

VALUE: Understanding Dialect Disparity in NLU

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, Diyi Yang



ACL 2022



Vern**A**cular **L**anguage **U**nderstanding **E**valuation:
Understanding Dialect Disparity in NLU

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, Diyi Yang



ACL 2022

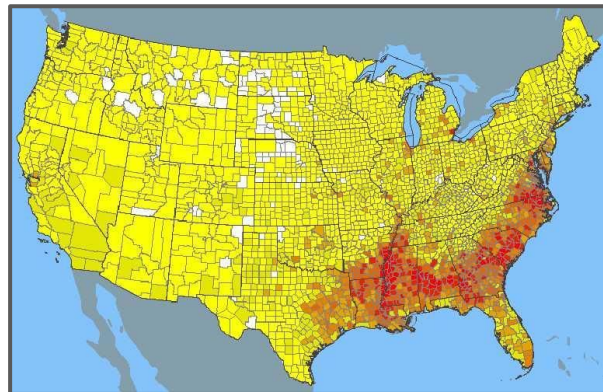


VALUE: Understanding Dialect Disparity in NLU

Motivation: *Dialect Disparity*

AAVE

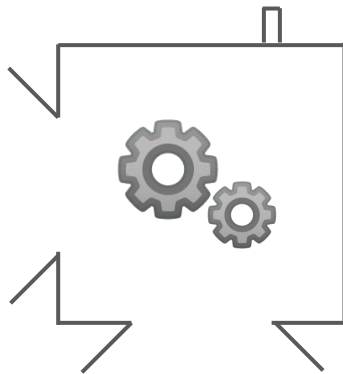
- Dependency Parsing ([Blodgett et al., 2018](#))
- Language ID ([Jurgens et al., 2017](#))
- POS Tagging ([Jørgensen et al., 2016](#))



What about more general NLU?



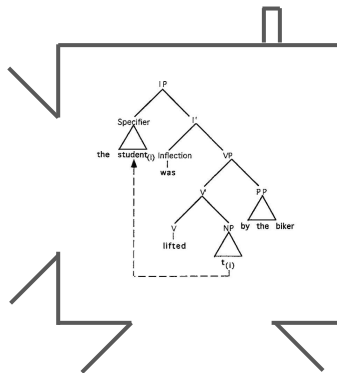
VALUE: Understanding Dialect Disparity in NLU



Dialect *Stress Test*



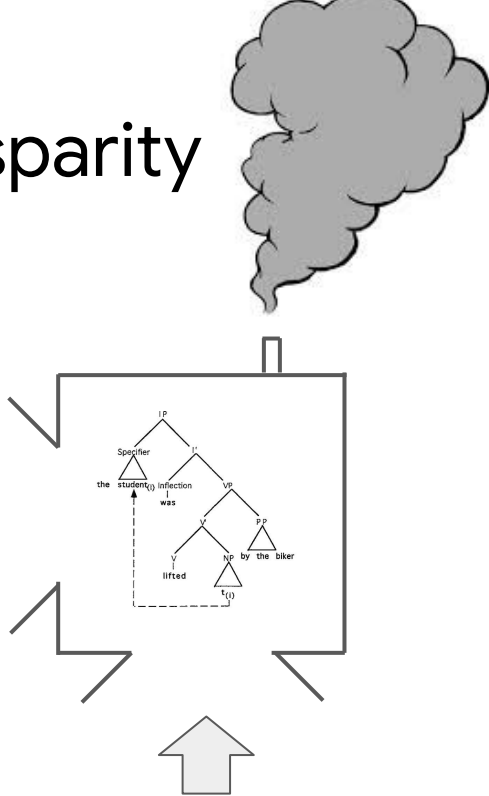
VALUE: Understanding Dialect Disparity in NLU



AAVE Stress Test



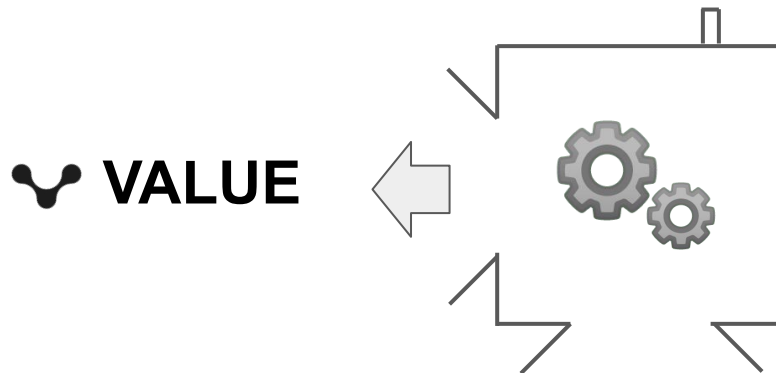
VALUE: Understanding Dialect Disparity



AAVE Stress Test



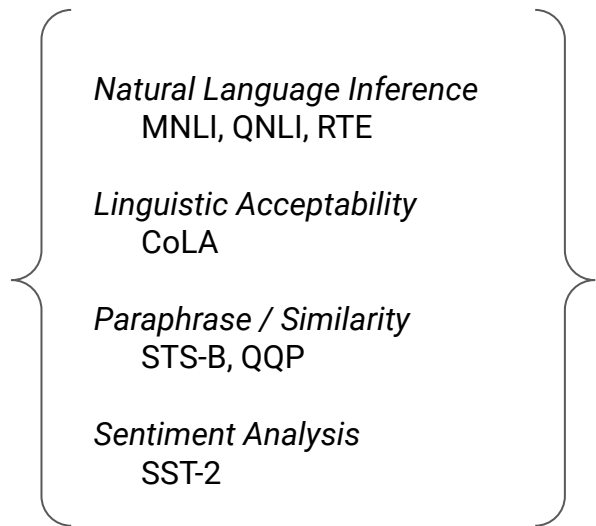
VALUE: Understanding Dialect Disparity in NLU



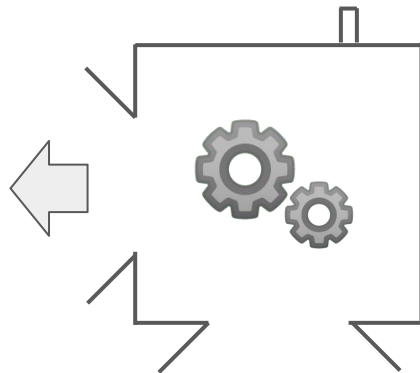
AAVE Stress Test

 **GLUE**

VALUE: Understanding Dialect Disparity in NLU



 **VALUE**



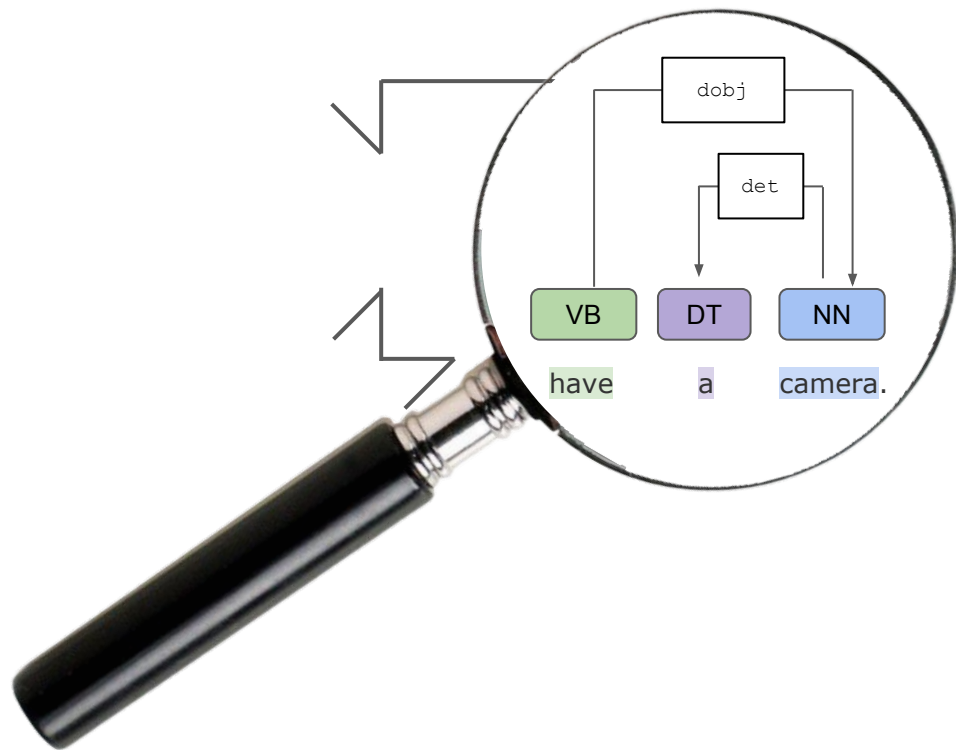
AAVE Stress Test

 **GLUE**

VALUE: Understanding Dialect Disparity in NLU

Advantages:

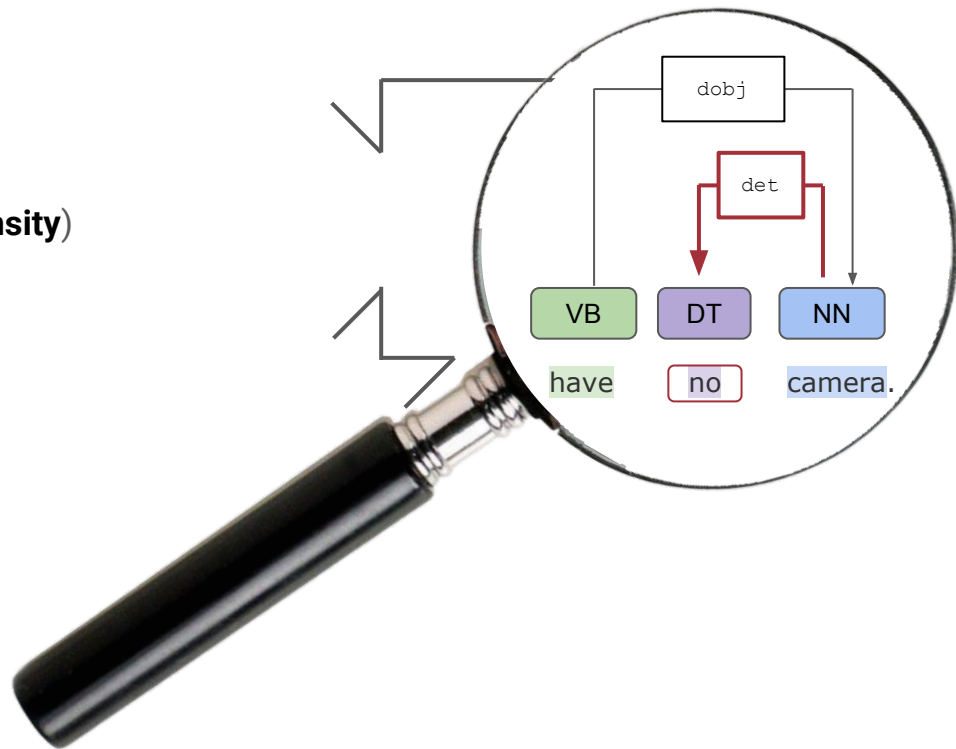
1. **Interpretable** (not **black-box**)



VALUE: Understanding Dialect Disparity in NLU

Advantages:

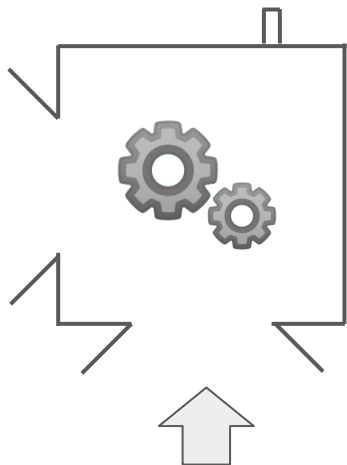
1. **Interpretable** (not **black-box**)
2. **Flexible** (tunable **feature-density**)



VALUE: Understanding Dialect Disparity in NLU

Advantages:

1. **Interpretable** (not **black-box**)
2. **Flexible** (tunable **feature-density**)
3. **Scalable** (**mix + match** datasets)



 **GLUE** SQuAD

 **SuperGLUE**

 **CoQA**
A Conversational Question Answering Challenge

 **WinoGrande**

VALUE: Understanding Dialect Disparity in NLU

Advantages:

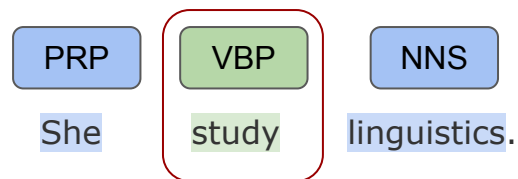
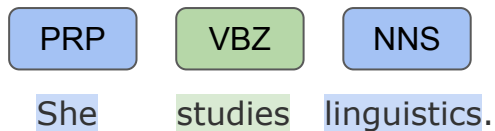
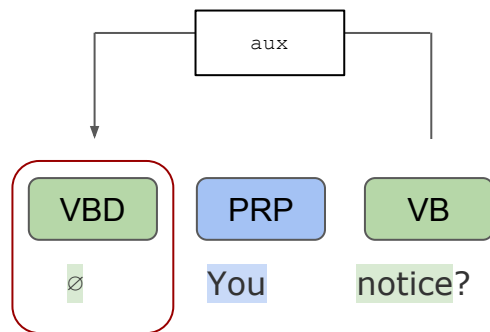
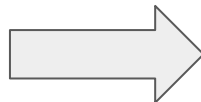
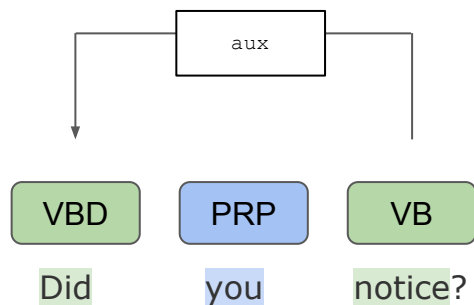
1. **Interpretable** (not **black-box**)
2. **Flexible** (tunable **feature-density**)
3. **Scalable** (**mix + match** datasets)
4. **Responsible** (**participatory design**)



Project Outline

1. **Transform:** Construct VALUE
2. **Validate:** Participatory Design and Gold-Standard
3. **Benchmark:** Test models on VALUE

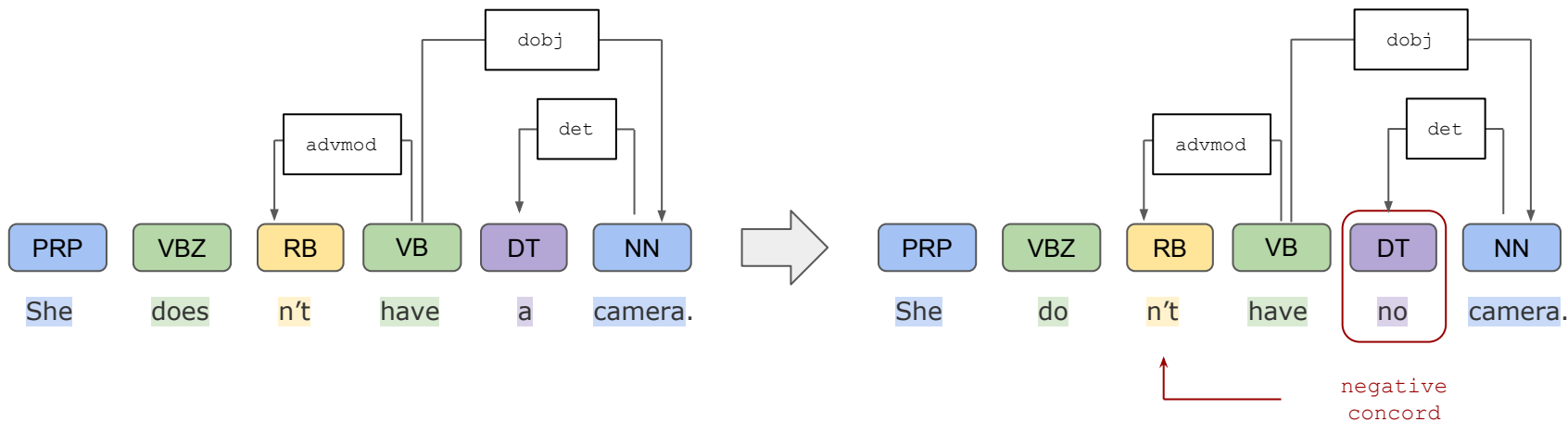
1. Transform: Morphosyntax



person = 0

1. Transform: Morphosyntax

Negative concord: AAVE speakers can use two negative morphemes to communicate a single negation.



1. Transform: Morphosyntax

Morphosyntactic Transformations:

auxiliary dropping · completive *done* / remote time *been* · existential *it* ·
future *gonna* · immediate future *finna* · *have/got* · inflection ·
negative concord · negative inversion · null complementizers · null genitives

1. Transform: Lexicon

Lexical Mapping: (one-to-many) [1]

1. Train `word2vec` on: **TwitterAAE** dataset
(Blodgett et al. 2016)

2. Linguistic code axis: $\mathbf{c} = \sum_{(\mathbf{x}_i, \mathbf{y}_i) \in S} \frac{\mathbf{x}_i - \mathbf{y}_i}{|S|}$

3. Rank candidate word pairs by:

$$\cos(\mathbf{c}, \mathbf{w}_i - \mathbf{w}_j)$$

4. Hand-filter any semantically unequal words

SAE	AAVE
arguing	<i>beefing, beefin, arguin</i>
anymore	<i>nomore, nomo</i>
brother	<i>homeboy</i>
classy	<i>fly</i>
dude	<i>n*ggah, manee, n*gga</i>
huge	<i>bigass</i>
probably	<i>prob, prolly, def, probly, deff</i>
rad	<i>dope</i>
remember	<i>rememba</i>
screaming	<i>screamin, yellin, hollering</i>
sister	<i>sista, sis</i>
these	<i>dese, dem</i>
with	<i>wit</i>

[1] Shoemark, P., Kirby, J., & Goldwater, S. (2018, November). [Inducing a lexicon of sociolinguistic variables from code-mixed text](#). In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text* (pp. 1-6).

2. Validate: Participatory Design

User-Centered Validation Protocol



Data for all.

2. Validate: Participatory Design

User-Centered Validation Protocol

DATA
WORKS

Georgia
Tech  College of
Computing

Sentence (1): can't nothing good happen

Sentence (2): nothing good can happen

○

Data for all.

2. Validate: Participatory Design

User-Centered Validation Protocol



Sentence (1): can't nothing good happen

Sentence (2): nothing good can happen

Understanding:

We have highlighted certain portions of **Sentence (2)** that are different in **Sentence (1)**. Do the words *and* the order of the words in **Sentence (1)** look like something you could reasonably say in AAVE?

Yes

No

2. Validate: Participatory Design

User-Centered Validation Protocol



Sentence (1): can't nothing good happen

Sentence (2): nothing good can happen

Understanding:

We have highlighted certain portions of **Sentence (2)** that are different in **Sentence (1)**. Do the words *and* the order of the words in **Sentence (1)** look like something you could reasonably say in AAVE?

Yes

No

If anything is confusing or strange, please let us know which of the highlighted segments were changed in a way that doesn't make sense _____

2. Validate: Participatory Design

User-Centered Validation Protocol



Sentence (1): can't nothing good happen

Sentence (2): nothing good can happen

2



Understanding:

We have highlighted certain portions of **Sentence (2)** that are different in **Sentence (1)**. Do the words *and* the order of the words in **Sentence (1)** look like something you could reasonably say in AAVE?

Yes

No

If anything is confusing or strange, please let us know which of the highlighted segments were changed in a way that doesn't make sense _____

2. Validate: Participatory Design

User-Centered Validation Protocol



Sentence (1): can't nothing good happen

Sentence (2): nothing good can happen

Rephrasing: (Gold Standard)

If possible, please provide a revised or alternative rephrasing of **Sentence (1)** that would be acceptable in the AAVE dialect. If no change is possible, leave this blank.

2. Validate: Participatory Design

User-Centered Validation Protocol



Sentence (1): can't nothing good happen

Sentence (2): nothing good can happen

Social Context:

If someone said this in your community, would it be (1) not very cool, (5) a bit sensitive, (7) passing, or (10) cool?



🙄 1 ○ 2 ○ 3 ○ 4 ○ 😐 5 ○ 6 ○ 7 ○ 8 ○ 9 ● 10 ○

2. Validate: Participatory Design



User-Centered Validation Protocol

Transformation	Accuracy (Maj. Vote)	Accuracy (Unanimous)	Size <i>n</i>
Ass constructions	-	-	-
Auxiliaries	96.6	77.4	638
Been / done	95.4	72.7	670
Existential dey/it	91.4	57.9	304
Gonna / finna	95.4	78.7	197
Have / got	96.2	84.8	290
Inflection	97.1	82.3	761
Negative concord	95.9	73.6	584
Negative inversion	95.0	69.3	101
Null genitives	97.9	85.3	573
Relative clause structures	94.1	58.3	489



3. Benchmark: Test NLU

 Train	Test 	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
GLUE: CoLA				
GLUE: MNLI				
GLUE: QNLI				
GLUE: RTE				
GLUE: SST-2				
GLUE: STS-B				
GLUE: QQP				



3. Benchmark: Test NLU

 Train	Test 	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
GLUE: CoLA		56.3		
GLUE: MNL1		83.6		
GLUE: QNLI		92.8		
GLUE: RTE		66.4		
GLUE: SST-2		94.6		
GLUE: STS-B		89.4		
GLUE: QQP		90.9		



3. Benchmark: Test NLU

 Train	Test 	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
GLUE: CoLA		56.3	55.6	
GLUE: MNL		83.6	82.5	
GLUE: QNLI		92.8	91.4	
GLUE: RTE		66.4	67.8	
GLUE: SST-2		94.6	92.4	
GLUE: STS-B		89.4	88.5	
GLUE: QQP		90.9	89.5	

3. Benchmark: Test NLU

 Train	Test 	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
GLUE: CoLA		56.3	55.6	-
GLUE: MNL1		83.6	82.5	82.1
GLUE: QNLI		92.8	91.4	91.2
GLUE: RTE		66.4	67.8	67.6
GLUE: SST-2		94.6	92.4	92.0
GLUE: STS-B		89.4	88.5	88.2
GLUE: QQP		90.9	89.5	89.2

3. Benchmark: Test NLU

 Train	Test 	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
GLUE: CoLA		56.3	55.6	-
GLUE: MNL1		83.6	82.5	82.1
GLUE: QNLI		92.8	91.4	91.2
GLUE: RTE		66.4	67.8	67.6
GLUE: SST-2		94.6	92.4	92.0
GLUE: STS-B		89.4	88.5	88.2
GLUE: QQP		90.9	89.5	89.2



3. Benchmark: Test NLU

↓ Train	Test →	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
AAVE: CoLA (Synth)		56.3	55.6	-
AAVE: MNL I (Synth)		83.6	82.5	82.1
AAVE: QNLI (Synth)		92.8	91.4	91.2
AAVE: RTE (Synth)		66.4	67.8	67.6
AAVE: SST-2 (Synth)		94.6	92.4	92.0
AAVE: STS-B (Synth)		89.4	88.5	88.2
AAVE: QQP (Synth)		90.9	89.5	89.2



3. Benchmark: Test NLU

↓ Train	Test →	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
AAVE: CoLA (Synth)		56.3 56.2	55.6 55.8	-
AAVE: MNL I (Synth)		83.6	82.5	82.1
AAVE: QNLI (Synth)		92.8	91.4	91.2
AAVE: RTE (Synth)		66.4	67.8	67.6
AAVE: SST-2 (Synth)		94.6	92.4	92.0
AAVE: STS-B (Synth)		89.4	88.5	88.2
AAVE: QQP (Synth)		90.9	89.5	89.2



3. Benchmark: Test NLU

 Train	Test 	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
AAVE: CoLA (Synth)		56.3 56.2	55.6 55.8	-
AAVE: MNL1 (Synth)		83.6 83.1	82.5 83.5	82.1 82.3
AAVE: QNLI (Synth)		92.8	91.4	91.2
AAVE: RTE (Synth)		66.4	67.8	67.6
AAVE: SST-2 (Synth)		94.6	92.4	92.0
AAVE: STS-B (Synth)		89.4	88.5	88.2
AAVE: QQP (Synth)		90.9	89.5	89.2



3. Benchmark: Test NLU

 Train	Test 	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
AAVE: CoLA (Synth)		56.3 56.2	55.6 55.8	-
AAVE: MNL1 (Synth)		83.6 83.1	82.5 83.5	82.1 82.3
AAVE: QNLI (Synth)		92.8 92.5	91.4 91.8	91.2 91.8
AAVE: RTE (Synth)		66.4	67.8	67.6
AAVE: SST-2 (Synth)		94.6	92.4	92.0
AAVE: STS-B (Synth)		89.4	88.5	88.2
AAVE: QQP (Synth)		90.9	89.5	89.2

3. Benchmark: Test NLU

 Train	Test 	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
AAVE: CoLA (Synth)		56.3 56.2	55.6 55.8	-
AAVE: MNL1 (Synth)		83.6 83.1	82.5 83.5	82.1 82.3
AAVE: QNLI (Synth)		92.8 92.5	91.4 91.8	91.2 91.8
AAVE: RTE (Synth)		66.4 67.1	67.8 66.1	67.6 67.3
AAVE: SST-2 (Synth)		94.6	92.4	92.0
AAVE: STS-B (Synth)		89.4	88.5	88.2
AAVE: QQP (Synth)		90.9	89.5	89.2

3. Benchmark: Test NLU

 Train	Test 	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
AAVE: CoLA (Synth)		56.3 56.2	55.6 55.8	-
AAVE: MNL1 (Synth)		83.6 83.1	82.5 83.5	82.1 82.3
AAVE: QNLI (Synth)		92.8 92.5	91.4 91.8	91.2 91.8
AAVE: RTE (Synth)		66.4 67.1	67.8 66.1	67.6 67.3
AAVE: SST-2 (Synth)		94.6 94.0	92.4 93.0	92.0 92.8
AAVE: STS-B (Synth)		89.4	88.5	88.2
AAVE: QQP (Synth)		90.9	89.5	89.2

3. Benchmark: Test NLU

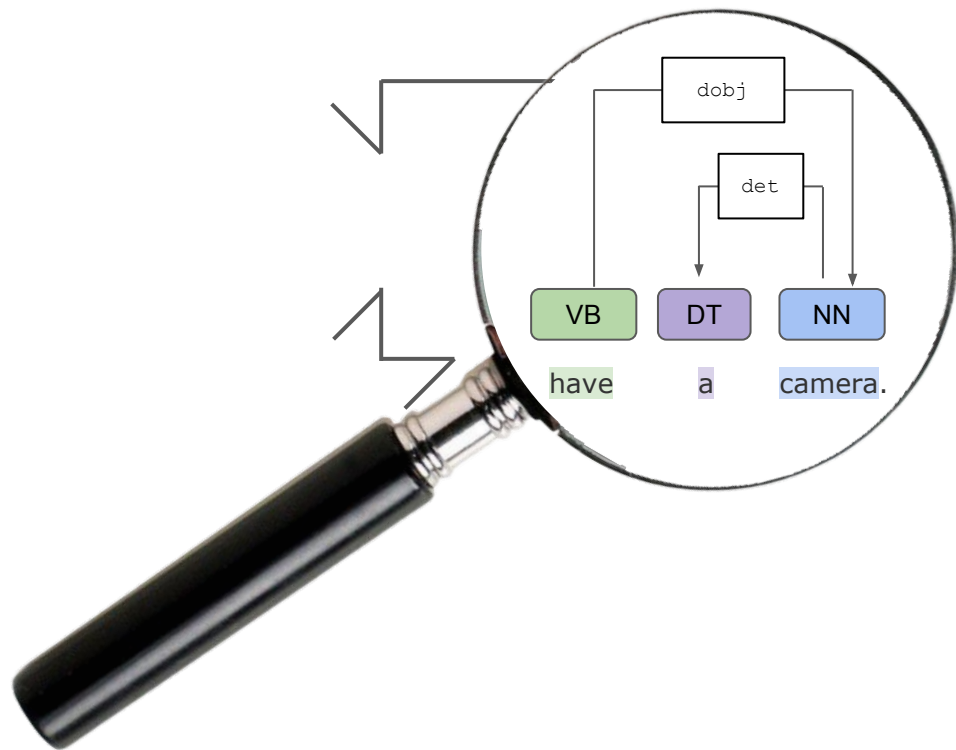
↓ Train	Test →	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
AAVE: CoLA (Synth)		56.3 56.2	55.6 55.8	-
AAVE: MNL1 (Synth)		83.6 83.1	82.5 83.5	82.1 82.3
AAVE: QNLI (Synth)		92.8 92.5	91.4 91.8	91.2 91.8
AAVE: RTE (Synth)		66.4 67.1	67.8 66.1	67.6 67.3
AAVE: SST-2(Synth)		94.6 94.0	92.4 93.0	92.0 92.8
AAVE: STS-B(Synth)		89.4 88.8	88.5 88.3	88.2 88.3
AAVE: QQP (Synth)		90.9	89.5	89.2

3. Benchmark: Test NLU

↓ Train	Test →	SAE (GLUE)	AAVE (Synthetic)	AAVE (Gold)
AAVE: CoLA (Synth)		56.3 56.2	55.6 55.8	-
AAVE: MNL1 (Synth)		83.6 83.1	82.5 83.5	82.1 82.3
AAVE: QNLI (Synth)		92.8 92.5	91.4 91.8	91.2 91.8
AAVE: RTE (Synth)		66.4 67.1	67.8 66.1	67.6 67.3
AAVE: SST-2(Synth)		94.6 94.0	92.4 93.0	92.0 92.8
AAVE: STS-B(Synth)		89.4 88.8	88.5 88.3	88.2 88.3
AAVE: QQP (Synth)		90.9 90.3	89.5 89.6	89.2 89.6

3. Benchmark: Test NLU

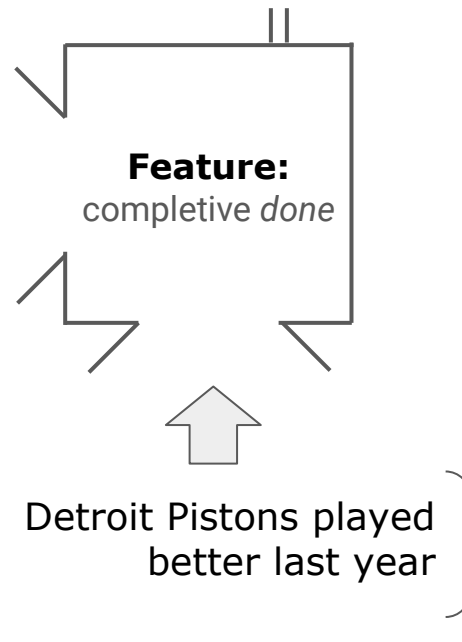
Perturbation Analysis (MNLI)



3. Benchmark: Test NLU

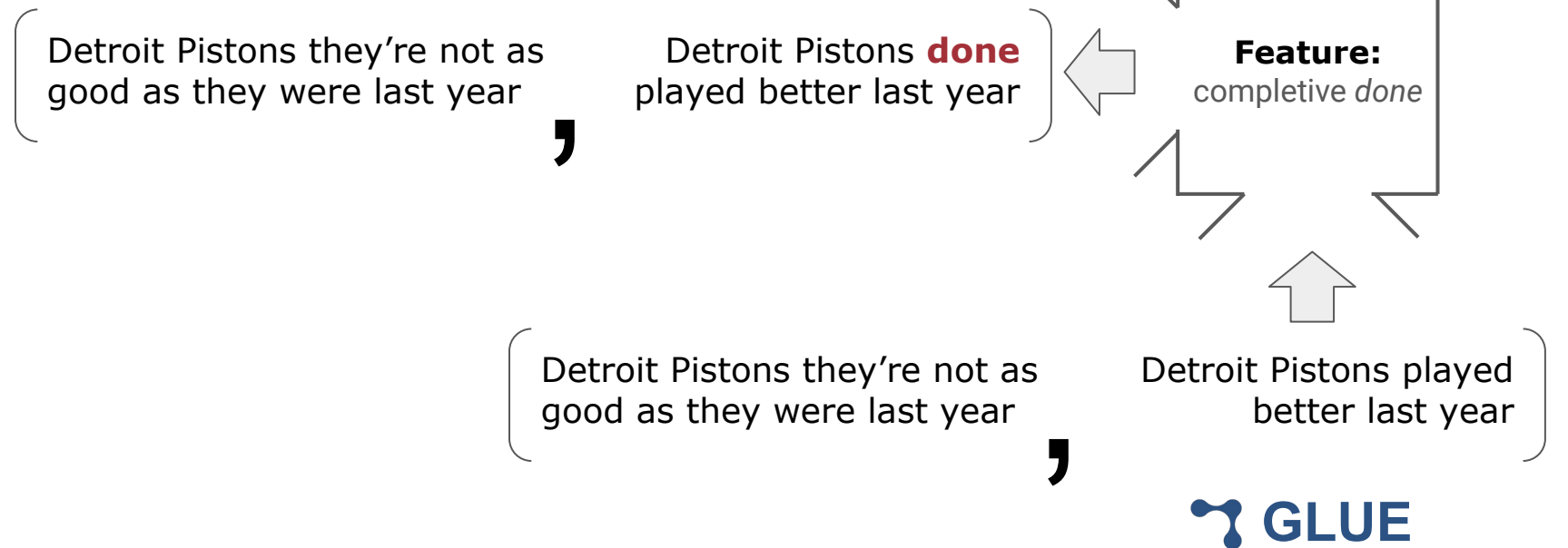
Perturbation Analysis (MNLI)

{ Detroit Pistons they're not as good as they were last year ,



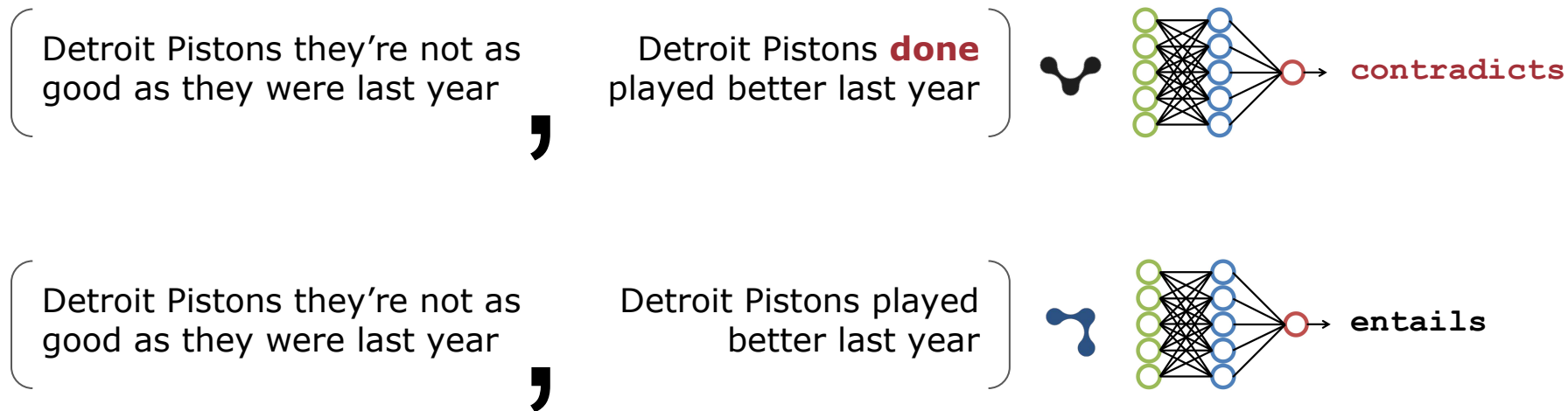
3. Benchmark: Test NLU

Perturbation Analysis (MNLI)



3. Benchmark: Test NLU

Perturbation Analysis (MNLI)



4. Conclusion: Limitations

1.  **VALUE** should not be considered *natural* AAVE

→ Exaggerated feature density [**stress test**]


→ Lacks social nuance

4. Conclusion: Limitations




1.  **VALUE** should not be considered *natural* AAVE

→ Exaggerated feature density [**stress test**]




→ Lacks social nuance

2.  Speech \neq orthography 

4. Conclusion: Limitations

1.  **VALUE** should not be considered *natural* AAVE
 - Exaggerated feature density [**stress test**]
 - Lacks social nuance
2.  Speech \neq orthography 
3. Synthetic test performance \nrightarrow real-world readiness

4. Conclusion: Limitations

1.  **VALUE** should not be considered *natural* AAVE
 - Exaggerated feature density [stress test]
 - Lacks social nuance
2.  Speech \neq orthography 
3. Synthetic test performance \nrightarrow real-world readiness
4. Misuse: hateful speech and appropriation

4. Conclusion: Contributions

1. **Transform:** Construct VALUE { **Flexible, Scalable** }
2. **Validate:** Participatory Design { **Responsible** }
3. **Benchmark:** Test models on VALUE { **Interpretable** }

4. Conclusion: Future Work

1. **Extend Scope:** Consider other tasks
2. **Extend Impact:** Reach other dialects

4. Conclusion: Future Work

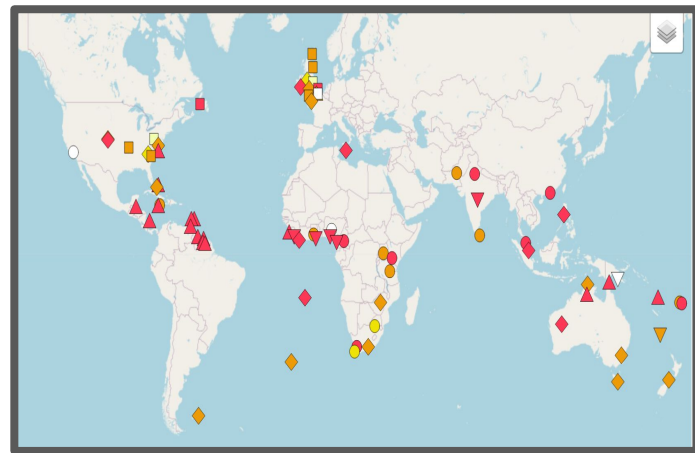
1. **Extend Scope:** Consider other tasks
2. **Extend Impact:** Reach other dialects

You like NLP?

You get the point?

feature: *No inversion / no auxiliaries in main clause yes/no questions*

pervasive in: *Colloquial AE, IrE, IE, SgE*



source:

<https://ewave-atlas.org/parameters/229#2/7.0/7.9>

4. Conclusion: Future Work

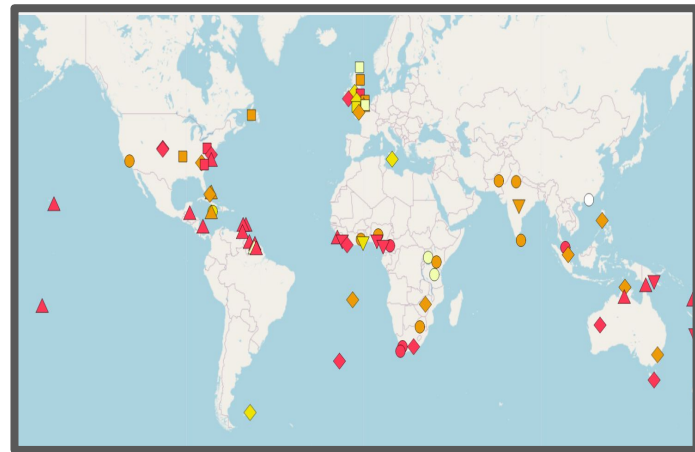
1. **Extend Scope:** Consider other tasks
2. **Extend Impact:** Reach other dialects

y'all

you'uns

feature: *Variants of the second-person pronoun*

pervasive in: *Colloquial AE, AppE, AusE*



source:

<https://ewave-atlas.org/parameters/229#2/7.0/7.9>

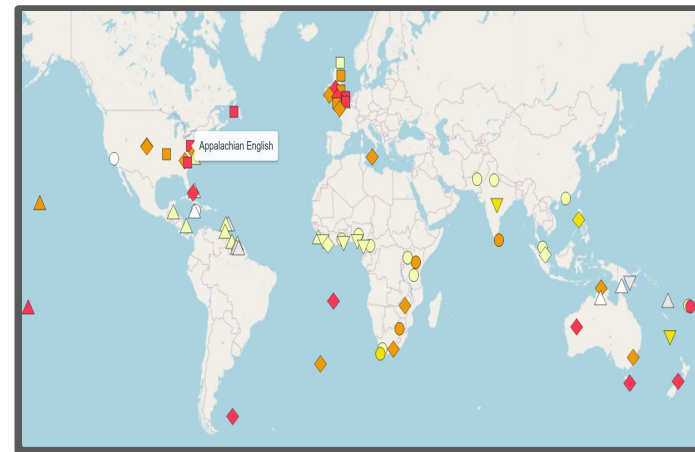
4. Conclusion: Future Work

1. **Extend Scope:** Consider other tasks
2. **Extend Impact:** Reach other dialects

Us kids used to pinch the sweets like hell.

feature: *us + NP in subject function*

pervasive in: *AppE*



source:

<https://ewave-atlas.org/parameters/229#2/7.0/7.9>

4. Conclusion: Future Work

1. **Extend Scope:** Consider other tasks
2. **Extend Impact:** Reach other dialects
3. **Build:** Dialect-Aware NLP systems

VALUE: Understanding Dialect Disparity in NLU

Caleb Ziems, Jiaao Chen, Camille Harris, Jessica Anderson, Diyi Yang



GT-SALT/[value](#)