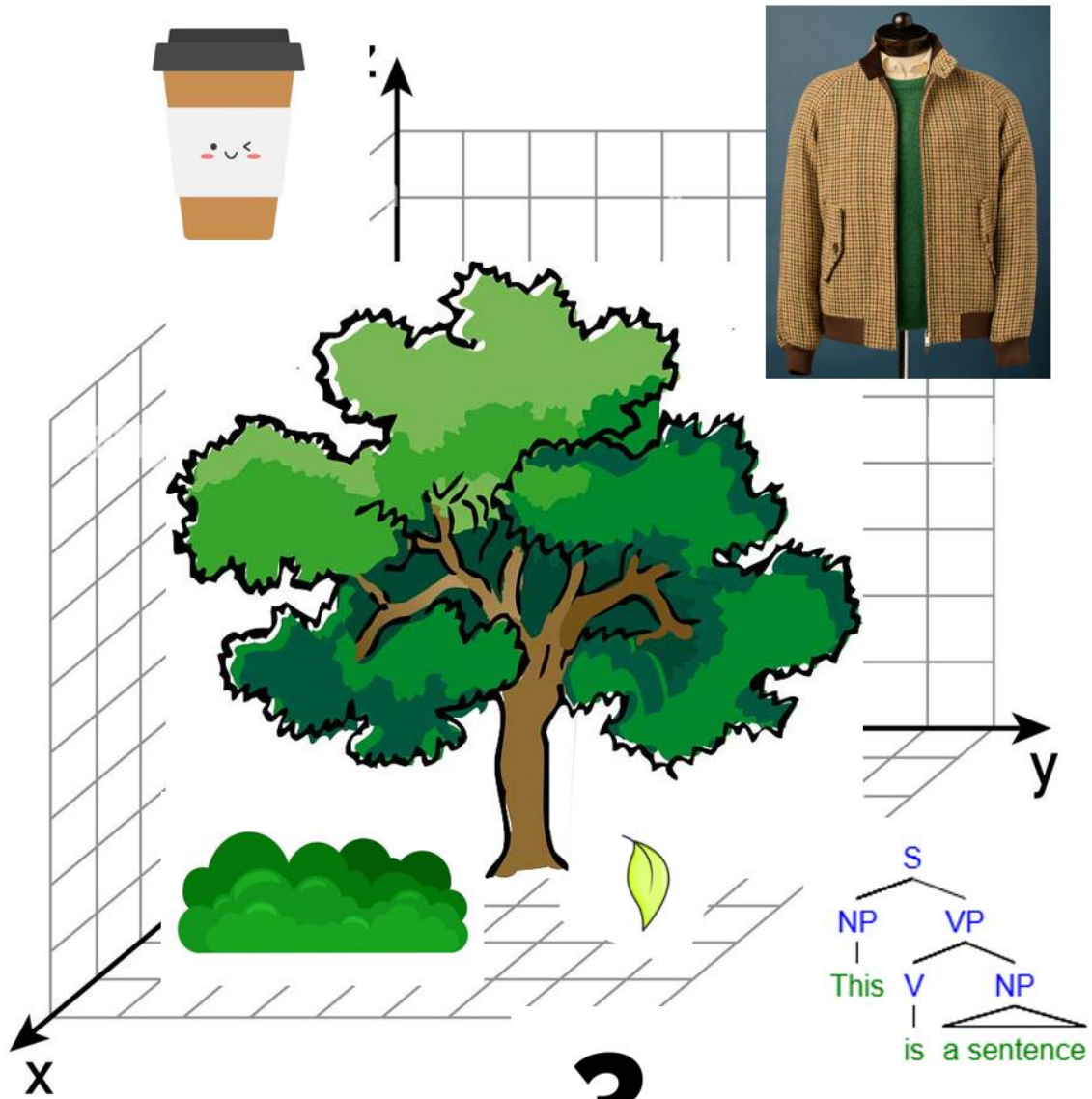
All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality

William Timkey and **Marten van Schijndel**
Department of Linguistics
Cornell University
{wpt25|mv443}@cornell.edu

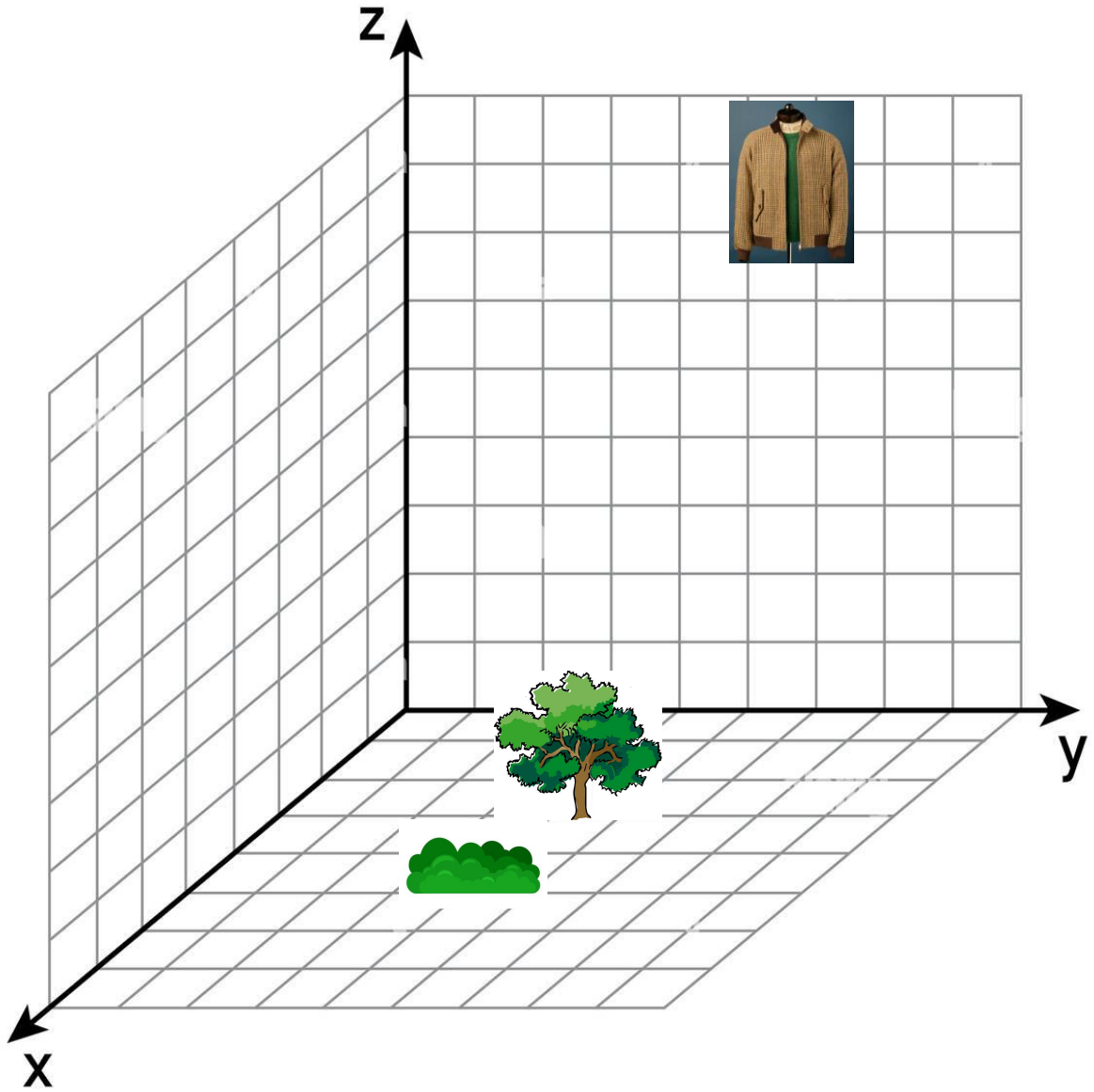
Will
(with long hair)

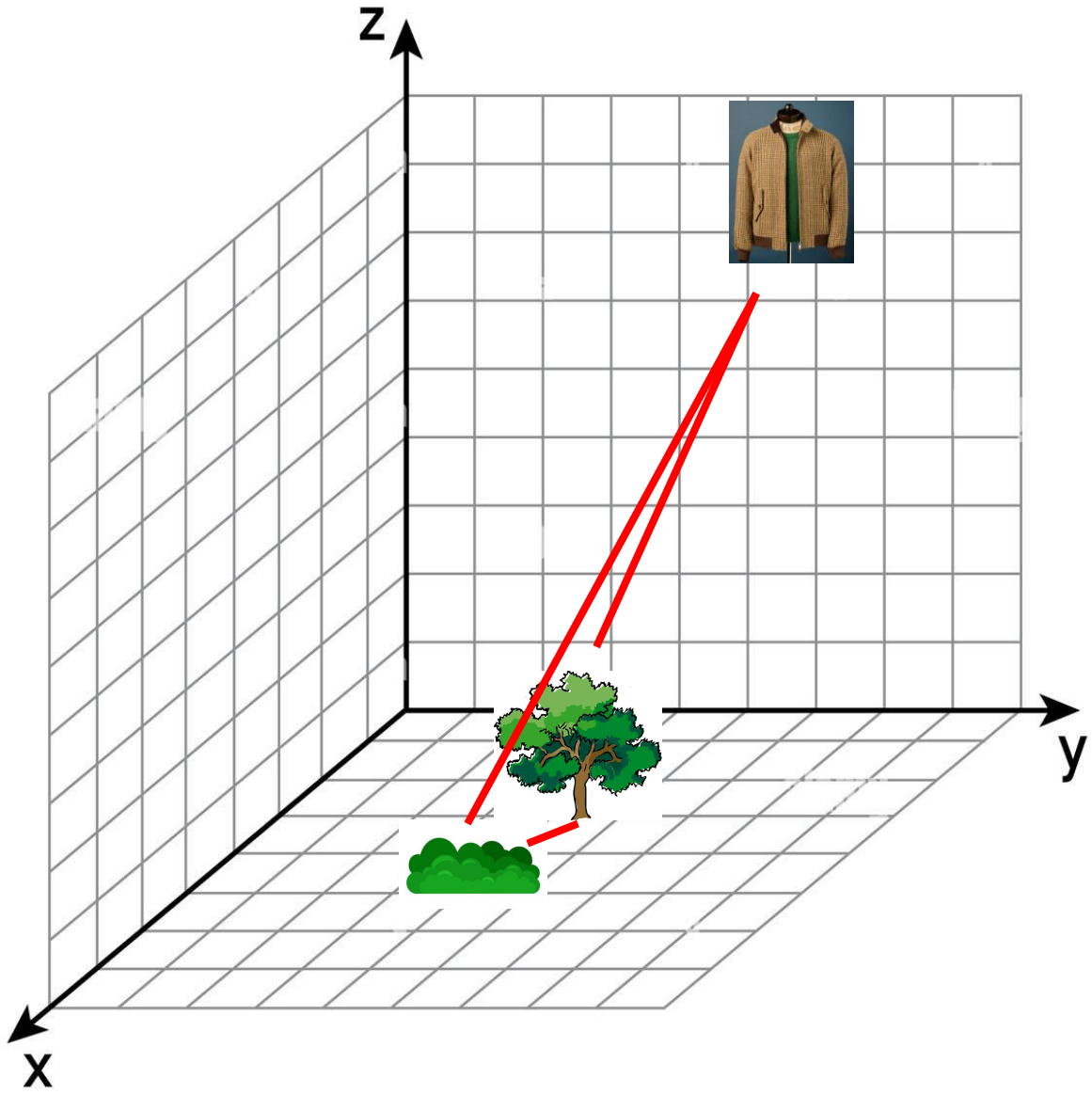
EMNLP 2021

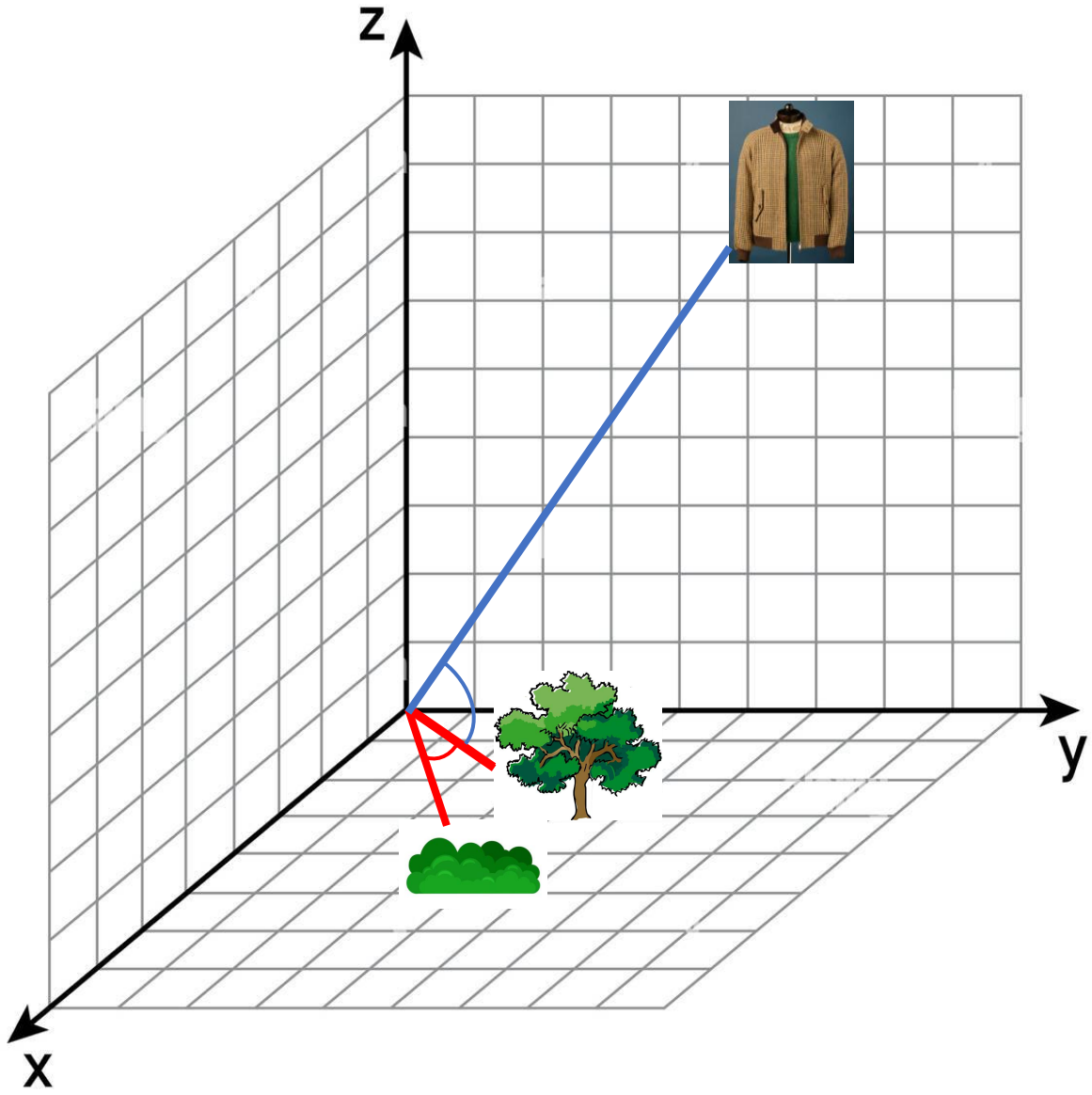
Marty
(with normal hair)



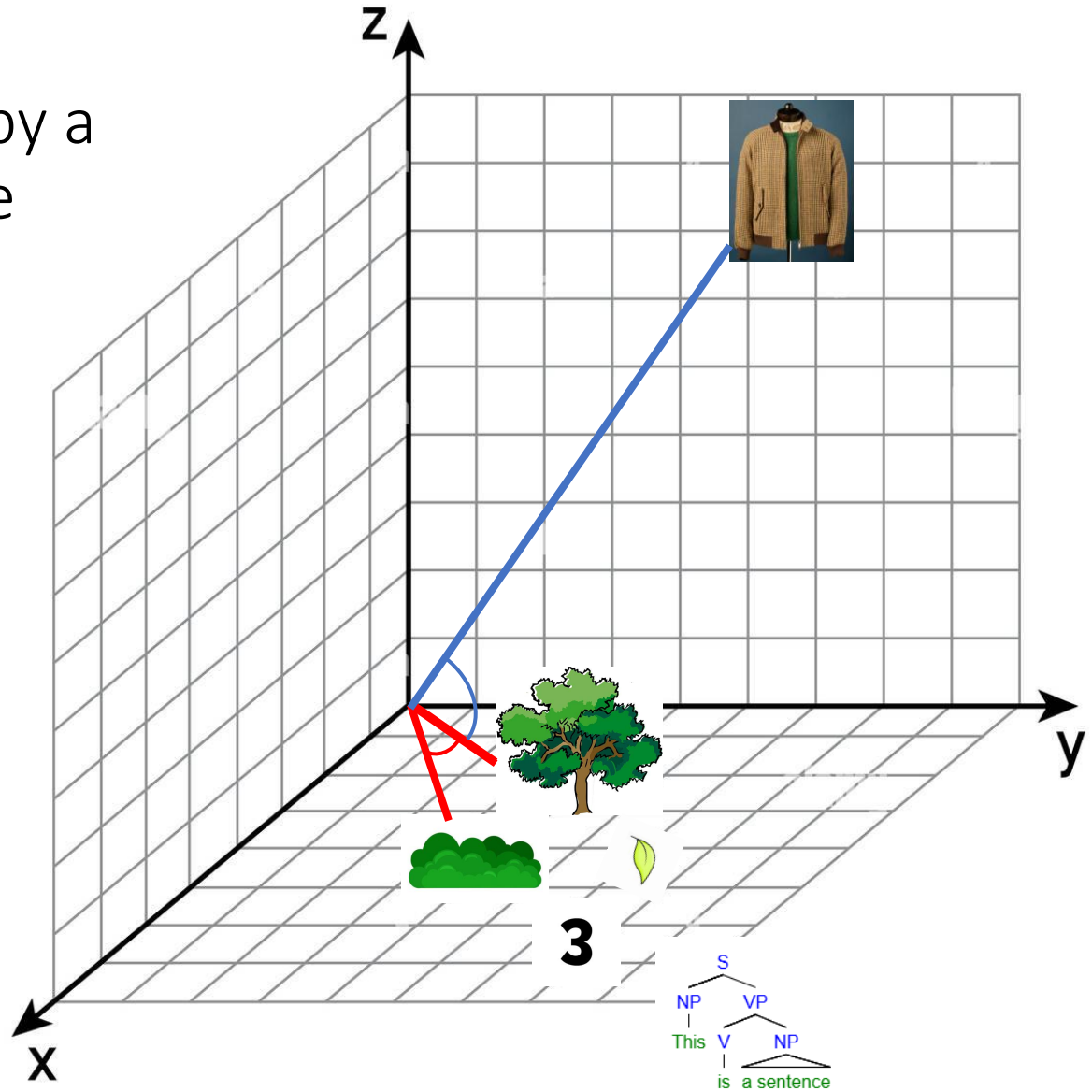
3


z
y
x


z
y
x


z
y
x

Anisotropy: representations occupy a (small) region in the space due to a few dimensions, so (nearly) everything is similar.



Mimno & Thompson (2017); Ethayarajh (2019)

Checking for anisotropy:

Cosine similarity:

$$\cos(u,v) = \frac{u \cdot v}{\|u\|\|v\|} = \sum_{i=1}^{d} \frac{u_i v_i}{\|u\|\|v\|}$$

… can be broken down into a dimension-wise product:

$$CC_i(u,v) = \frac{u_i v_i}{\|u\|\|v\|}$$

… and we can measure the relative contribution of each dimension across a corpus:

$$CC(f_\ell^i) = \frac{1}{n} \cdot \sum_{\{x_\alpha, y_\alpha\} \in S} CC_i(f_\ell(x_\alpha), f_\ell(y_\alpha))$$

Model	Layer	1	2	3	$\hat{A}(f_\ell)$
GPT-2					
BERT					
RoBERTa					
XLNet					
Word2Vec					
GloVe					

$$CC(f_\ell^i) = \frac{1}{n} \cdot \sum_{\{x_\alpha, y_\alpha\} \in S} CC_i(f_\ell(x_\alpha), f_\ell(y_\alpha))$$

Model	Layer	1	2	3	$\hat{A}(f_\ell)$
GPT-2					
BERT					
RoBERTa					
XLNet					
Word2Vec		0.031	0.023	0.023	0.130
GloVe		0.105	0.096	0.095	0.104

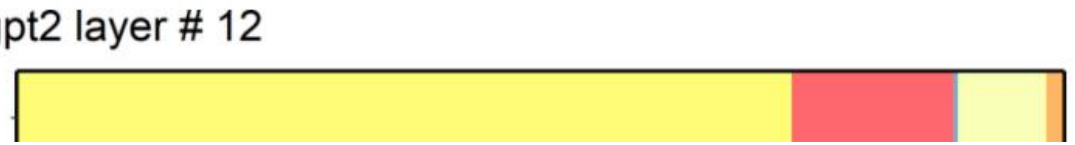
$$CC(f_\ell^i) = \frac{1}{n} \cdot \sum_{\{x_\alpha, y_\alpha\} \in S} CC_i(f_\ell(x_\alpha), f_\ell(y_\alpha))$$

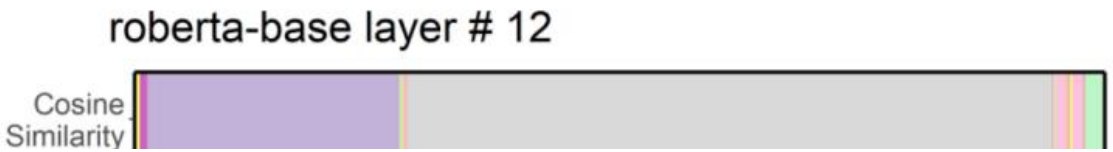
The vector space is very anisotropic…

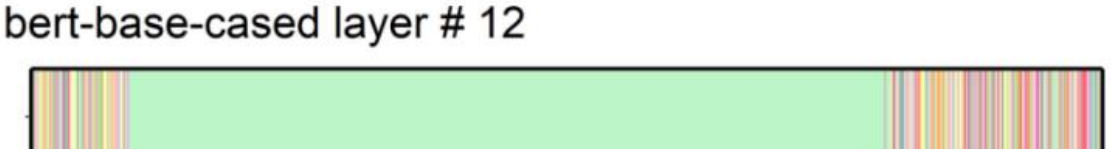
Model	Layer	1	2	3	$\hat{A}(f_\ell)$
GPT-2	11	0.275	0.269	0.265	0.640
	12	0.763	0.131	0.078	0.885
BERT	10	0.817	0.004	0.003	0.396
	11	0.884	0.003	0.002	0.506
RoBERTa	7	0.726	0.193	0.032	0.705
	12	0.663	0.262	0.020	0.745
XLNet	10	0.990	0.000	0.000	0.887
	11	0.996	0.001	0.000	0.981
Word2Vec		0.031	0.023	0.023	0.130
GloVe		0.105	0.096	0.095	0.104

$$CC(f_\ell^i) = \frac{1}{n} \cdot \sum_{\{x_\alpha, y_\alpha\} \in S} CC_i(f_\ell(x_\alpha), f_\ell(y_\alpha))$$

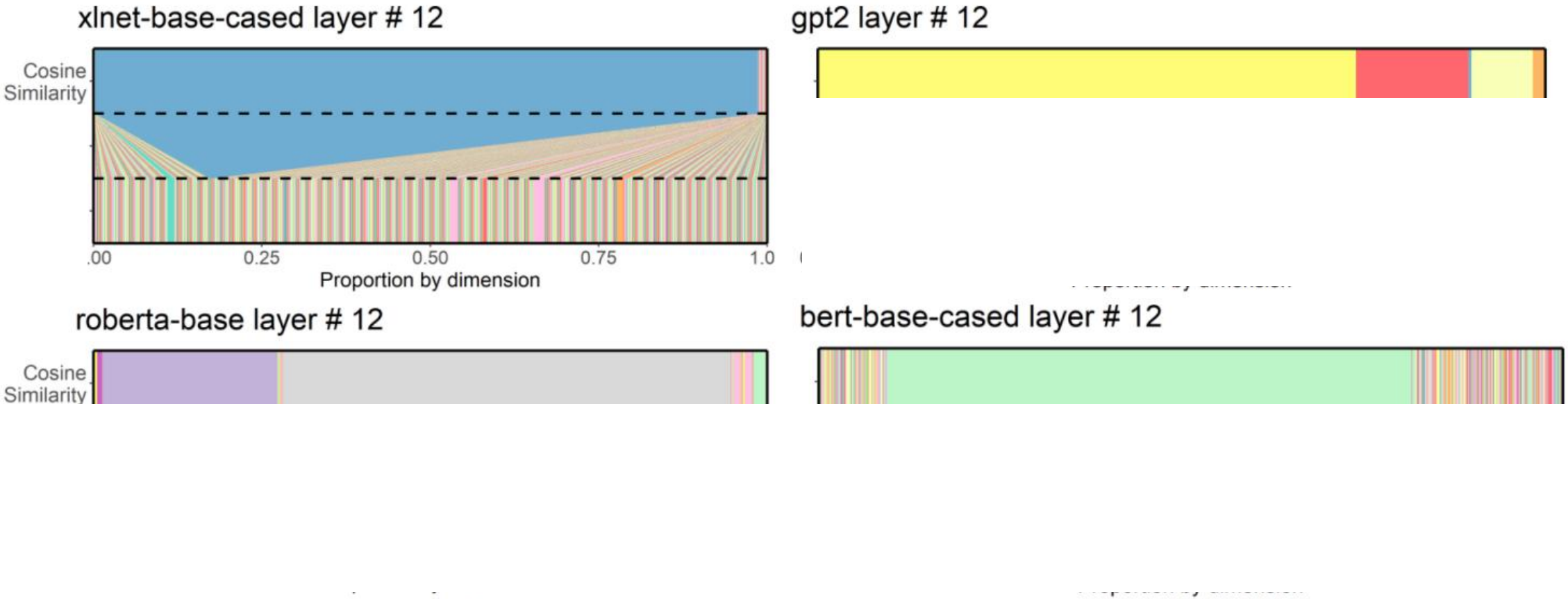
… meaning cosine only uses 1-5 dimensions:







… meaning cosine only uses 1-5 dimensions:



Standardization is a possible fix:

For each dimension:

$$\mu = \frac{1}{|\mathcal{O}|} \cdot \sum_{x \in \mathcal{O}} x \qquad \sigma = \sqrt{\frac{1}{|\mathcal{O}|} \cdot \sum_{x \in \mathcal{O}} (x - \mu)^2}$$

Mean vector

Standard deviation

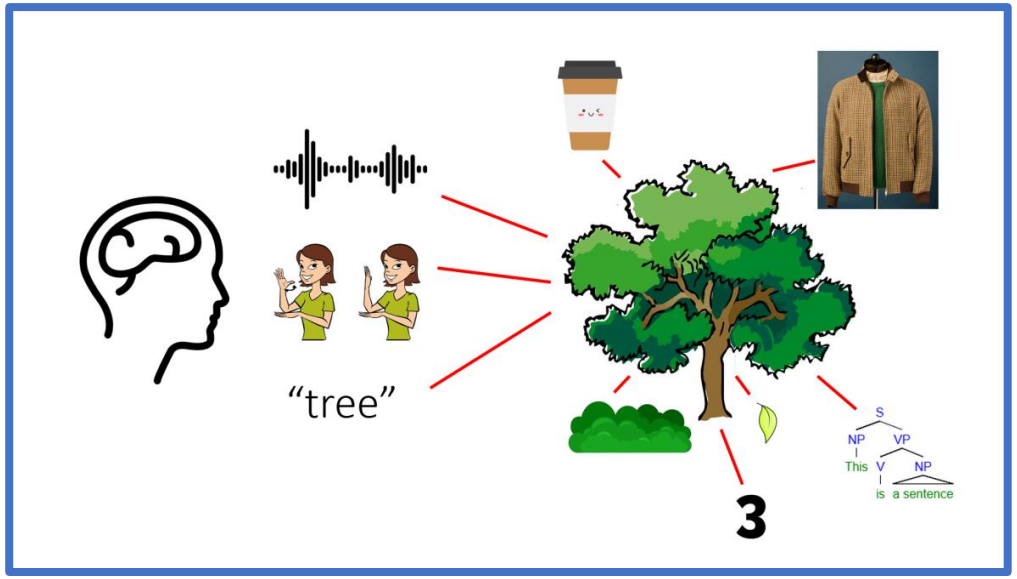
$$z = \frac{x - \mu}{\sigma}$$

Z-scored vector

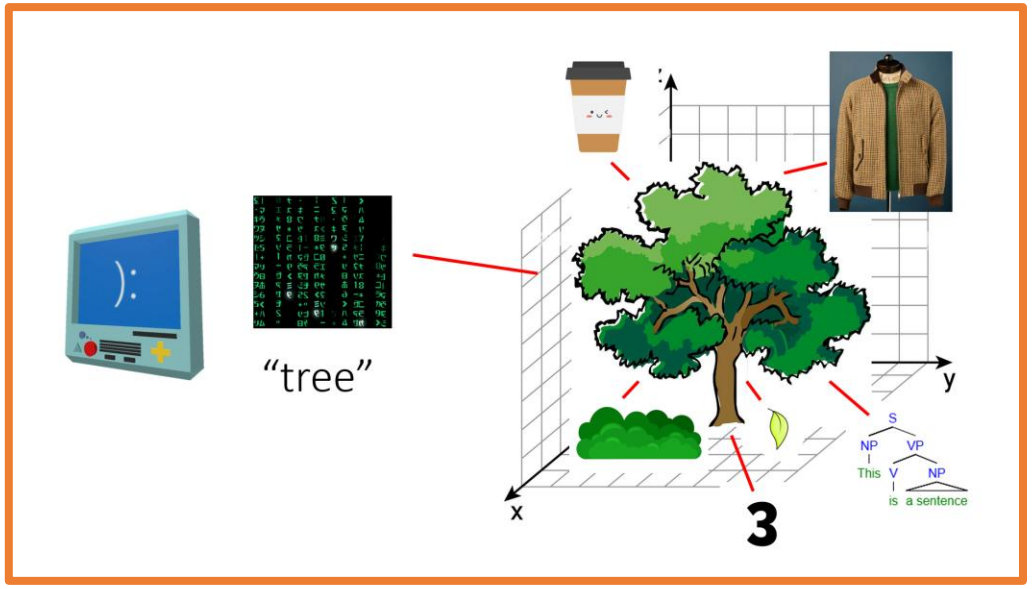
Takeaway:

Transformer representation spaces are highly anisotropic, and raw cosine similarity is not a reliable similarity measure.

Human

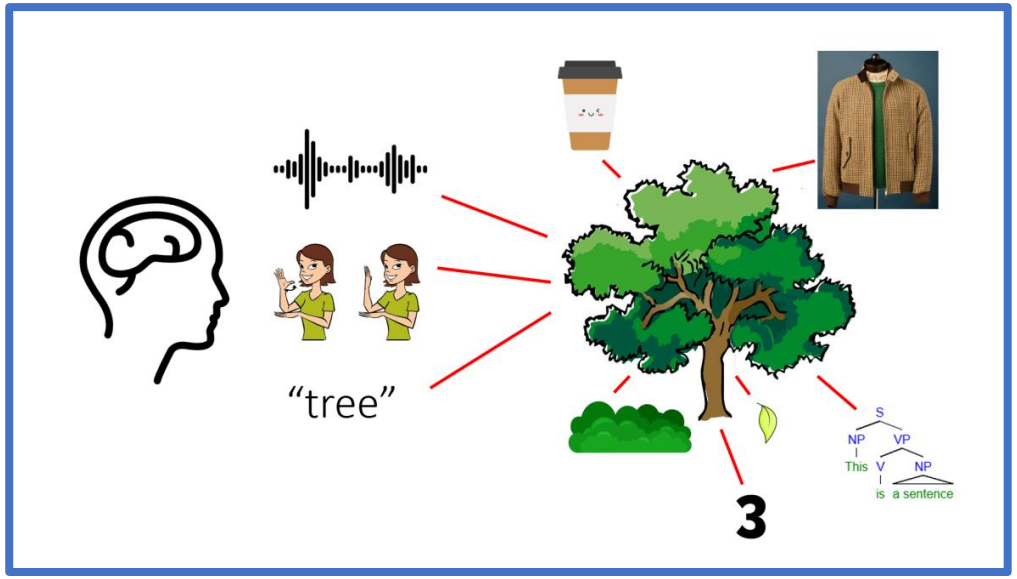


Model

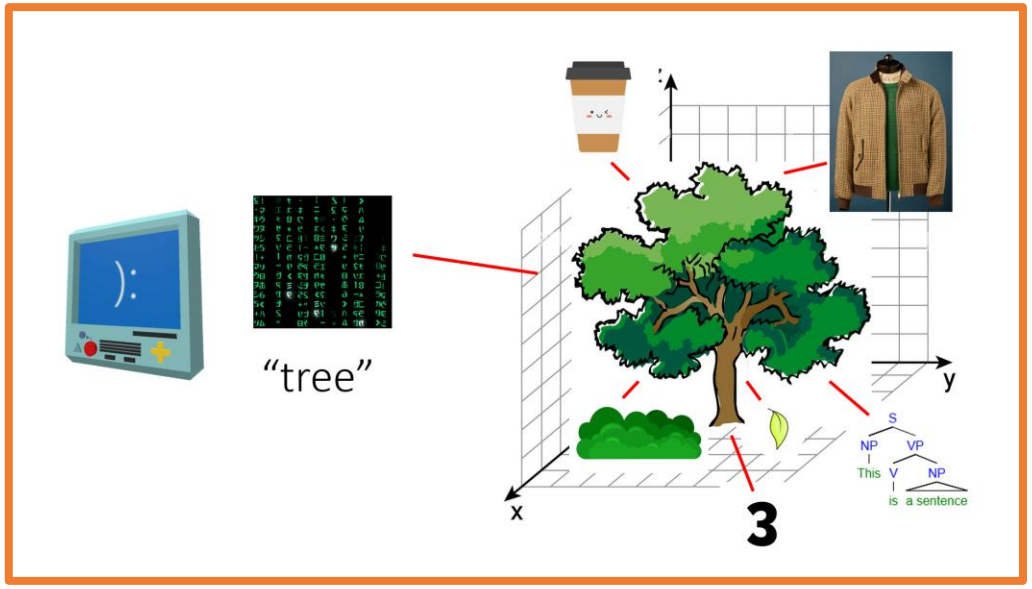


1. Interpretable
2. Robust

Human



Model



1. Interpretable
2. Robust

3. How do we test for “robust” representations?

Semantics or spelling? Probing contextual word embeddings with orthographic noise

Jacob A. Matthews John R. Starr Marten van Schijndel
Cornell University
{jam963, jrs673, mv443}@cornell.edu

Jacob
(with long hair)

ACL 2024
Bangkok, Thailand

Marty

Stars with cooler outer atmospheres , including the Sun , can form various diatomic and polyatomic molecules . = = = Diameter = = = Due to their great distance from the Earth , all stars except the Sun appear to the unaided eye as shining points in the night sky that twinkle because of the effect of the Earth 's atmosphere . The Sun is also a star , but it is close enmugh to the Earth to appear as a disk instead , and to provide daylight . Other than the Sun , the star with the largest

Stars with cooler outer atmospheres , including the Sun , can form various diatomic and polyatomic molecules . = = = Diameter = = = Due to their great distance from the Earth , all stars except the Sun appear to the unaided eye as shining points in the night sky that twinkle because of the effect of the Earth 's atmosphere . The Sun is also a star , but it is close **enmugh** to the Earth to appear as a disk instead , and to provide daylight . Other than the Sun , the star with the largest

enough

Stars with cooler outer atmospheres , including the Sun , can form various diatomic and polyatomic molecules . = = = Diameter = = = Due to their great distance from the Earth , all stars except the Sun appear to the unaided eye as shi points in the night sky that twinkle be the effect of the Earth 's atmosphere . Th also a star , but is close to t to appear as a disk instead , and to daylight . Other than the Sun , the star largest

en-mu-gh

The Set-Up:

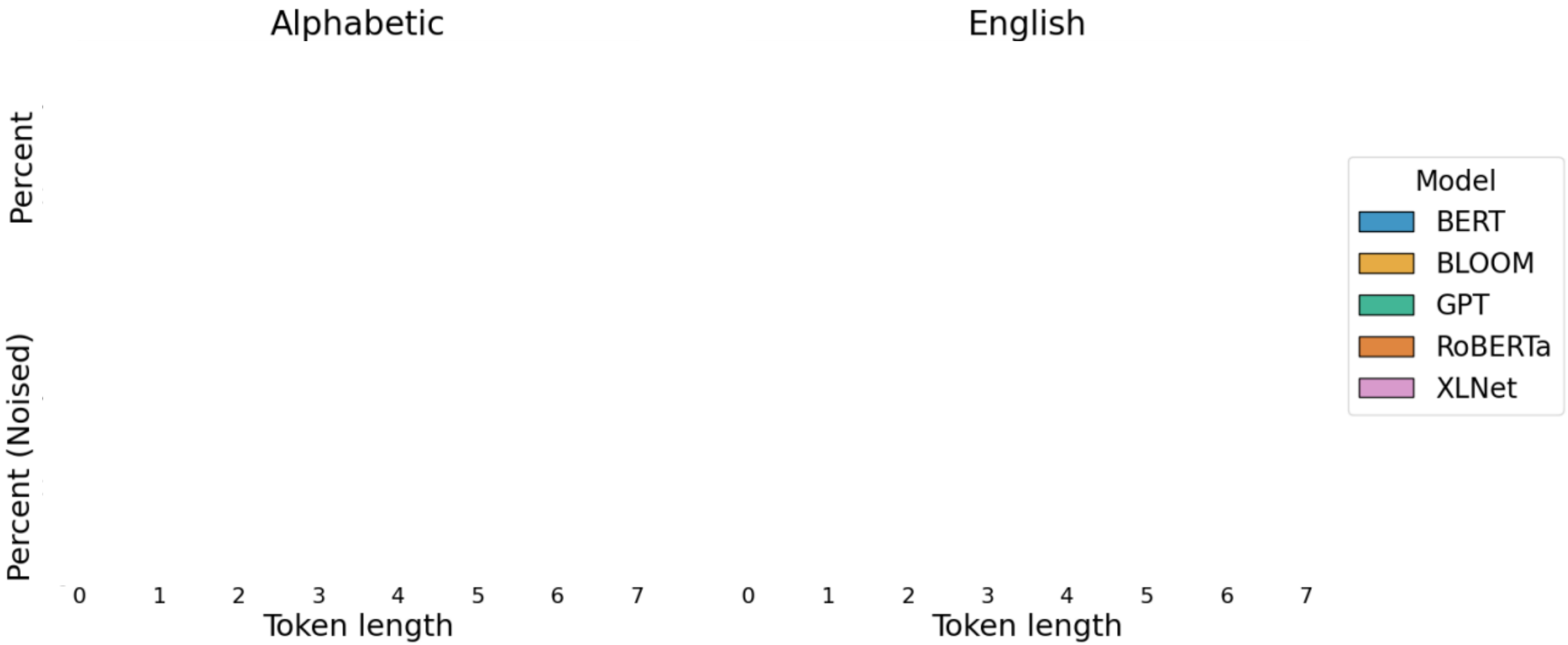
Model	Word	Edited	Word Tokens	Edited Tokens
GPT-2	contenders	contelders	“contenders”	“cont”, “e”, “ld”, “ers”
BERT			“contender”, “s”	“con”, “tel”, “ders”
XLNet			“contenders”	“con”, “tel”, “der”, “s”

Five models:
1. BERT
2. BLOOM
3. GPT-2
4. RoBERTa
5. XLNet

Data: `wikitext-2`

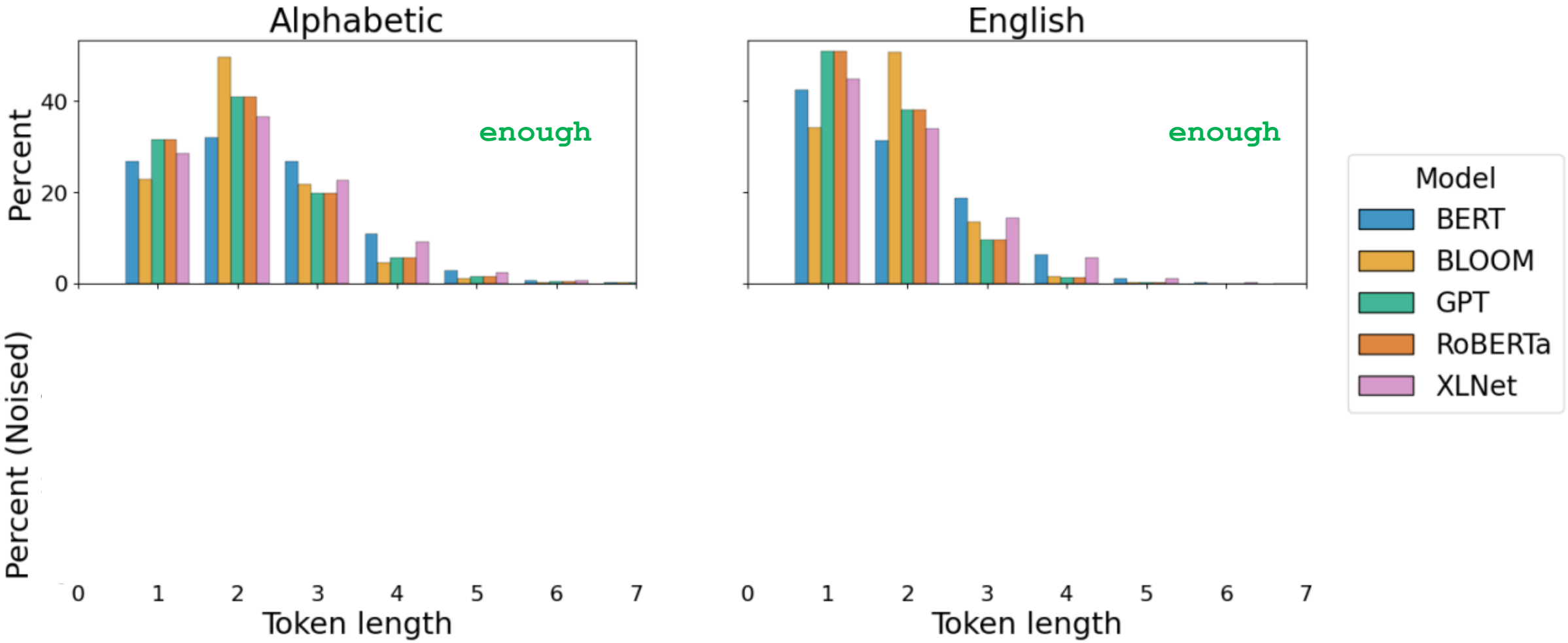
Two kinds of analyses:
1. Distribution shifts
    1. Alphabetic
    2. English
2. Similarity
    1. Without context
    2. With context

Distribution of token length shifts higher...



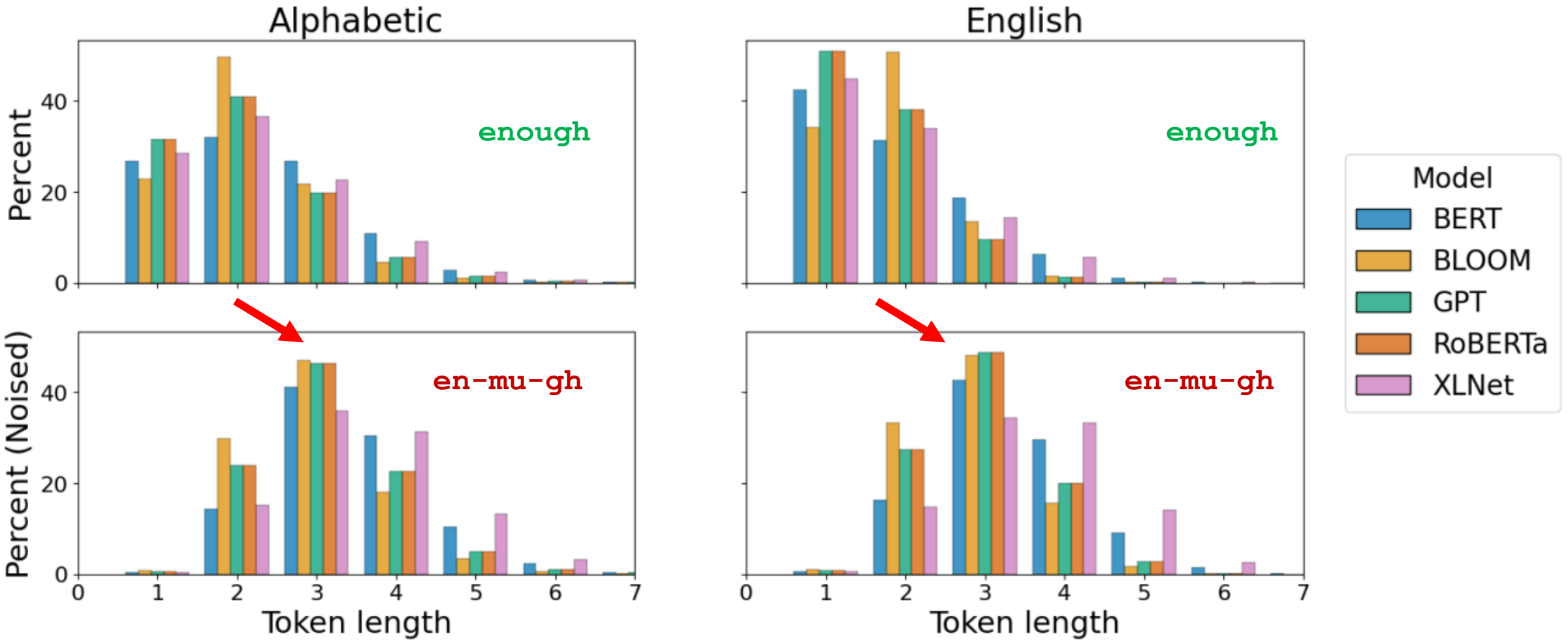
Token Length -> token length of the word

Distribution of token length shifts higher...



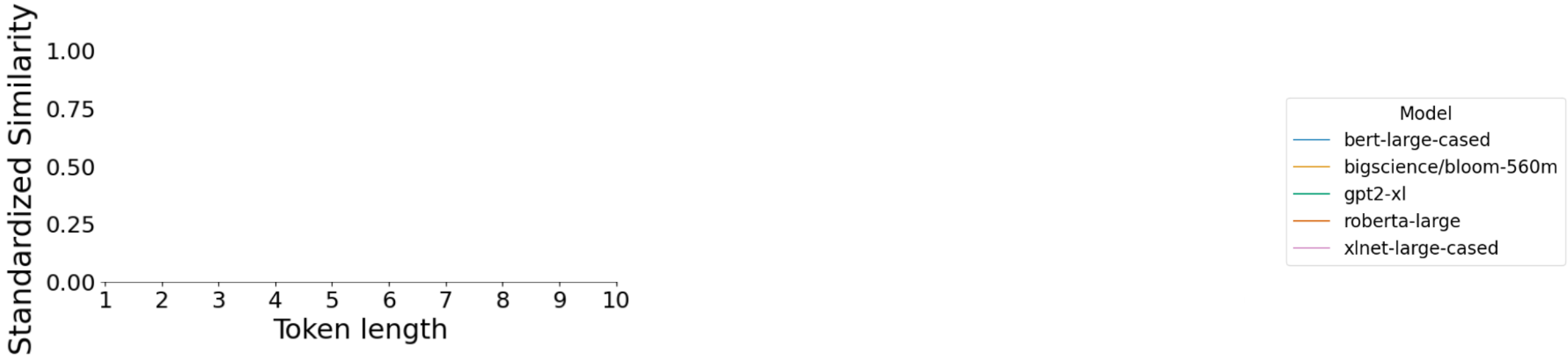
Token Length -> token length of the word

Distribution of token length shifts higher...



Token Length -> token length of the word

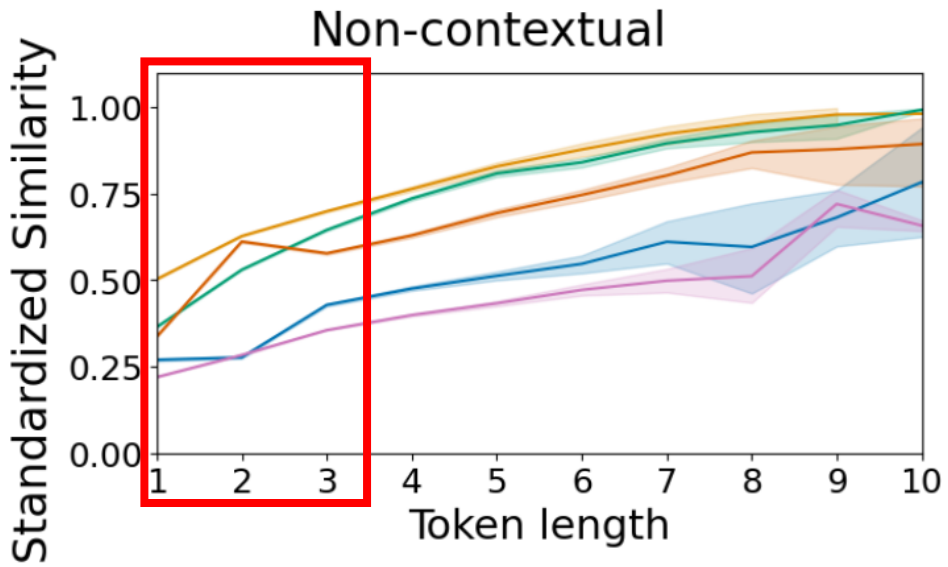
… and words with more tokens are more robust!

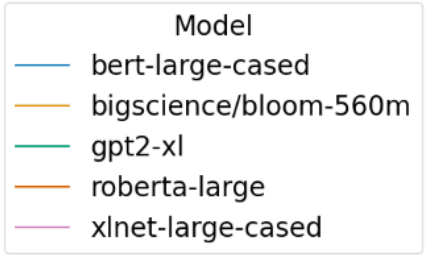


Token Length -> token length of original (unedited) word

Standardized Similarity -> sim(original, edited)

… and words with more tokens are more robust!

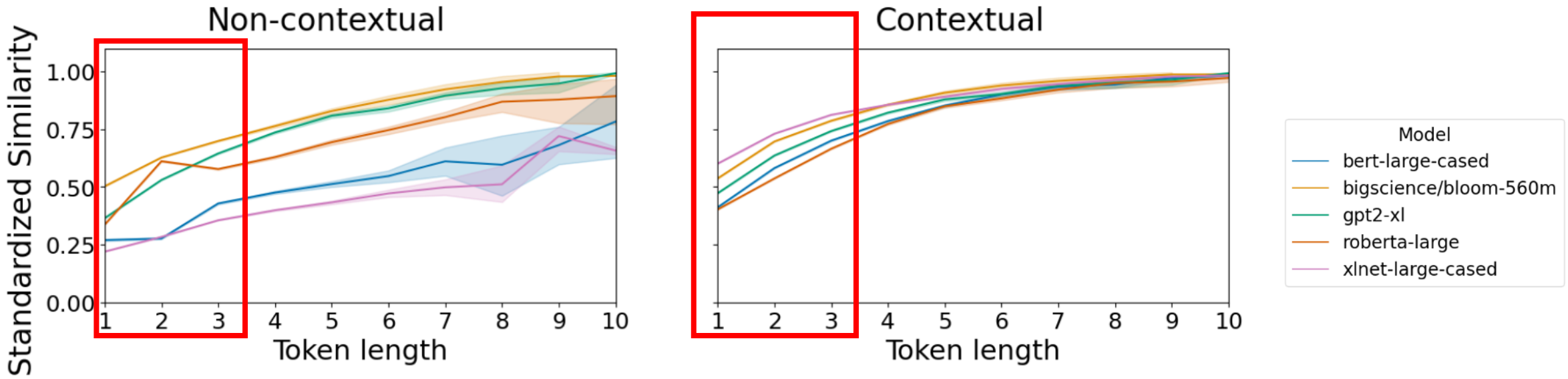




Token Length -> token length of original (unedited) word

Standardized Similarity -> sim(original, unedited)

… and words with more tokens are more robust!

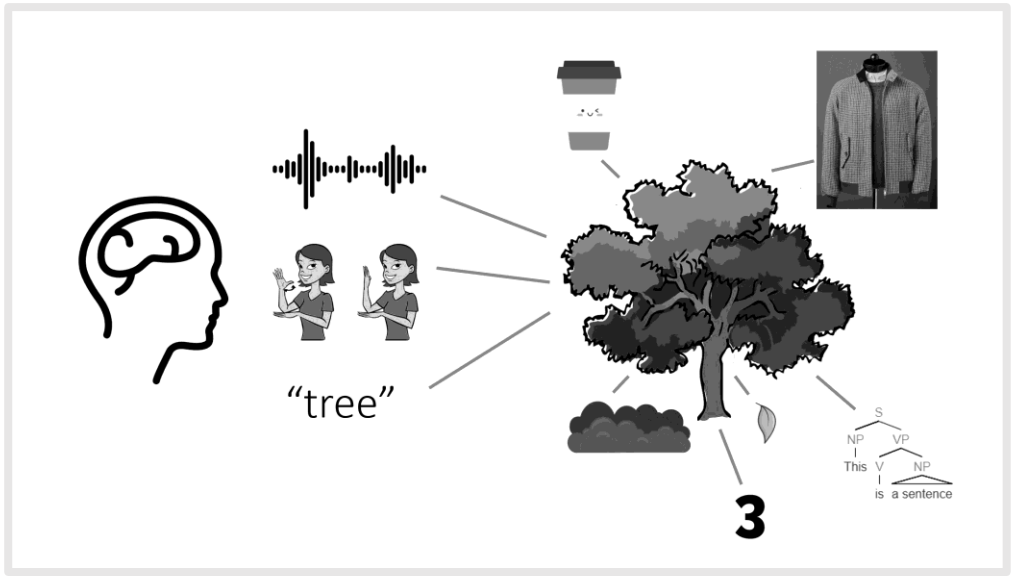


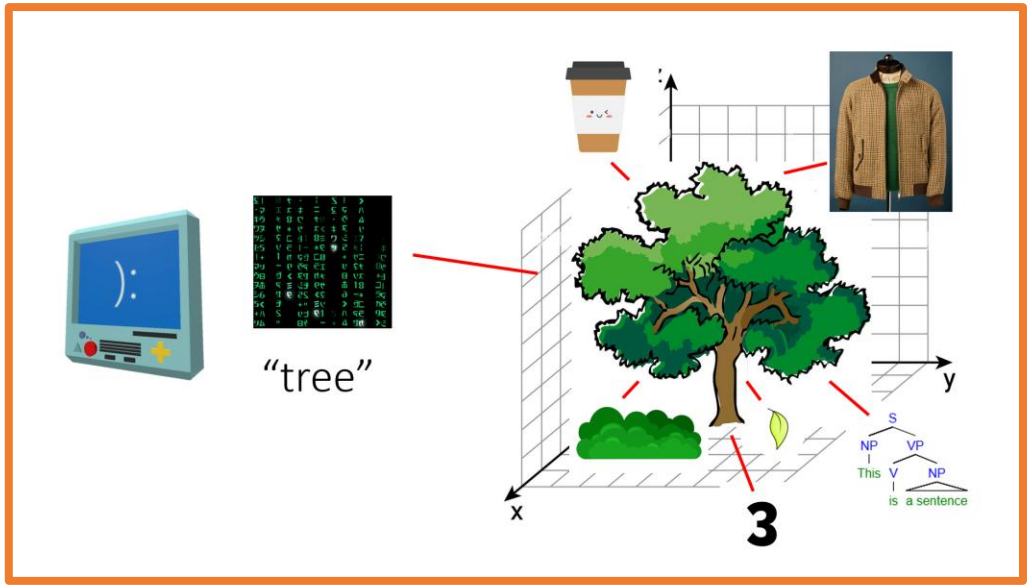
Token Length -> token length of original (unedited) word

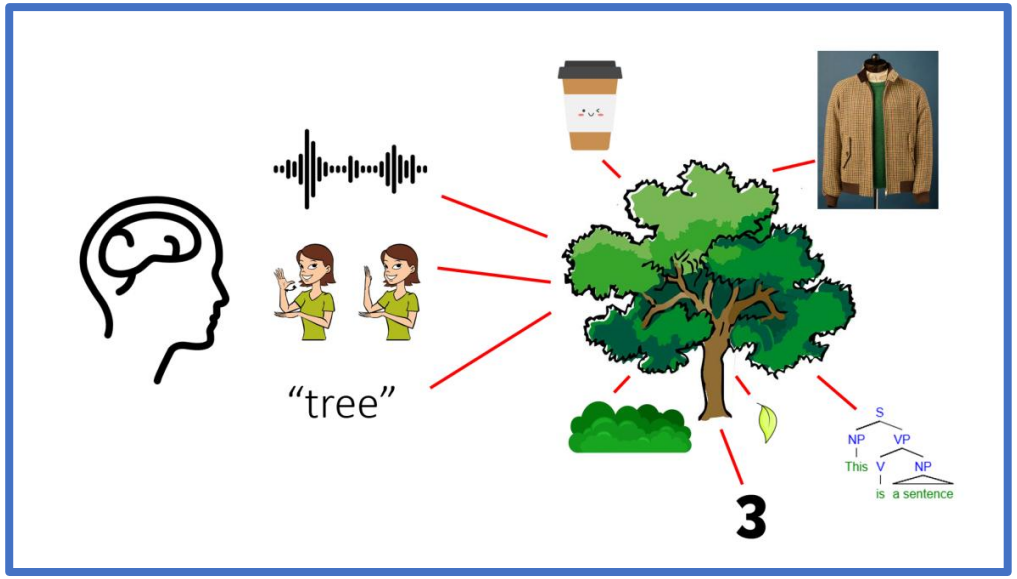
Standardized Similarity -> sim(original, unedited)

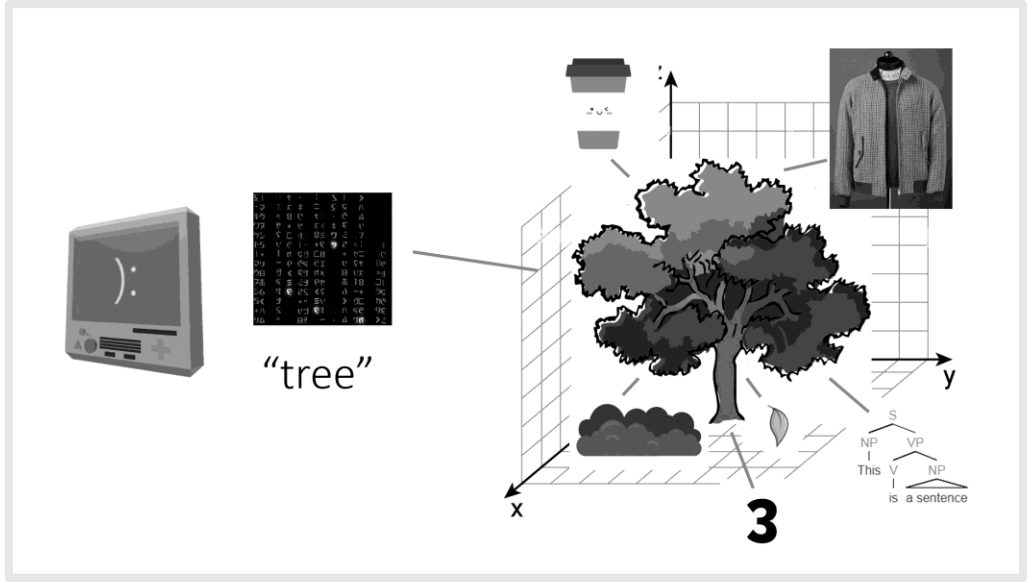
Takeaway:

Much of a word’s semantic identity is lost when a single character is changed, challenging the assumption that CWEs robustly capture word-level semantic information.


"tree"
S
NP
VP
This
V
NP
is
a sentence
3


"tree"
x
y
S
NP
VP
This
V
NP
is
a sentence
3


"tree"
S
NP
VP
This
V
NP
is
a sentence
3


"tree"
x
y
S
NP
VP
This
V
NP
is
a sentence
3

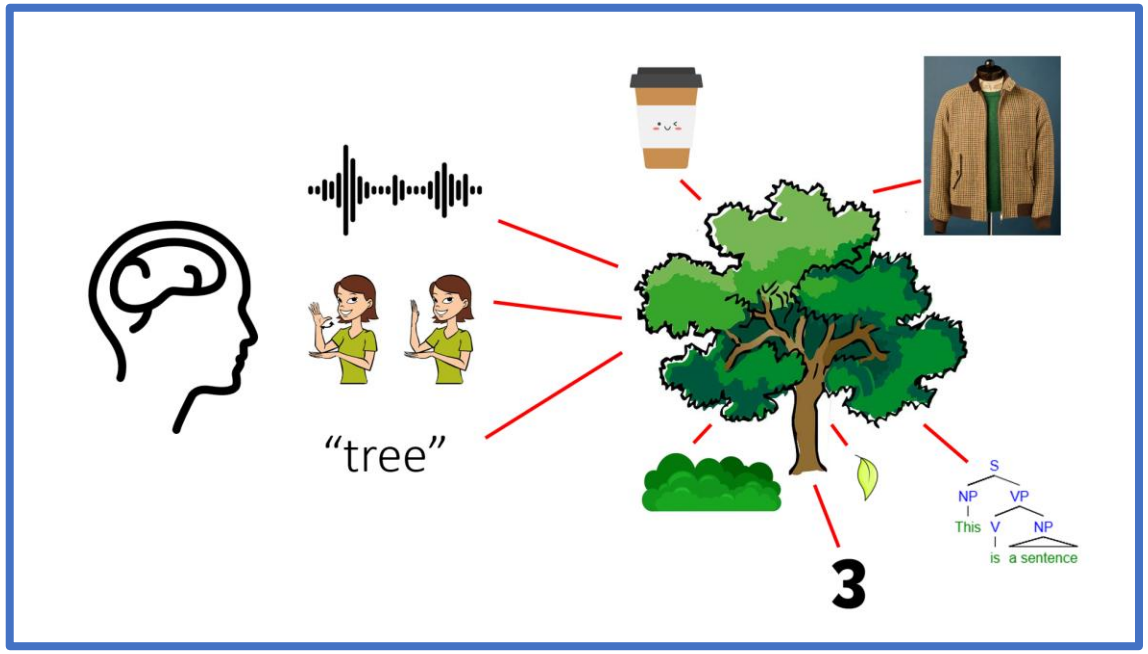
5. How do humans determine similarity?

Me!

Ashlyn
(normal hair)

Marty

Capturing human similarity judgments:



a) How similar are *trees* & *bushes* on dimension X?

b) Are *bushes* or *leaves* more similar to *trees* on dimension X?

The issues with these methods:

1. Judgments are isolated from *context* → similarity is *contextual*
2. Judgments are isolated from *each other* → similarity is *multi-faceted*

Tversky (1977); Goldstone et al. (1991); Medin et al. (1993); *inter alia*

GRIS: A new psycholinguistic paradigm

(Generating Representations In Space)

- Place objects (text, images, audio) onto labeled canvases.
- Tracks when & where each object was placed.
- Many customizable features!*

*Feel free to ask about these features!

Our test bed: The NYT Connections!

KAYAK	SNOW	BUCKS	HAIL
OPTION	TAB	MOM	NETS
LEVEL	RAIN	HEAT	RETURN
JAZZ	SHIFT	RACE CAR	SLEET

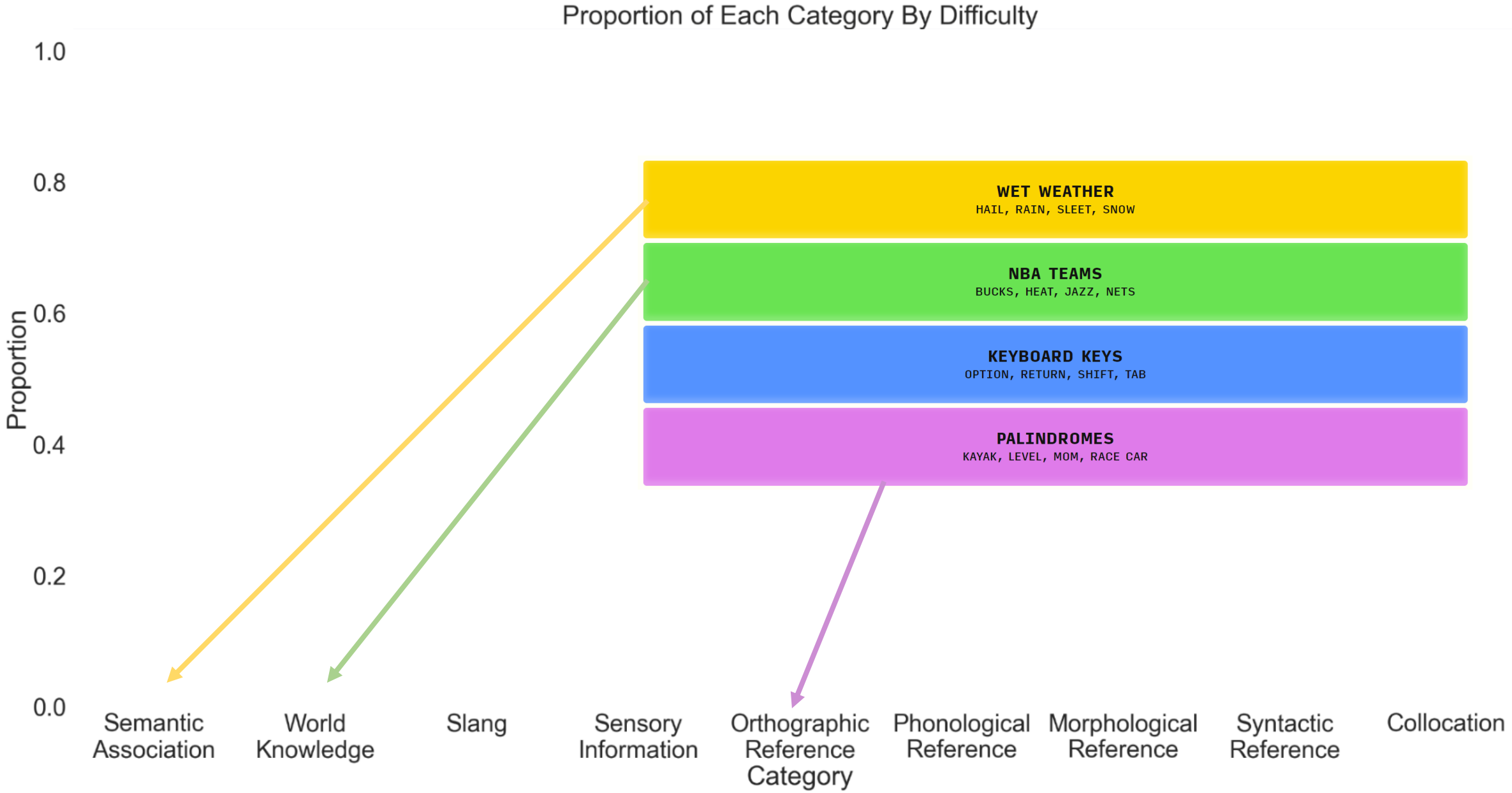
Samadarshi et al. (2024)

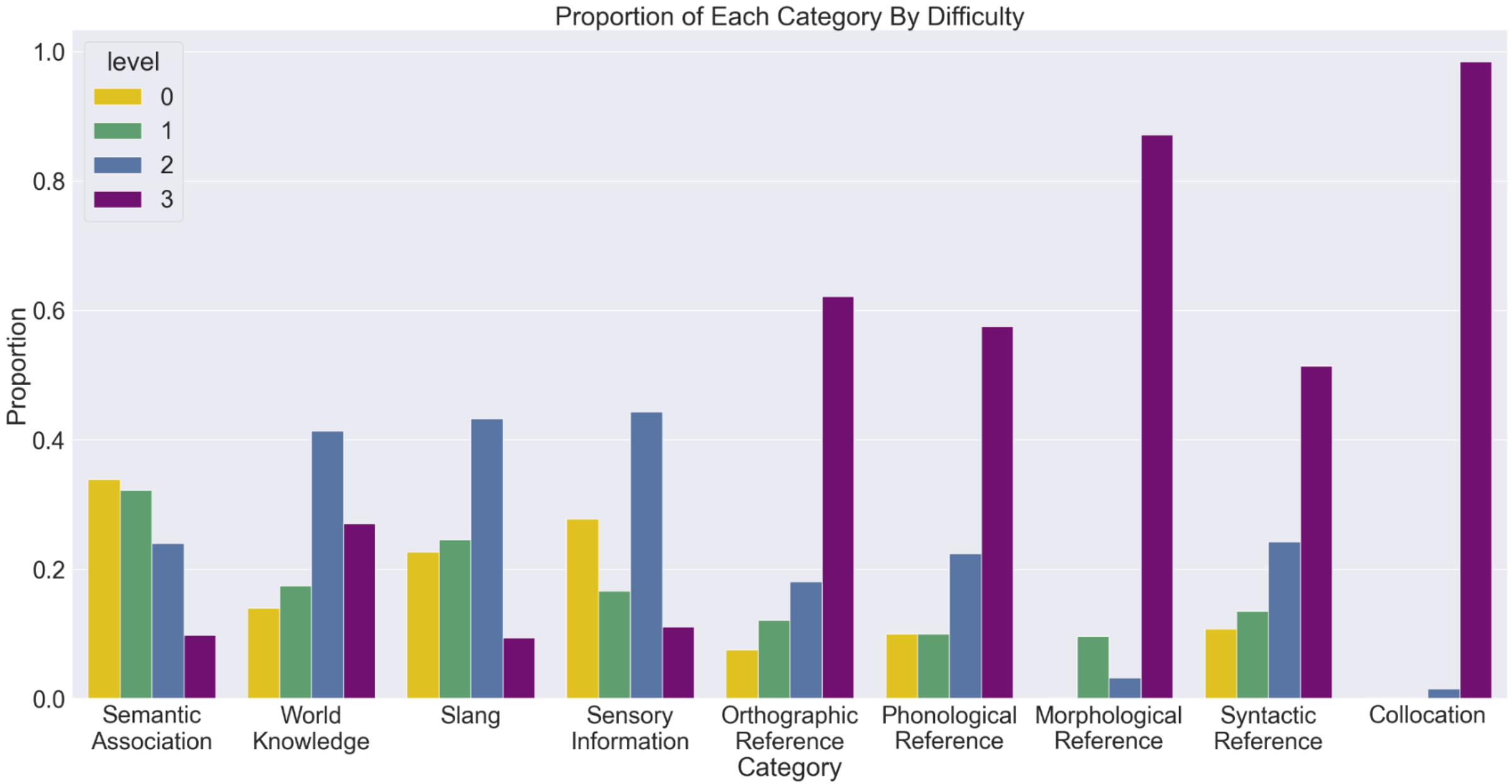
Our test bed: The NYT Connections!

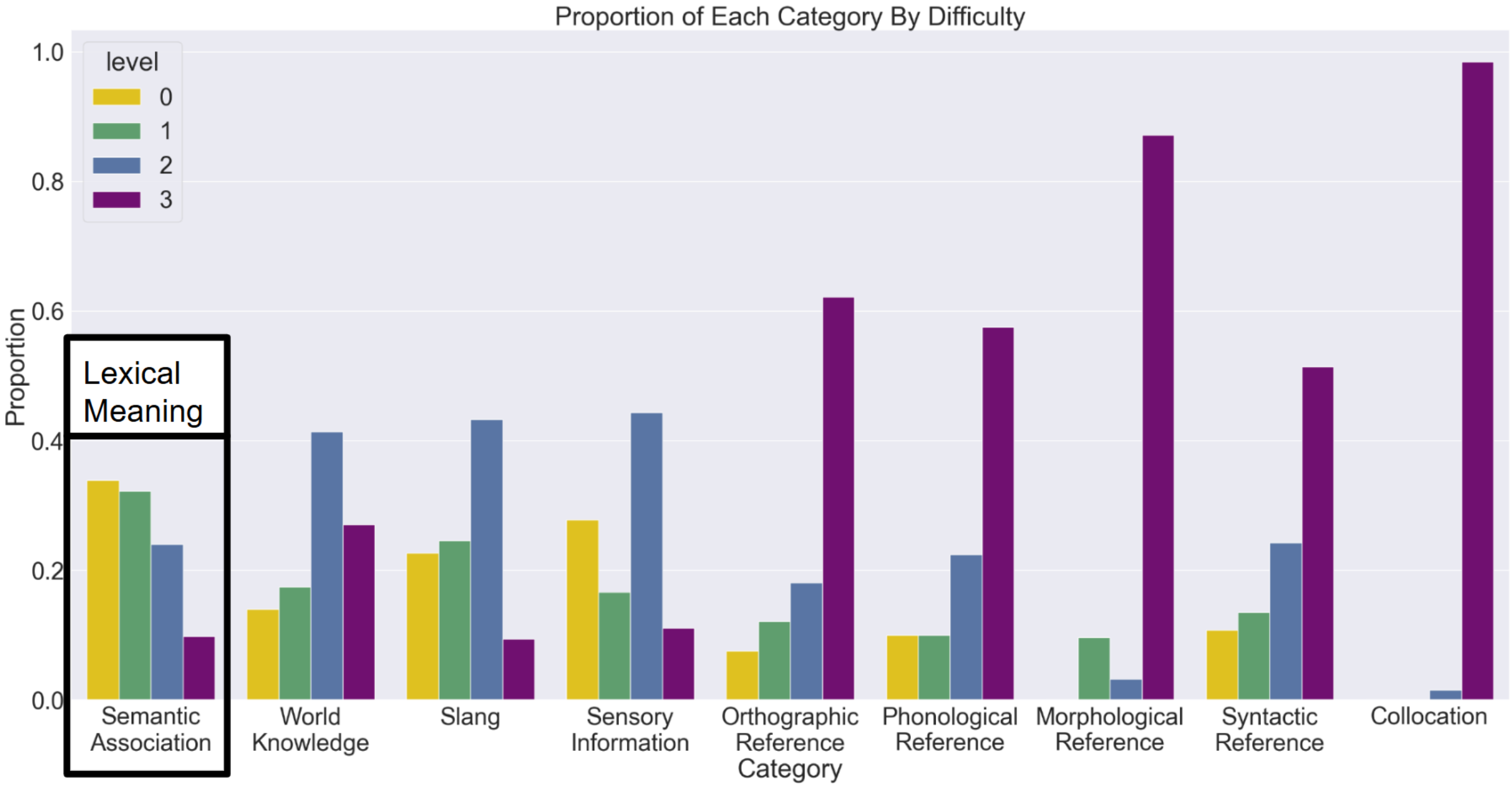
KAYAK	SNOW	BUCKS	HAIL	**WET WEATHER** HAIL, RAIN, SLEET, SNOW
OPTION	TAB	MOM	NETS	**NBA TEAMS** BUCKS, HEAT, JAZZ, NETS
LEVEL	RAIN	HEAT	RETURN	**KEYBOARD KEYS** OPTION, RETURN, SHIFT, TAB
JAZZ	SHIFT	RACE CAR	SLEET	**PALINDROMES** KAYAK, LEVEL, MOM, RACE CAR

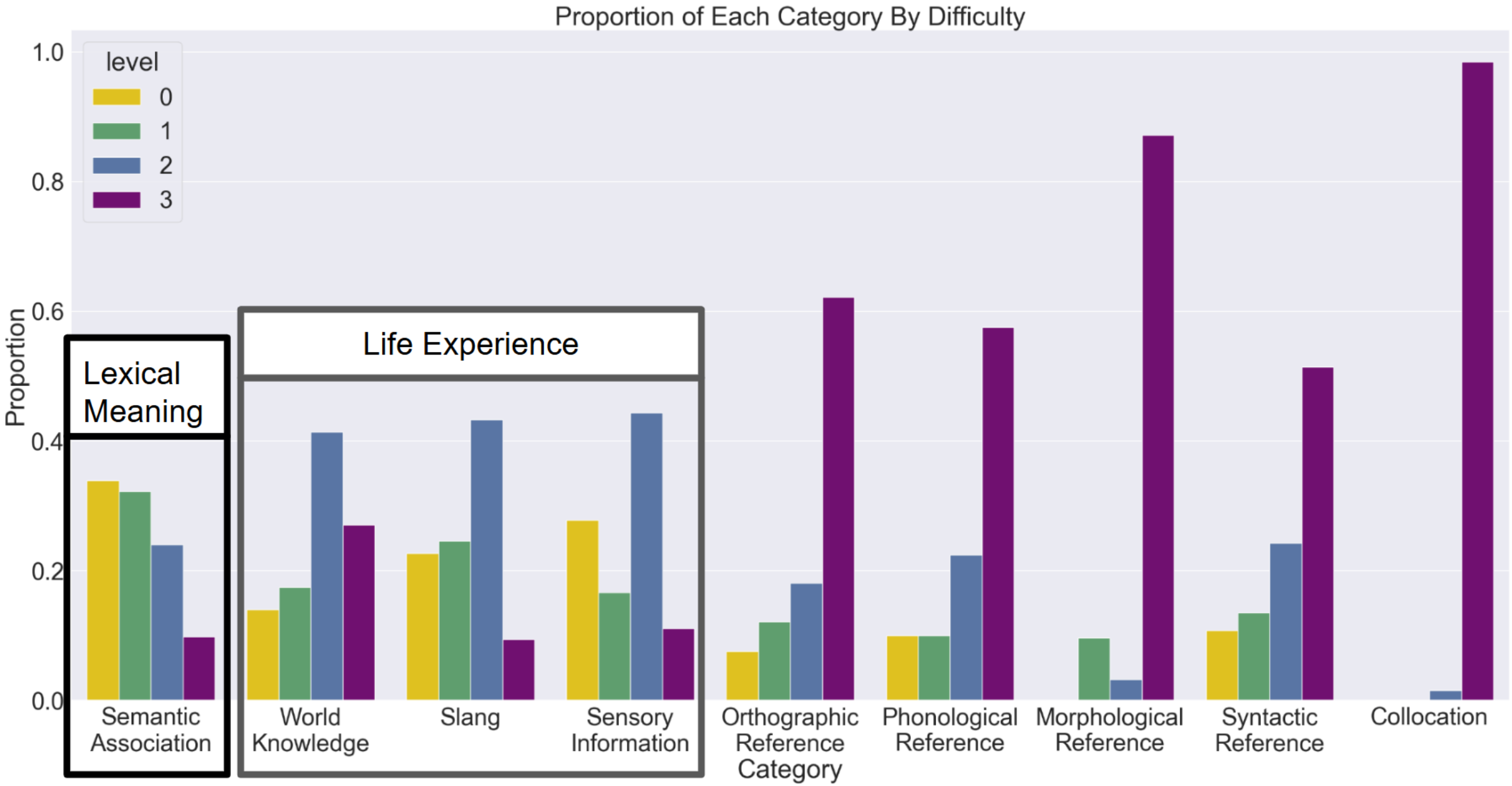
- A very difficult task (for humans and models!)
- >300 puzzles with labels and difficulty (yellow < green < blue < purple)

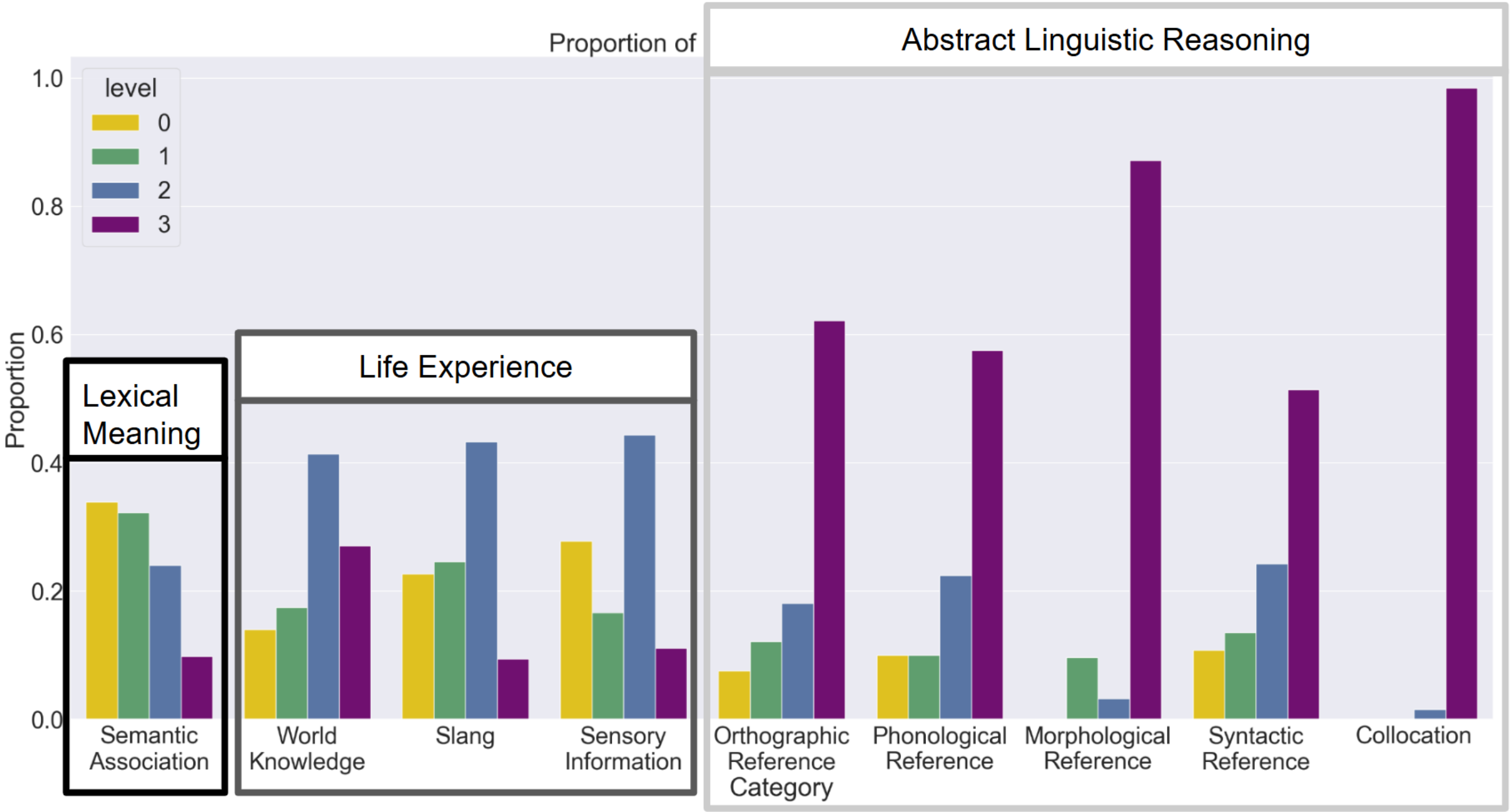
Samadarshi et al. (2024)


Proportion of Each Category By Difficulty
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
WET WEATHER
HAIL, RAIN, SLEET, SNOW
NBA TEAMS
BUCKS, HEAT, JAZZ, NETS
KEYBOARD KEYS
OPTION, RETURN, SHIFT, TAB
PALINDROMES
KAYAK, LEVEL, MOM, RACE CAR
Semantic Association
World Knowledge
Slang
Sensory Information
Orthographic Reference
Phonological Reference
Morphological Reference
Syntactic Reference
Collocation
Category

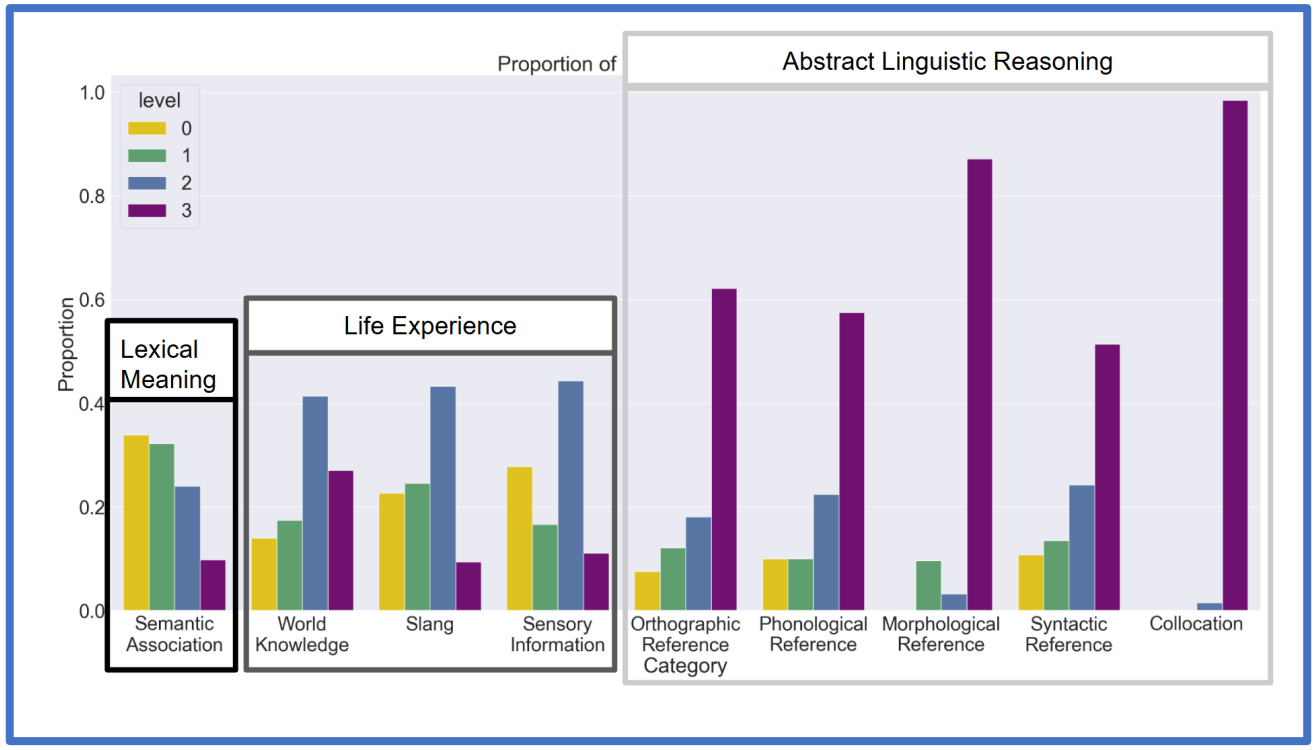

Proportion of Each Category By Difficulty
level
0
1
2
3
Proportion
1.0
0.8
0.6
0.4
0.2
0.0
Semantic Association
World Knowledge
Slang
Sensory Information
Orthographic Reference
Phonological Reference
Morphological Reference
Syntactic Reference
Collocation
Category


Proportion of Each Category By Difficulty
level
0
1
2
3
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
Lexical Meaning
Semantic Association
World Knowledge
Slang
Sensory Information
Orthographic Reference
Phonological Reference
Morphological Reference
Syntactic Reference
Collocation
Category


Proportion of Each Category By Difficulty
level
0
1
2
3
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
Lexical Meaning
Life Experience
Semantic Association
World Knowledge
Slang
Sensory Information
Orthographic Reference
Phonological Reference
Morphological Reference
Syntactic Reference
Collocation
Category


Proportion of
Abstract Linguistic Reasoning
level
0
1
2
3
1.0
0.8
0.6
0.4
0.2
0.0
Proportion
Lexical Meaning
Life Experience
Semantic Association
World Knowledge
Slang
Sensory Information
Orthographic Reference Category
Phonological Reference
Morphological Reference
Syntactic Reference
Collocation

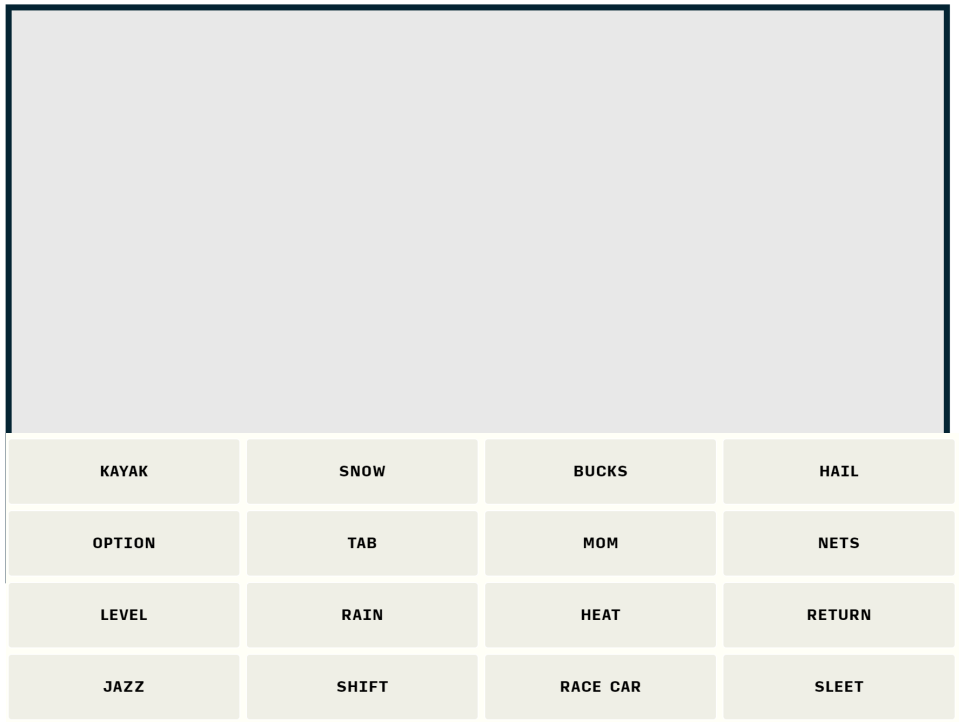
Some similarities look more difficult than others…



1) Are purple categories *actually* more difficult than others?

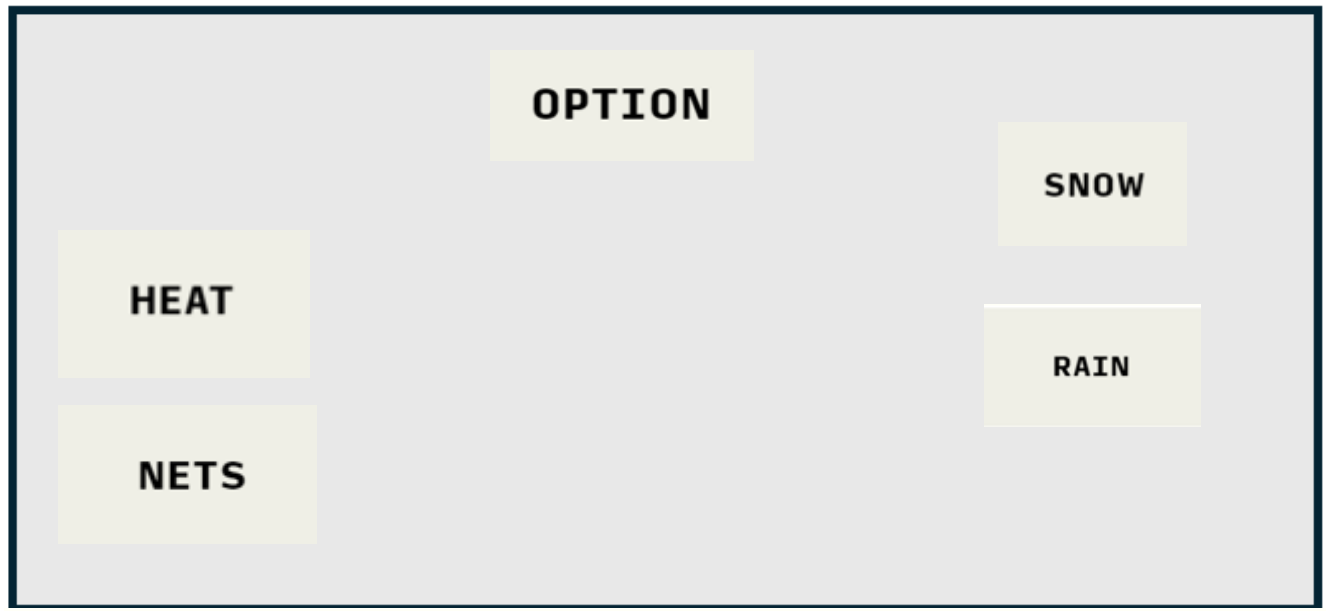
2) Do people find linguistic reasoning more difficult than other kinds of reasoning?

Measuring Similarity with GRIS:



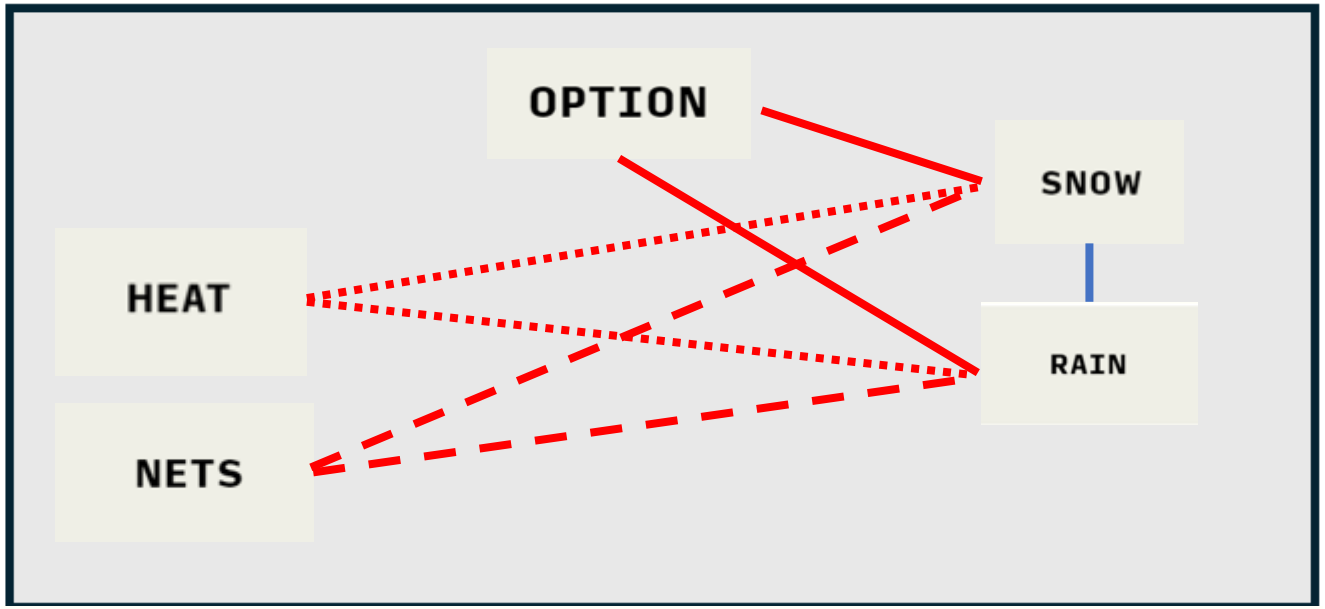
- People complete the task that you were asked to complete, with very similar instructions.
- We track incremental and final word positions.

KAYAK	SNOW	BUCKS	HAIL
OPTION	TAB	MOM	NETS
LEVEL	RAIN	HEAT	RETURN
JAZZ	SHIFT	RACE CAR	SLEET



KAYAK		BUCKS	HAIL
	TAB	MOM	
LEVEL			RETURN
JAZZ	SHIFT	RACE CAR	SLEET

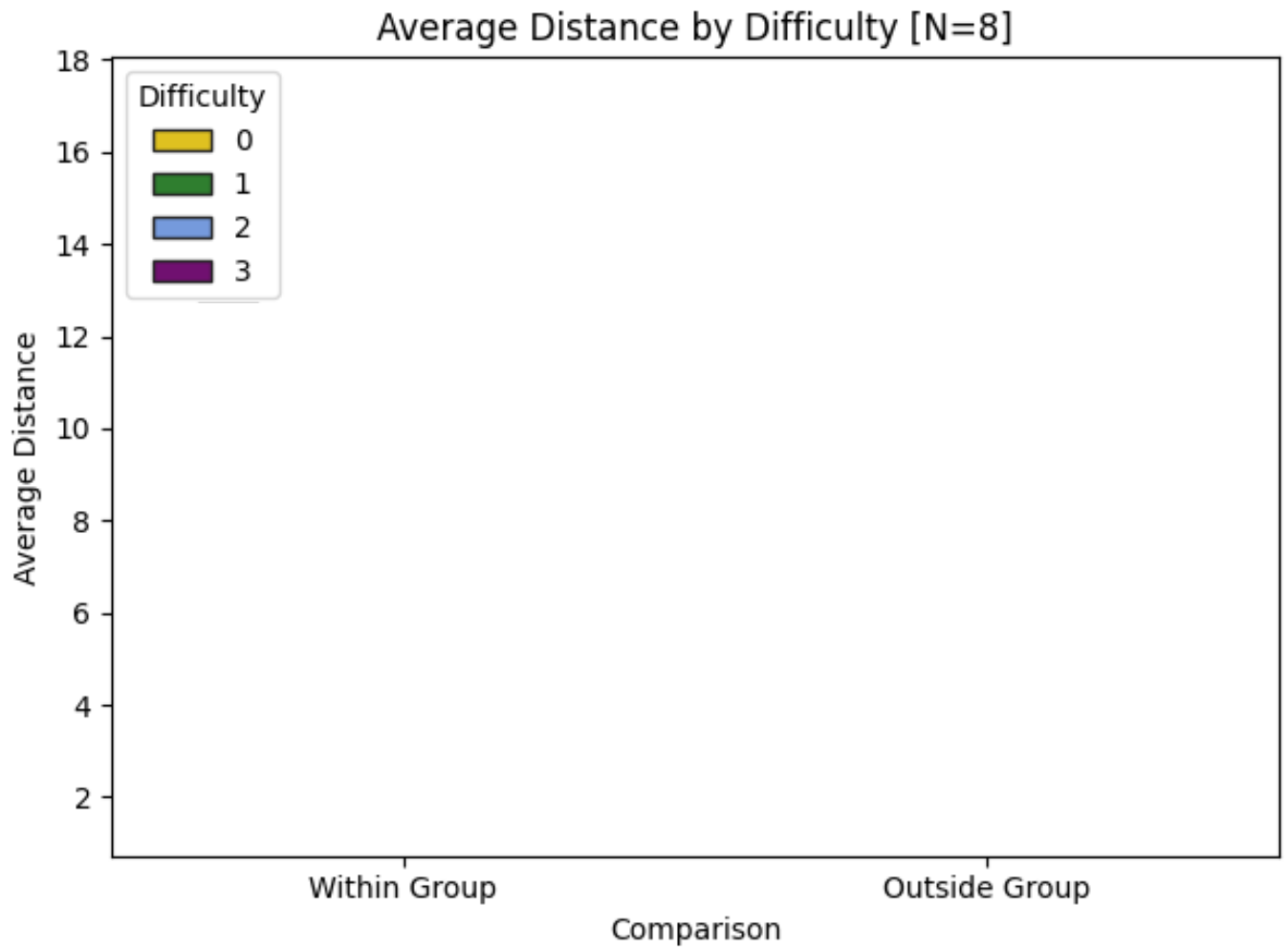
Average outside group distance



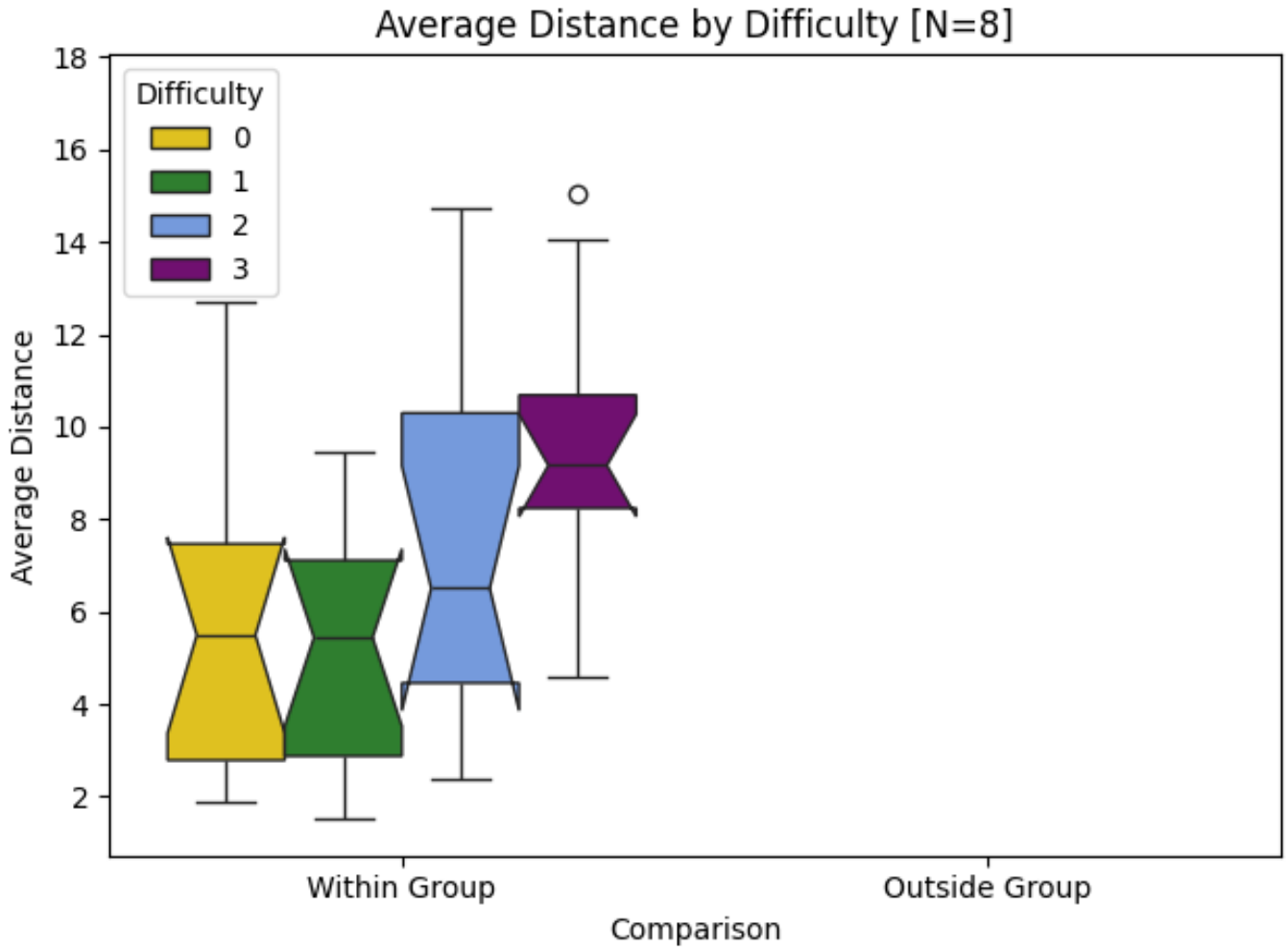
Average within group distance!

KAYAK		BUCKS	HAIL
	TAB	MOM	
LEVEL			RETURN
JAZZ	SHIFT	RACE CAR	SLEET

Higher difficulties *are* harder to group:

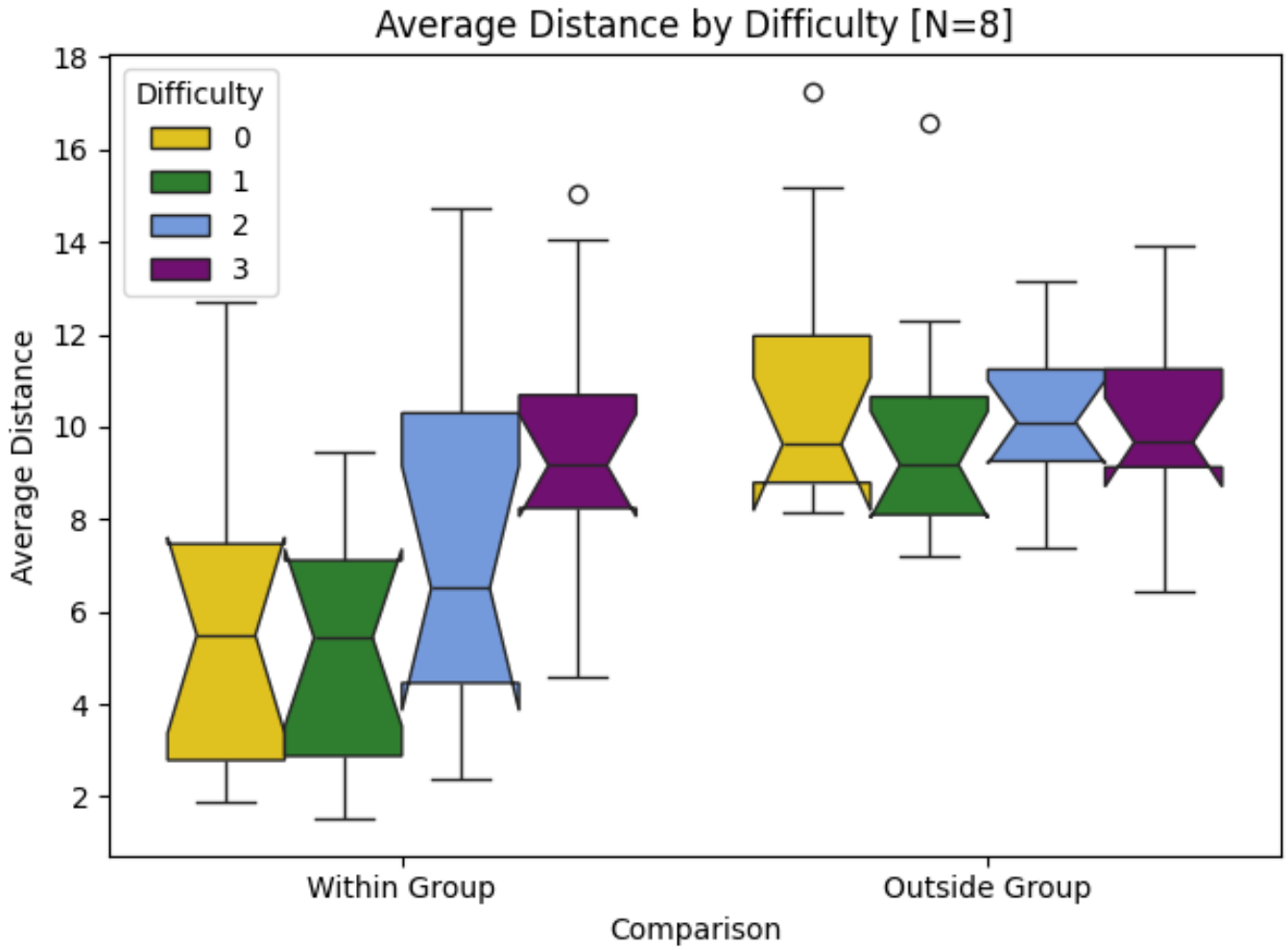


Higher difficulties *are* harder to group:



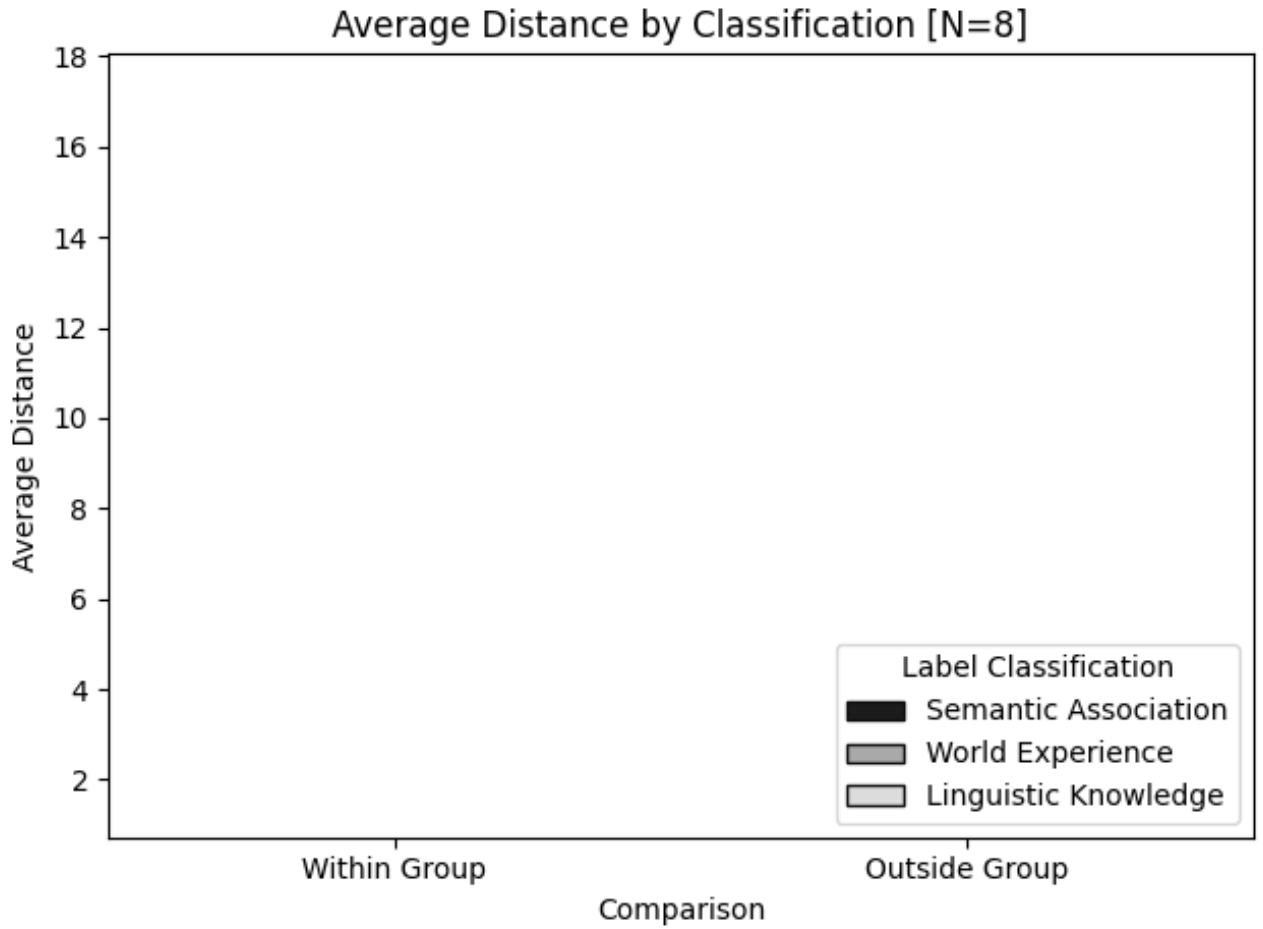
- Words within a category are often close together…
  - … though difficulty modulates this distance gradually.

Higher difficulties *are* harder to group:

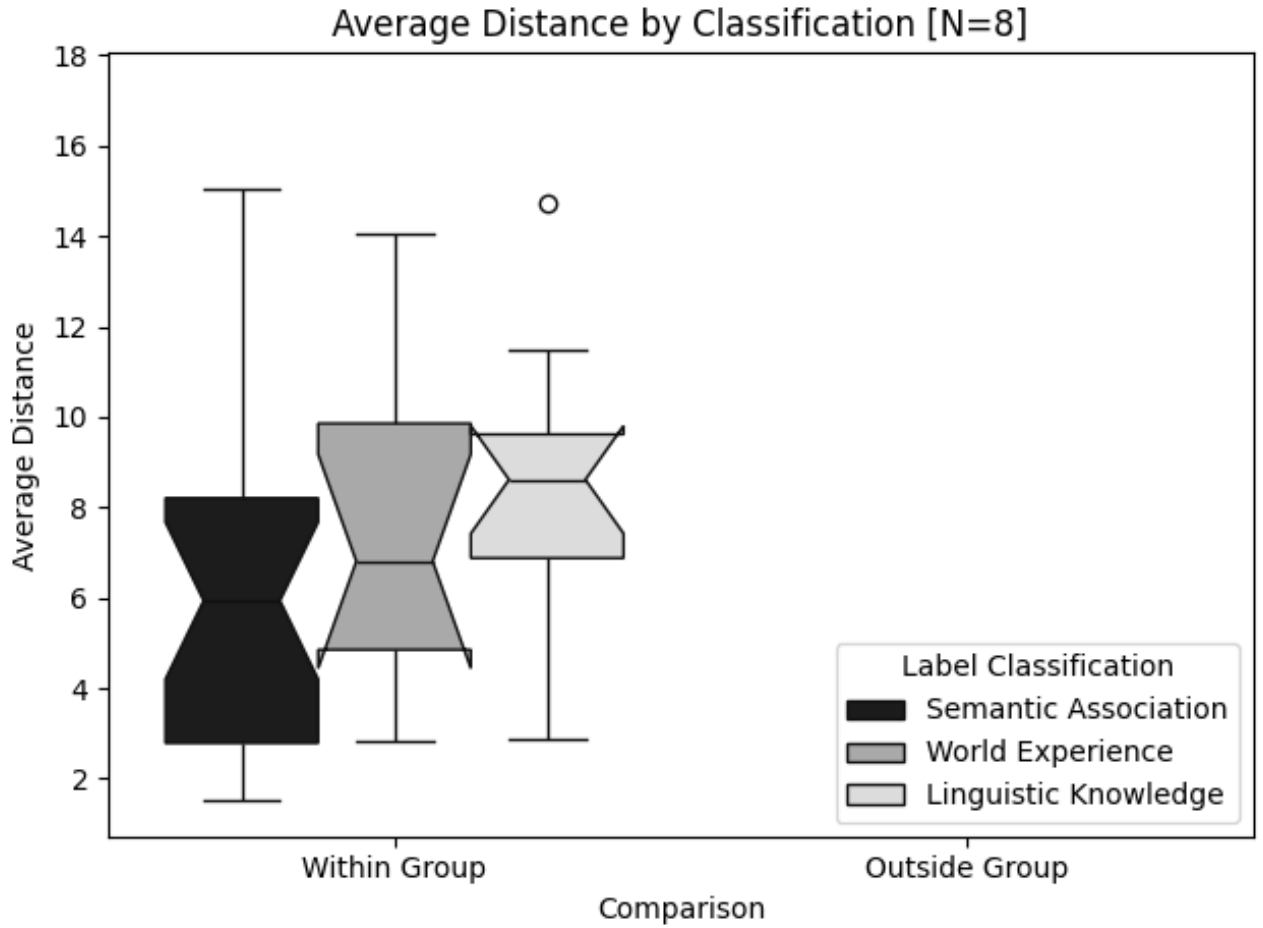


- Words within a category are often close together…
  - … though difficulty modulates this distance gradually.
- Words outside a category are often far…
  - … though note the purple within-group!

Some kinds of similarities *are* easier:

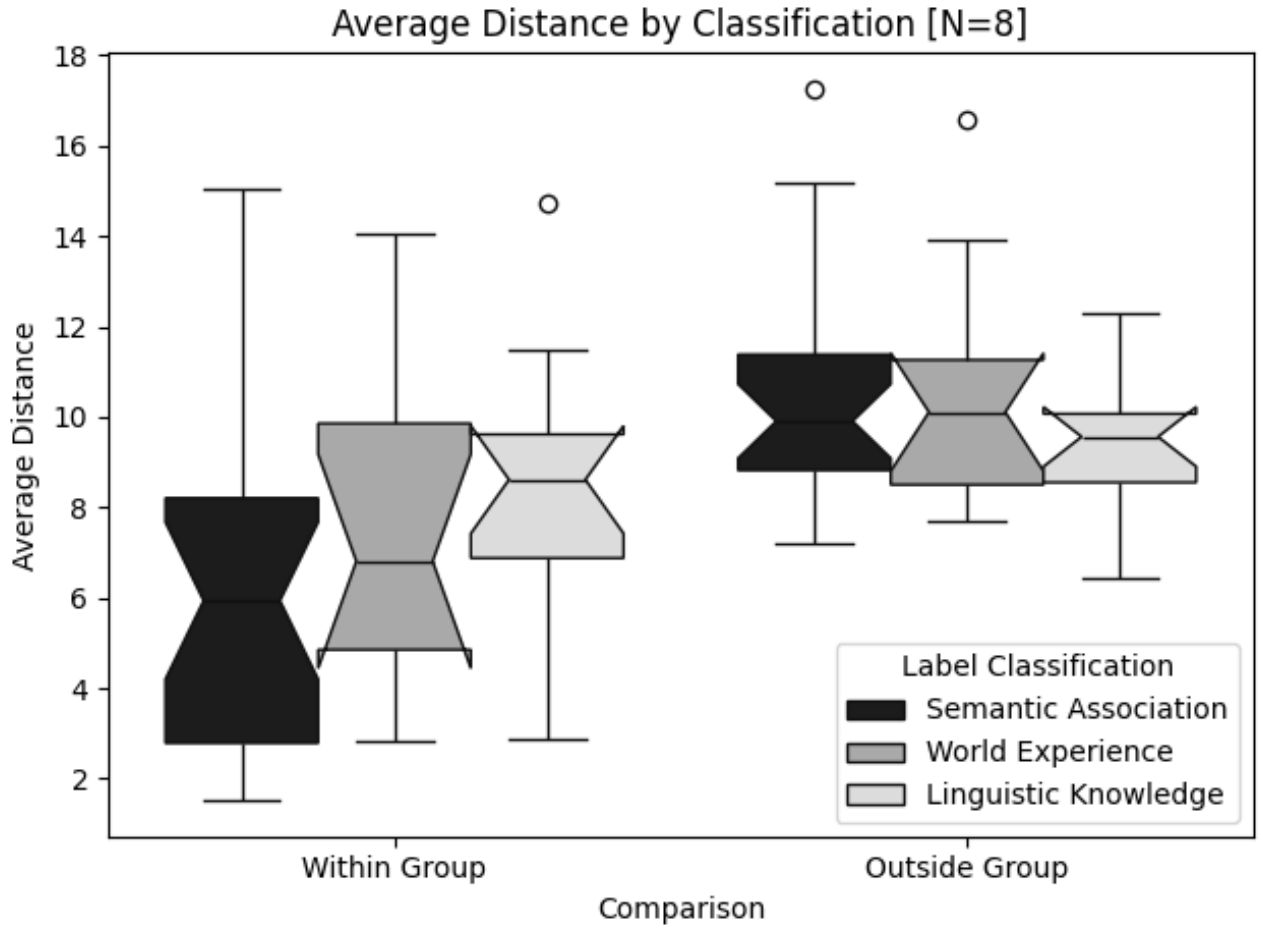


Some kinds of similarities *are* easier:



- Simple semantic association is easy.
- Other kinds of associations are difficult.

Some kinds of similarities *are* easier:



- Simple semantic association is easy.
- Other kinds of associations are difficult.

Takeaway:

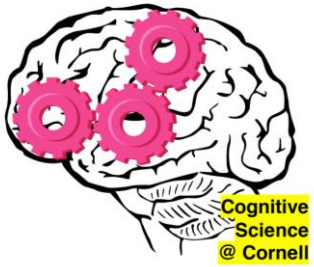
Within their representational spaces,
humans display similarity asymmetries.

(and… we hope that you can try out GRIS soon!)

Thanks!

C.Psyd

Cornell NLP



Will

Marty

Magnolia

Ashlyn

Jacob

KAYAK	SNOW	BUCKS	HAIL
OPTION	TAB	MOM	NETS
LEVEL	RAIN	HEAT	RETURN
JAZZ	SHIFT	RACE CAR	SLEET

WET WEATHER (hail, rain, sleet, snow)

NBA TEAMS (Bucks, Heat, Jazz, Nets)

KEYBOARD KEYS (option, return, shift, tab)

PALINDROMES (kayak, level, mom, race car)

KAYAK	SNOW	BUCKS	HAIL
OPTION	TAB	MOM	NETS
LEVEL	RAIN	HEAT	RETURN
JAZZ	SHIFT	RACE CAR	SLEET

WET WEATHER (hail, rain, sleet, snow)

NBA TEAMS (Bucks, Heat, Jazz, Nets)

KEYBOARD KEYS (option, return, shift, tab)

PALINDROMES (kayak, level, mom, race car)