Morphological Generalization by Children & Computers

Sarah Brogden Payne

sarah.payne@stonybrook.edu | paynesa.github.io

MIT Department of Brain & Cognitive Sciences

April 19, 2024

Morphological Inflection

- Patterns of **word formation** to express **grammatical categories**
  - **English:** ***walk***+Past → ***walked***
  - **Mandarin:** **3**+Pl → ***tāmen*** ‘they’
  - **Hebrew:** √***ħtl***+Dim+Sg+Def → ***haħataltul*** ‘the kitty’
  - **Latin:** ***amic***+Fem+Sg+Gen → ***amīcae*** ‘the friend’s’
  - **Shona:** ***bik***+1sg.Subj+6cl.Obj+Past+Caus+Pass→ ***ndakachibikiswa*** ‘I was made to cook it’

Morphological Inflection

- Patterns of **word formation** to express **grammatical categories**
  - **Roots/stems** modified by many processes
    - **Suffixation/prefixation/circumfixation**, **stem mutations**, **reduplication**
  - Express **number, tense, mood, voice, aspect, evidentiality,…**
  - Common across the world's languages
    - Vary dramatically in terms of **complexity** or **"richness"**
  - Poses a learning challenge for both **machines** and **humans**

Morphological Inflection: Applications

Cognitive Modeling

- Insight into the **cognitive computations** underlying morphological learning
- **Past Tense Debate**
  - Early **connectionist** account (Rumelhart & McClelland 1986)
  - Several shortcomings
- Recent advances in **ANN architectures**
  - Renewed interest in the plausibility of ANNs as **cognitive models**

Natural Language Processing

- **Traditionally:** downstream tasks
  - In settings where **pipelining** is still common (e.g., **low-resource**)
  - Particularly for languages with **lots of inflectional morphology**
- May provide insight into the behavior of **ANN architectures**
  - A particular kind of **string-to-string mapping** problem
  - Varying performance may reflect **divergent properties** of different architectures

Morphological Inflection: Solved?

- **Kirov & Cotterell (2018):** encoder-decoder network can **overcome practical limitations** of older ANNs
  - Near **100% test accuracy**
  - Learn **several inflectional classes** at once
- **Corkerey et al. (2019):** K&C model still fails empirically
  - Predictions **don't match well** with human nonce word judgments
    - **Over-irregularizes** compared to humans!
  - Massive **variability** in model rankings between seeds
    - **Correlation with human ratings** also varies massively

Morphological Inflection: Solved?

Best systems on a subset of the 2018 **CoNLL-SIGMORPHON** shared task

	High	Medium	Low
Adyghe	100.00(uzh-2)	94.40(uzh-1)	90.60(ua-8)
Albanian	98.90(bme-2)	88.80(iitbhu-iiith-2)	36.40(uzh-1)
Arabic	93.70(uzh-1)	79.40(uzh-1)	45.20(uzh-1)
Armenian	96.90(bme-2)	92.80(uzh-1)	64.90(uzh-1)
Asturian	98.70(uzh-1)	92.40(iitbhu-iiith-2)	74.60(uzh-2)
Azeri	100.00(axsemantics-2)	96.00(iitbhu-iiith-2)	65.00(iitbhu-iiith-2)
Bashkir	99.90(uzh-2)	97.30(uzh-2)	77.80(iitbhu-iiith-1)
Basque	98.90(bme-2)	88.10(iitbhu-iiith-2)	13.30(uzh-1)
Belarusian	94.90(uzh-1)	70.40(uzh-1)	33.40(ua-8)
Bengali	99.00(bme-3)	99.00(uzh-2)	72.00(uzh-2)
Breton	100.00(waseda-1)	96.00(uzh-2)	72.00(uzh-1)
Bulgarian	98.30(uzh-2)	83.80(uzh-2)	62.90(ua-8)
Catalan	98.90(uzh-2)	92.80(waseda-1)	72.50(ua-8)
Classical-syriac	100.00(axsemantics-1)	100.00(axsemantics-2)	96.00(uzh-2)
Cornish	—	70.00(uzh-1)	40.00(ua-4)
Crimean-tatar	100.00(iit-varanasi-1)	98.00(uzh-2)	91.00(iitbhu-iiith-2)
Czech	94.70(uzh-1)	87.20(uzh-1)	46.50(uzh-2)
Danish	95.50(uzh-1)	80.40(uzh-1)	87.70(ua-6)
Dutch	97.90(uzh-1)	85.70(uzh-1)	69.30(ua-6)
English	97.10(uzh-2)	94.50(uzh-1)	91.80(ua-8)

Very good performance on medium and high training

Morphological Inflection: Solved?

Performance on **closely-related languages** is **highly variable**

Azeri	100.00(axsemantics-2)	96.00(iitbhu-iiith-2)	65.00(iitbhu-iiith-2)
Turkish	98.50(uzh-2)	90.70(uzh-1)	39.50(iitbhu-iiith-2)
Turkmen	—	98.00(iitbhu-iiith-1)	90.00(uzh-2)

Belarusian	94.90(uzh-1)	70.40(uzh-1)	33.40(ua-8)
Russian	94.40(uzh-2)	86.90(uzh-1)	53.50(uzh-1)
Ukrainian	96.20(uzh-2)	81.40(uzh-1)	57.10(ua-6)

Finnish	95.40(uzh-1)	82.80(uzh-1)	25.70(uzh-1)
Ingrian	—	92.00(uzh-2)	46.00(iitbhu-iiith-2)
Karelian	—	100.00(uzh-2)	94.00(ua-5)

Kashubian	—	88.00(bme-2)	68.00(ua-5)
Lower-sorbian	97.80(uzh-1)	85.10(uzh-1)	54.30(ua-6)
Polish	93.40(uzh-2)	82.40(uzh-2)	49.40(ua-6)

Danish	95.50(uzh-1)	80.40(uzh-1)	87.70(ua-6)
Norwegian-bokmaal	92.10(uzh-2)	84.10(uzh-1)	90.10(ua-6)
Swedish	93.30(uzh-1)	79.80(uzh-1)	79.00(ua-8)

Czech	94.70(uzh-1)	87.20(uzh-1)	46.50(uzh-2)
Slovak	97.10(uzh-1)	78.60(uzh-1)	51.80(uzh-2)

Galician	99.50(uzh-1)	90.80(uzh-1)	61.10(uzh-2)
Portuguese	98.60(uzh-2)	94.80(uzh-2)	75.80(uzh-2)

Irish	91.50(uzh-2)	77.10(uzh-1)	37.70(uzh-1)
Scottish-gaelic	—	94.00(iitbhu-iiith-1)	74.00(iitbhu-iiith-2)

Morphological Inflection isn't solved!

Morphological Inflection: Outstanding Issues

- ANNs are trained on **unrealistically large/saturated data**
- ANNs are rarely evaluated against **child learning trajectories** and **error patterns**

Belth, Payne et al. (2021, Cogsci)
Kodner, Payne et al. (2023, ACL)
Kodner, Khalifa, Payne, & Liu (2023, Cogsci)

- Current evaluation metrics fail to control for:
  - **Overlap** between train and test
  - Performance **variation** across multiple splits
  - **Frequency effects** in uniform sampling

Kodner, Payne et al. (2023, ACL)
Kodner, Khalifa & Payne (2023, EMNLP)

Outline

- **Background**
  - Defining the task
  - Input sparsity
  - Developmental trajectories & error patterns
- **Developmentally-grounded** evaluation
- Another approach: **Abduction of Tolerable Productivity**
- Revisiting the train-test overlap
- Probing feature-based generalization
- Conclusions

Outline

- **Background**
  - **Defining the task**
  - Input sparsity
  - Developmental trajectories & error patterns
- **Developmentally-grounded** evaluation
- Another approach: **Abduction of Tolerable Productivity**
- Revisiting the train-test overlap
- Probing feature-based generalization
- Conclusions

Morphological Inflection as an NLP Task

- **Training:** (`lemma`, `inflected form`, `feature set`)

swim	swam	V;PST
eat	eats	V;PRS;3;SG
cat	cats	N;PL
…	…	…

- **Testing:** (`lemma`, `feature set`) → `inflected form`

swim	?	V;PRS;3;SG
box	?	N;PL
cat	?	N;SG
…	…	…

Morphological Inflection as an NLP Task

- **Training:** (`lemma`, `inflected form`, `feature set`)

swim	swam	V;PST
eat	eats	V;PRS;3;SG
cat	cats	N;PL
…	…	…

- **Testing:** (`lemma`, `feature set`) → `inflected form`

swim	swims	V;PRS;3;SG
box	boxes	N;PL
cat	cat	N;SG
…	…	…

Outline

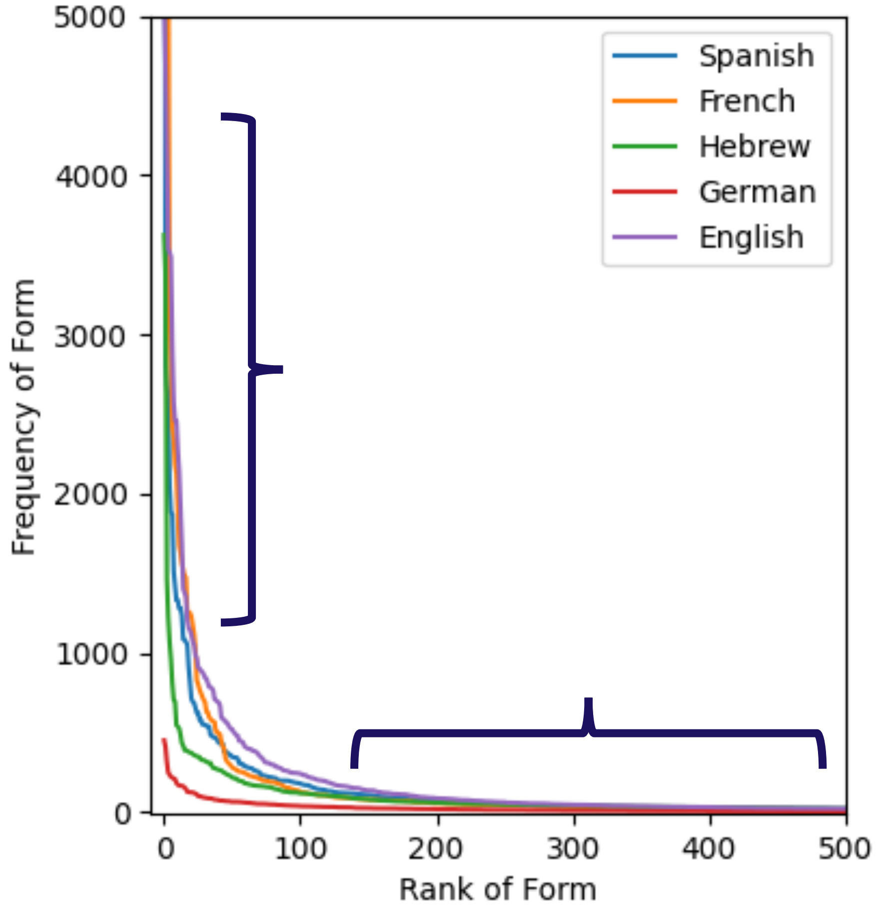
- **Background**
  - Defining the task
  - **Input sparsity**
  - Developmental trajectories & error patterns
- **Developmentally-grounded** evaluation
- Another approach: **Abduction of Tolerable Productivity**
- Revisiting the train-test overlap
- Probing feature-based generalization
- Conclusions

Input Sparsity: Zipf's Law

- **Zipf's law:** word ***rank*** inversely proportional to ***frequency***

$$f(r) \propto \frac{1}{r}$$

- **Consequences:**
  - A **few forms** occur very **frequently**
  - Most occur very **rarely** (long tail)



(data from Payne et al 2021, Belth et al 2021, Payne 2022, and Payne 2023)

Input Sparsity: Paradigm Saturation

- Long-tailed distributions in morphology: **Paradigm Saturation**
  - How many possible inflected forms does a lemma actually occur in?

$$saturation = \frac{\#\ seen}{\#\ possible}$$

	Present	Preterite	Imperfect	Conditional	Future
1SG	*amo*	*amé*	*amaba*	*amaría*	*amaré*
2SG	*amas*	*amaste*	*amabas*	*amarías*	*amarás*
3SG	*ama*	*amó*	*amaba*	*amaría*	*amará*
1PL	*amamos*	*amamos*	*amábamos*	*amaríamos*	*amaremos*
2PL	*amáis*	*amasteis*	*amabais*	*amaríais*	*amaréis*
3PL	*aman*	*amaron*	*amaban*	*amarían*	*amarán*

(Chan 2008, Lignos & Yang 2016)

Input Sparsity: Paradigm Saturation

- Long-tailed distributions in morphology: **Paradigm Saturation**
  - How many possible inflected forms does a lemma actually occur in?

$$saturation = \frac{\#\, seen}{\#\, possible}$$

$$= \frac{\mathbf{7}}{\#\, possible}$$

	Present	Preterite	Imperfect	Conditional	Future
1SG	*amo*		*amaba*		*amaré*
2SG		*amaste*			
3SG	*ama*		*amaba*		
1PL	*amamos*				
2PL					
3PL					

(Chan 2008, Lignos & Yang 2016)

Input Sparsity: Paradigm Saturation

- Long-tailed distributions in morphology: **Paradigm Saturation**
  - How many possible inflected forms does a lemma actually occur in?

$$saturation = \frac{\#\ seen}{\#\ possible}$$

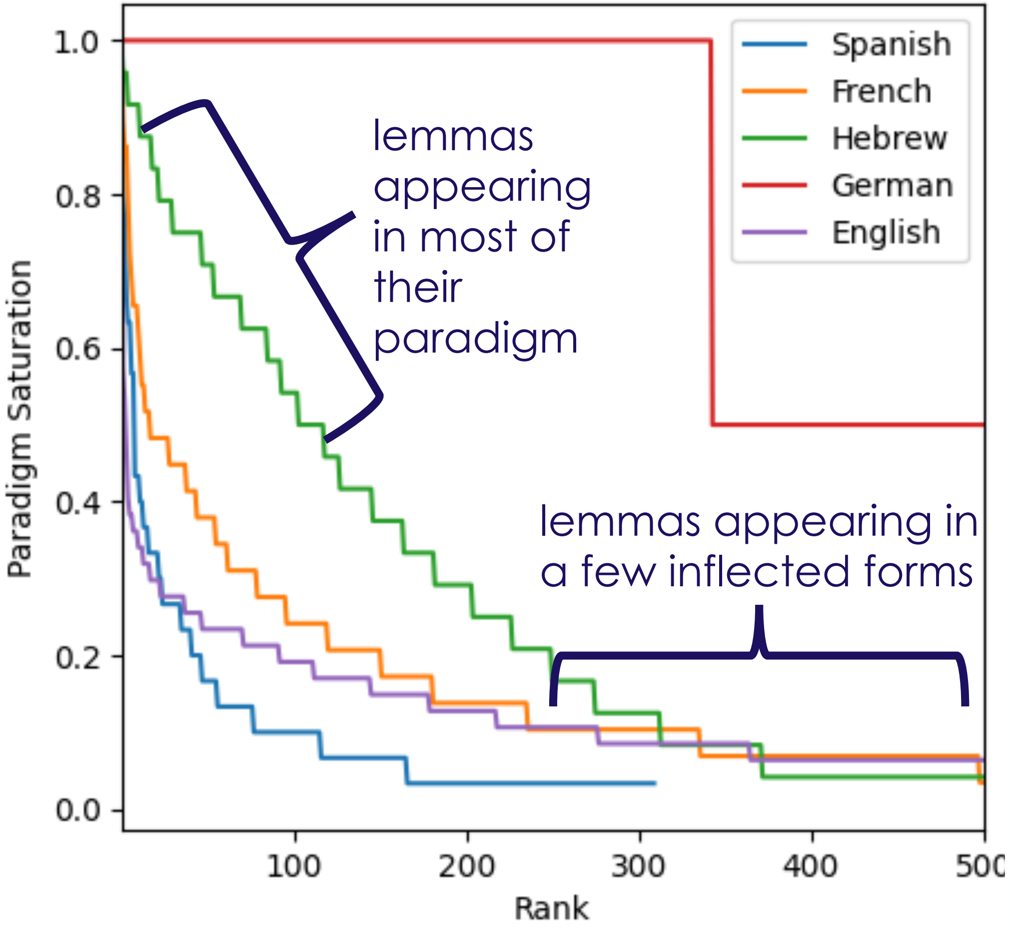
	Present	Preterite	Imperfect	Conditional	Future
1SG	***amo***	*trabajé*	***amaba***	*trabajía*	***amaré***
2SG	*tomas*	***amaste***	*mirabas*	*mirarías*	*esperás*
3SG	***ama***	*esperó*	***amaba***	*espería*	*tomará*
1PL	***amamos***	*miramos*	*mirabamos*	*tomaríamos*	*miraremos*
2PL	*tratáis*				
3PL	*esperan*	*miraron*	*entraban*	*tratarían*	*entrarán*

$$= \frac{\mathbf{7}}{\#\ possible}$$

$$= \frac{\mathbf{7}}{\mathbf{26}} \approx 27\%$$

(Chan 2008, Lignos & Yang 2016)

Input Sparsity: Paradigm Saturation



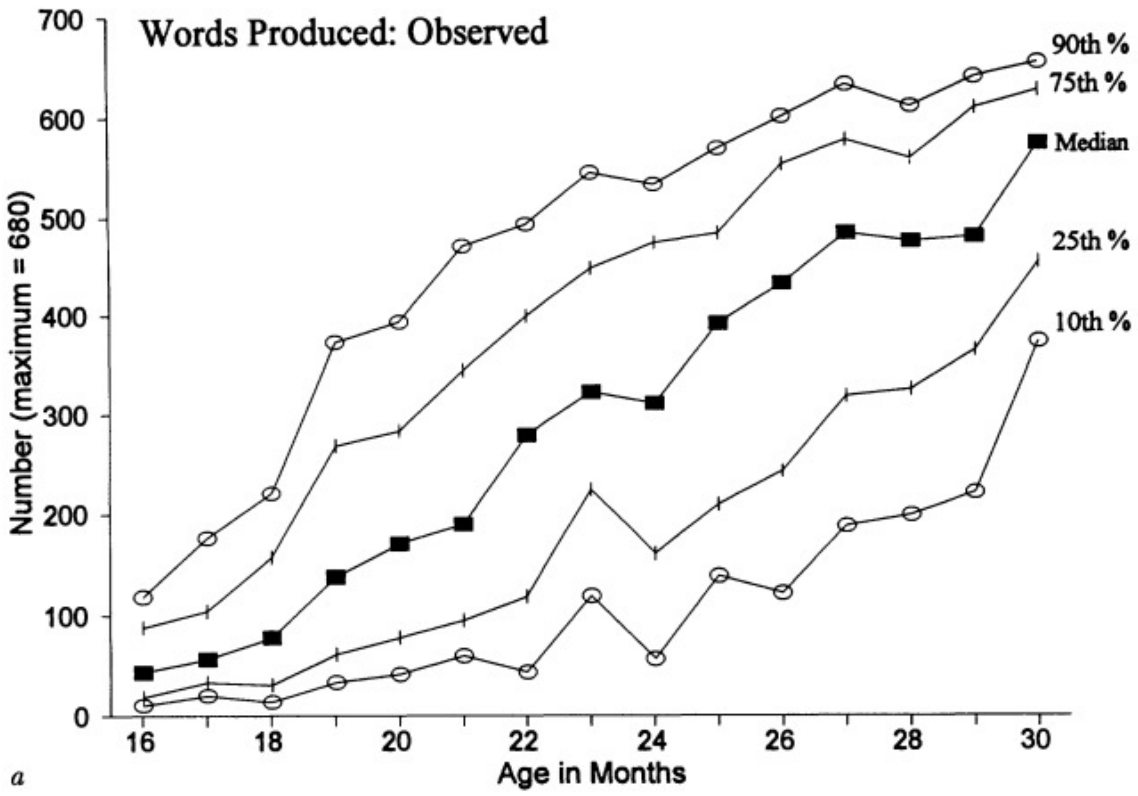
(data from Payne et al 2021, Belth et al 2021, Payne 2022, and Payne 2023)

Input Sparsity: Early Vocabulary

- **At 2;0:** **200-500 words** cross-linguistically
- **At 3;0:** **<1000 words** cross-linguistically
- Early vocabulary makeup:
  - ~50% **nouns**
  - ~25% **verbs**
- More **frequent** words learned earlier

Bornstein et al. (2004)

age of interest



(from Fenson et al 1994)

Input Sparsity: Summary

- Children must **generalize from small, sparse input**
  - From a **few hundred** of the **most-frequent** forms
  - To unseen **lemmas**
  - To unseen **feature sets**, especially in **highly-inflected languages**

Input Sparsity: Summary

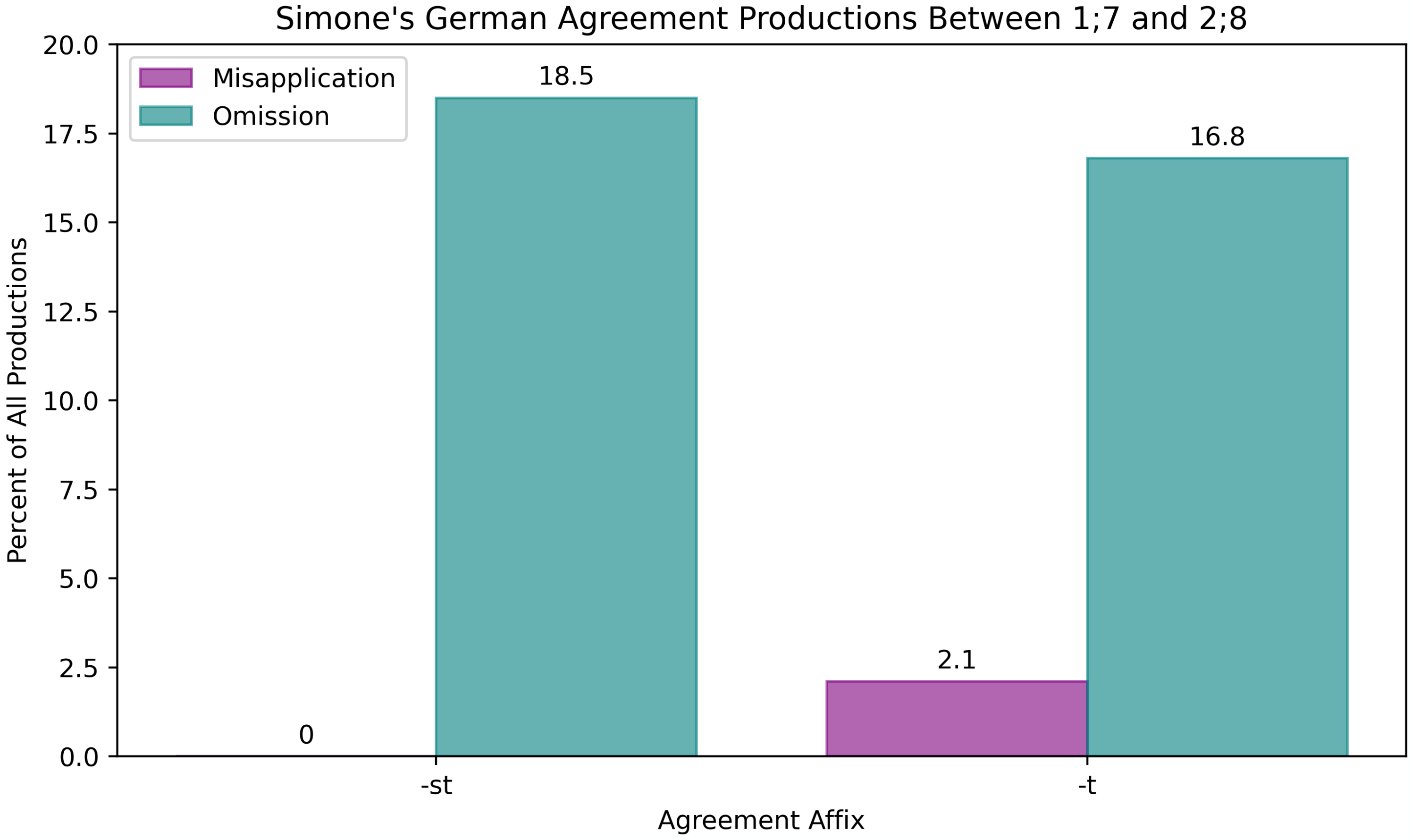
- Children must **generalize from small, sparse input**
- **Previous training data: too much, too saturated**
    - **Kirov & Cotterell: > 3,500** verbs in **entire paradigm**
    - Children know **< 350** verbs at 3;0
    - Would need to see **> 15k lemmas** to see 3,500 in entire paradigm
- **Previous training data:** sampled **uniformly** from UniMorph
    - Kirov & Cotterell, SIGMORPHON shared task, etc.
    - Unnatural **bias towards low-frequency** items
    - Frequency correlated with **irregularity** and **order of acquisition**

Outline

- **Background**
  - Defining the task
  - Input sparsity
  - **Developmental trajectories & error patterns**
- **Developmentally-grounded** evaluation
- Another approach: **Abduction of Tolerable Productivity**
- Revisiting the train-test overlap
- Probing feature-based generalization
- Conclusions

Child Production Errors: Omissions

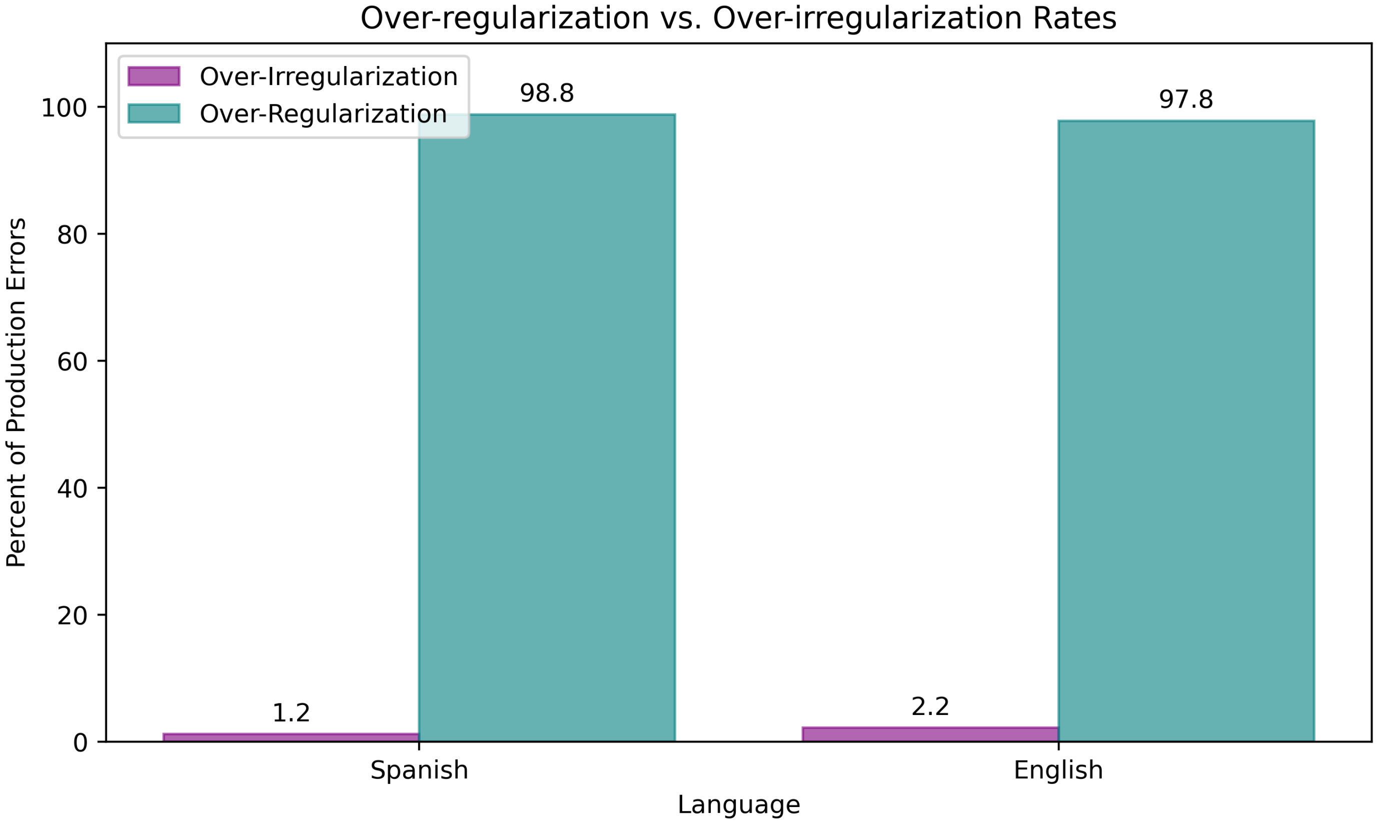
- **Omissions:** ***Root Infinitives***
  - e.g. "Papa have it"
- **Substitutions:** incorrect overt affix
  - e.g. "I has it"



(Clahsen & Penke 1992, Philips 1995, Legate & Yang 2007)

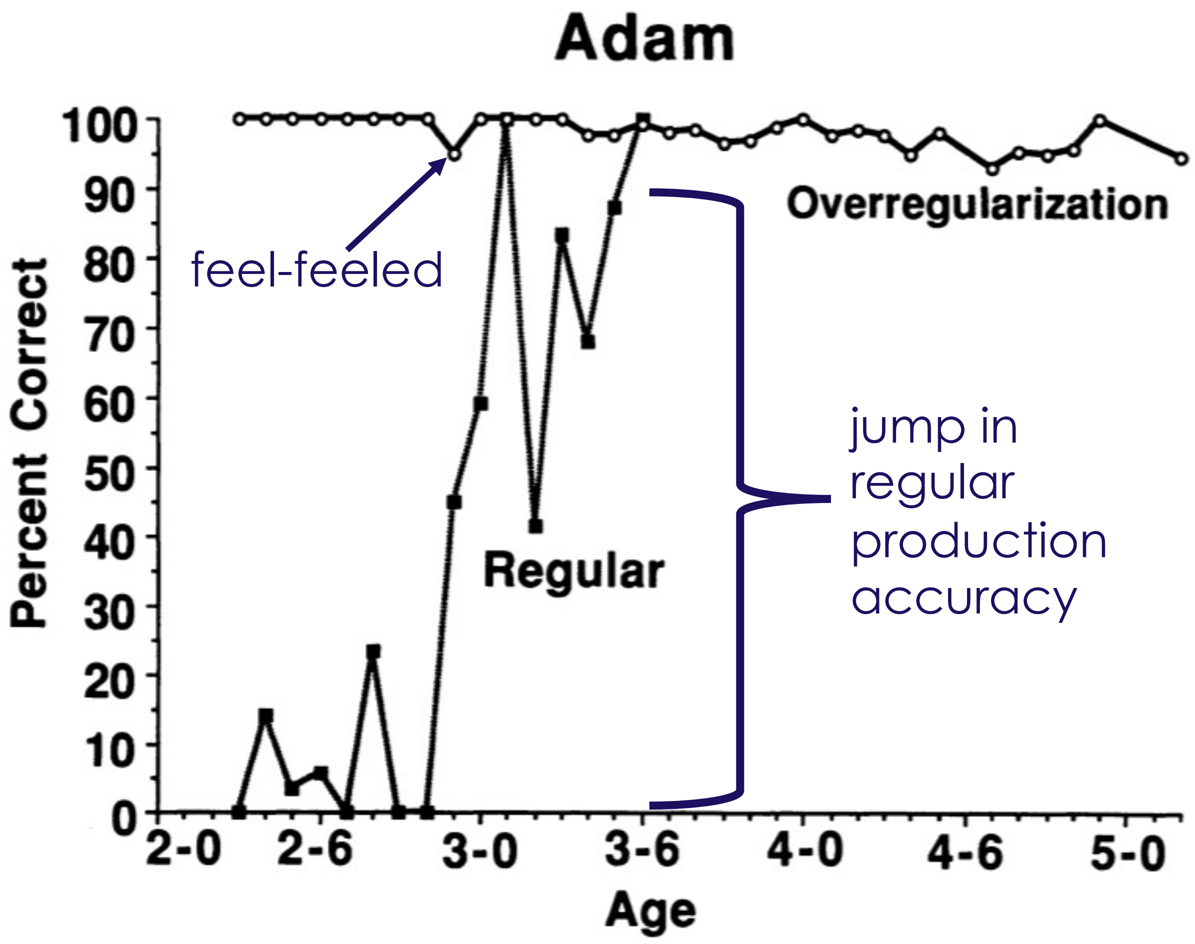
Child Production Errors: Over-regularization

- **Over-regularization**
  - e.g. ***feel-feeled***
- **Over-irregularization**
  - e.g. ***bite-bote***



(Maslen et al 2004, Xu & Pinker 1995, Clahsen et al 2002)

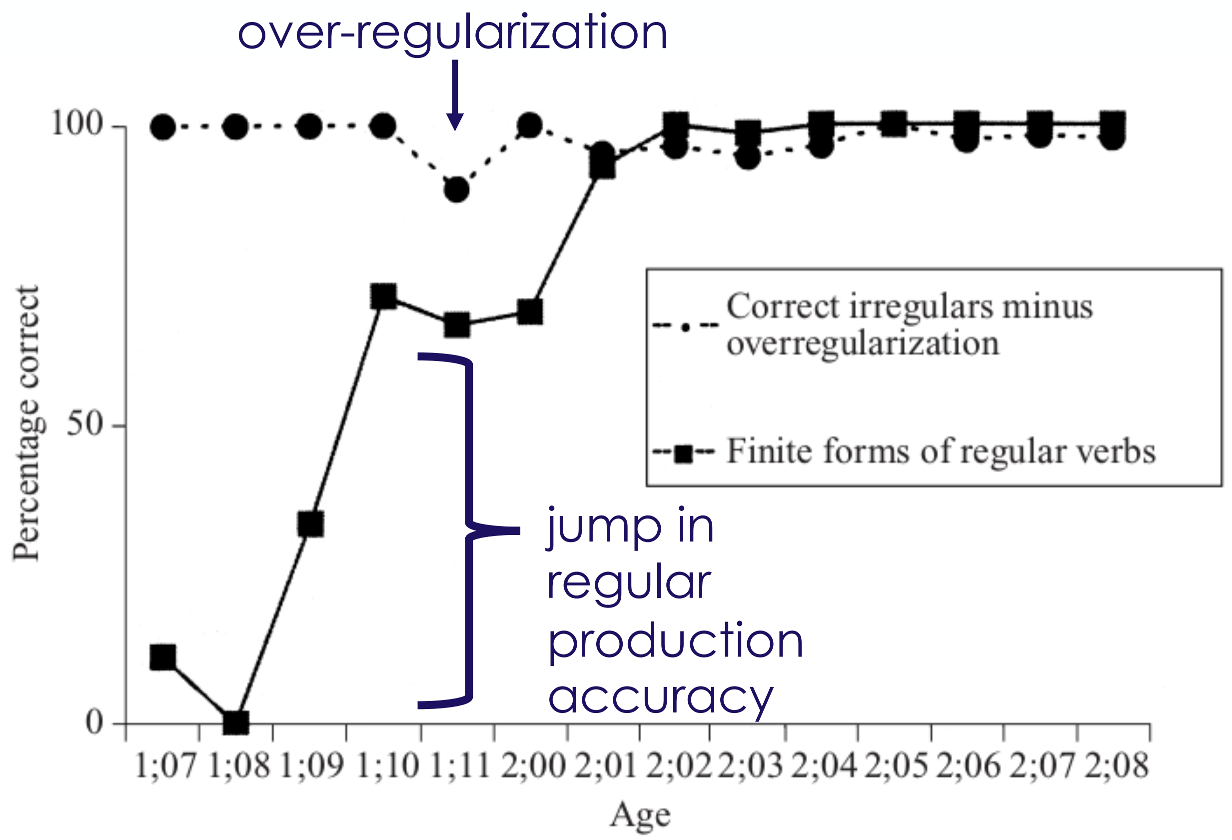
Developmental Trajectories: Regression



(from Marcus et al 1992)

Developmental Trajectories: Regression

Spanish Finiteness Marking



(from Clahsen, Aveledo, and Roca 2002)

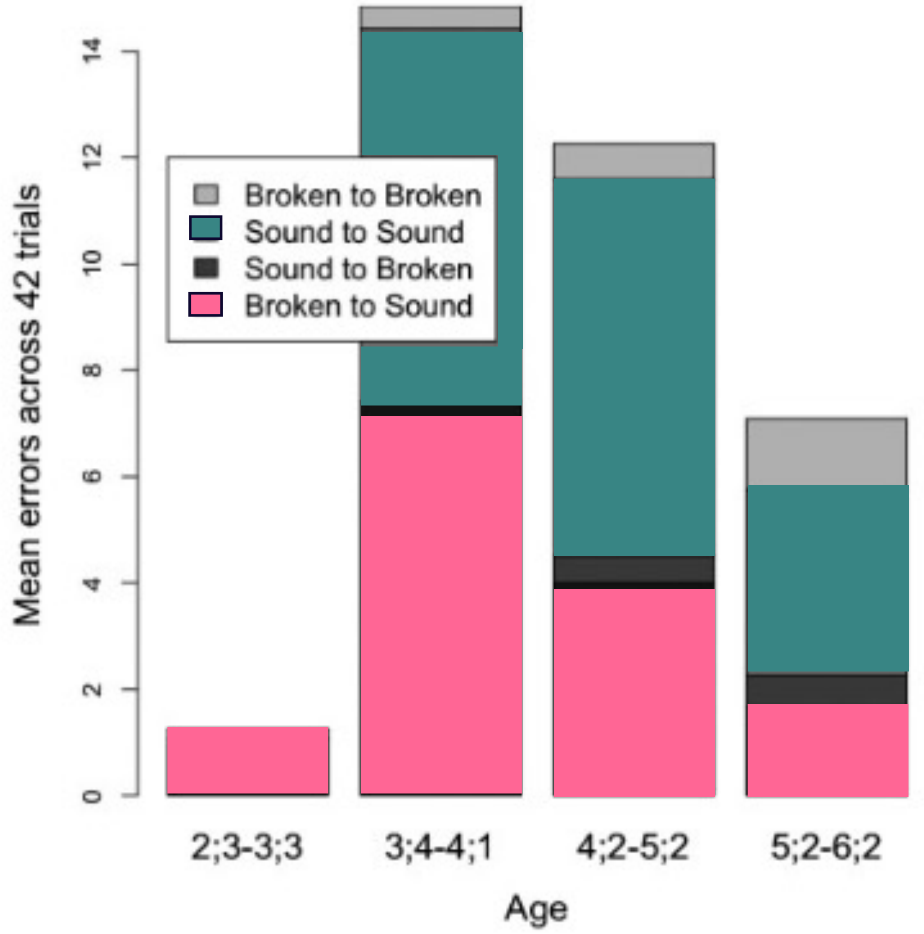
Developmental Trajectories: Regression

Two kinds of **developmental regression** for children learning **Palestinian Arabic noun plurals:**

Masc sound → Fem sound

Broken → Fem sound

Pluralization Errors in Ravid & Farah (1999)



Outline

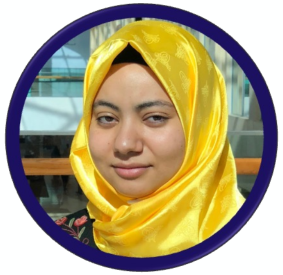
- **Background**
  - Defining the task
  - Input sparsity
  - Developmental trajectories & error patterns
- **Developmentally-grounded evaluation**
- Another approach: **Abduction of Tolerable Productivity**
- Revisiting the train-test overlap
- Probing feature-based generalization
- Conclusions

Kodner, Khalifa, Payne & Liu (Cogsci, 2023)

- Do ANNs match **developmental trajectories** and **error patterns** of children?
- Detailed analysis of **3 well-studied developmental phenomena**:
  - **English past tense** (800 train + 200 ftune)
    - Children learn English past tense on **< 300 verbs**
  - **German noun plurals** (480 train, 120 ftune)
  - **Arabic noun plurals** (800 train + 200 ftune)

Frequency-weighted sampling

Jordan Kodner



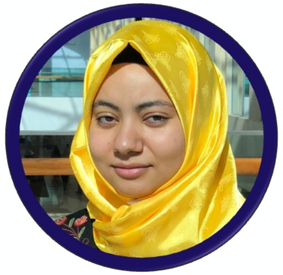
Salam Khalifa

Zoey Liu

Kodner, Khalifa, Payne & Liu (Cogsci, 2023)

- Do ANNs match **developmental trajectories** and **error patterns** of children?
- Detailed analysis of **3 well-studied developmental phenomena**
- **4 models:**
  - **CLUZH-B4:** character-level **transducer** that significantly outperformed the 2022 SIGMORPHON baseline, with **beam decoding**
  - **CLUZH-GR:** character-level **transducer** with **greedy decoding**
  - **CHR-TRM:** character-level **transformer** that was used as a baseline in 2021 and 2022 SIGMORPHON shared tasks
  - **NonNeur:** non-neural baseline using a **majority classifier**

Jordan Kodner

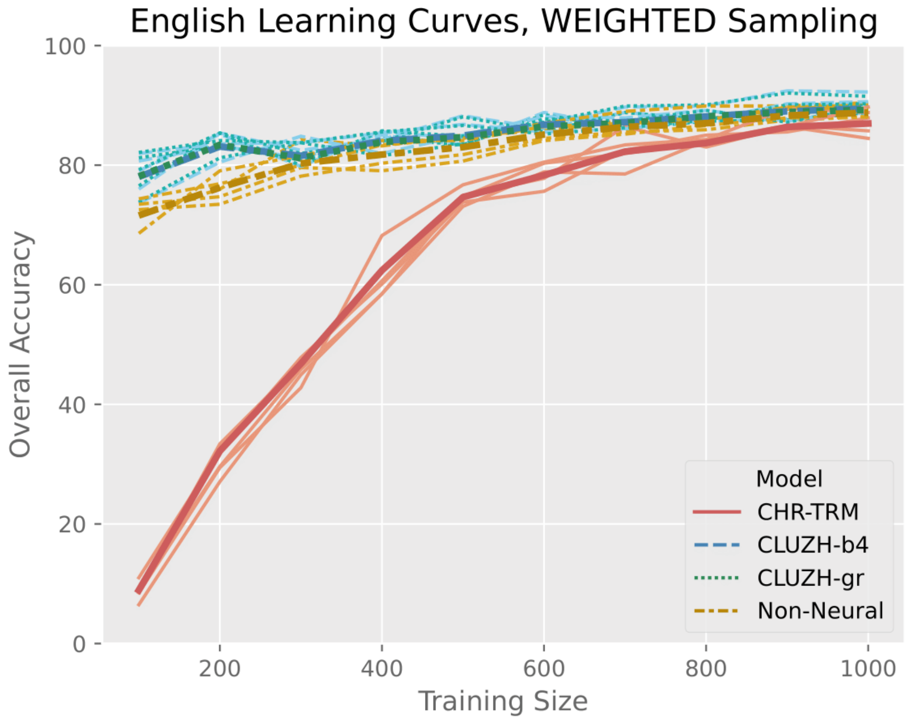


Salam Khalifa

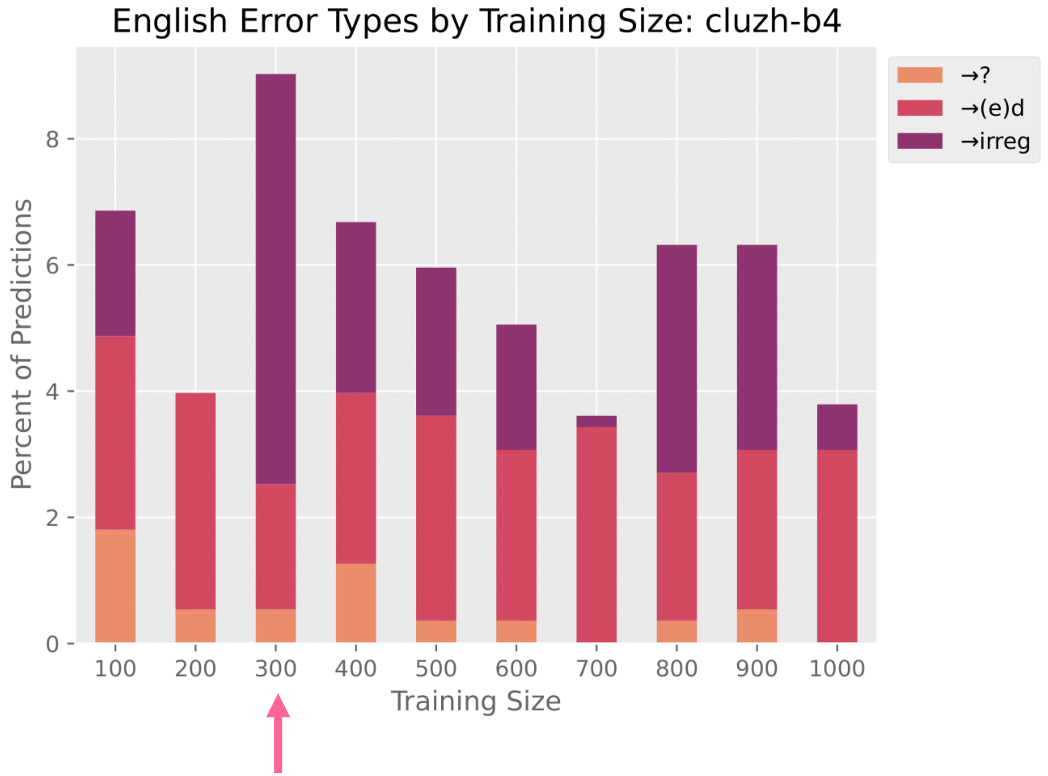
Zoey Liu

Wehrli et al. (2022); Wu et al. (2021); Cotterell et al. (2017)

Model Results: English Past Tense

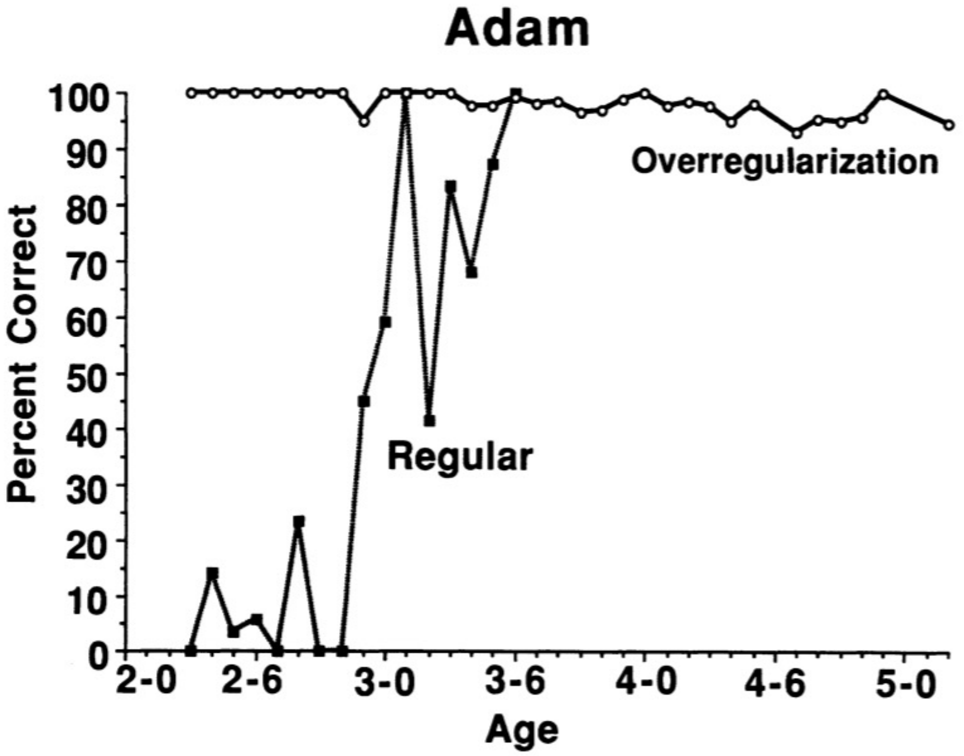


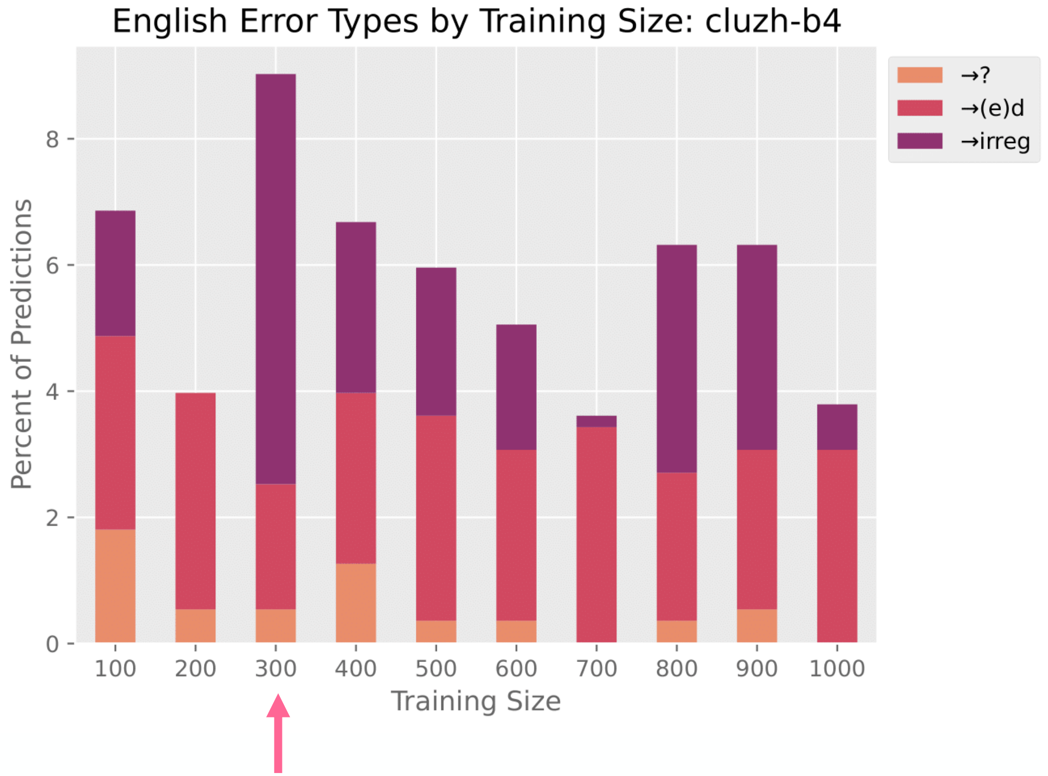
✅ **High overall accuracy**



🚨 **No developmental regression:** spike in error rate caused by **over-irregularization**

Model Results: English Past Tense



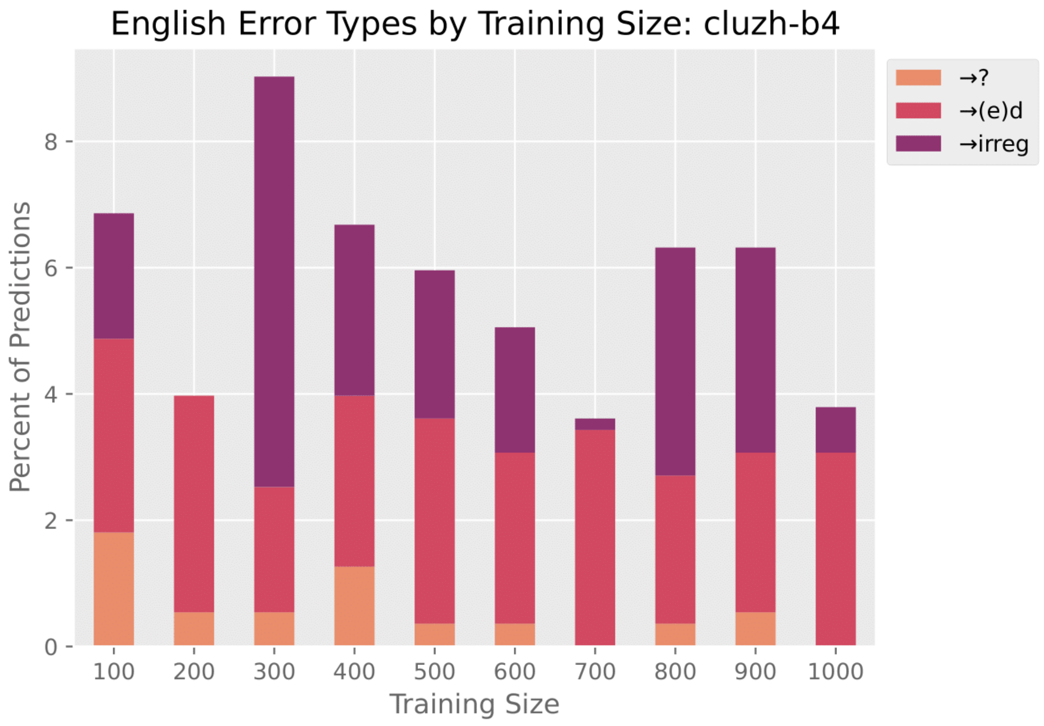


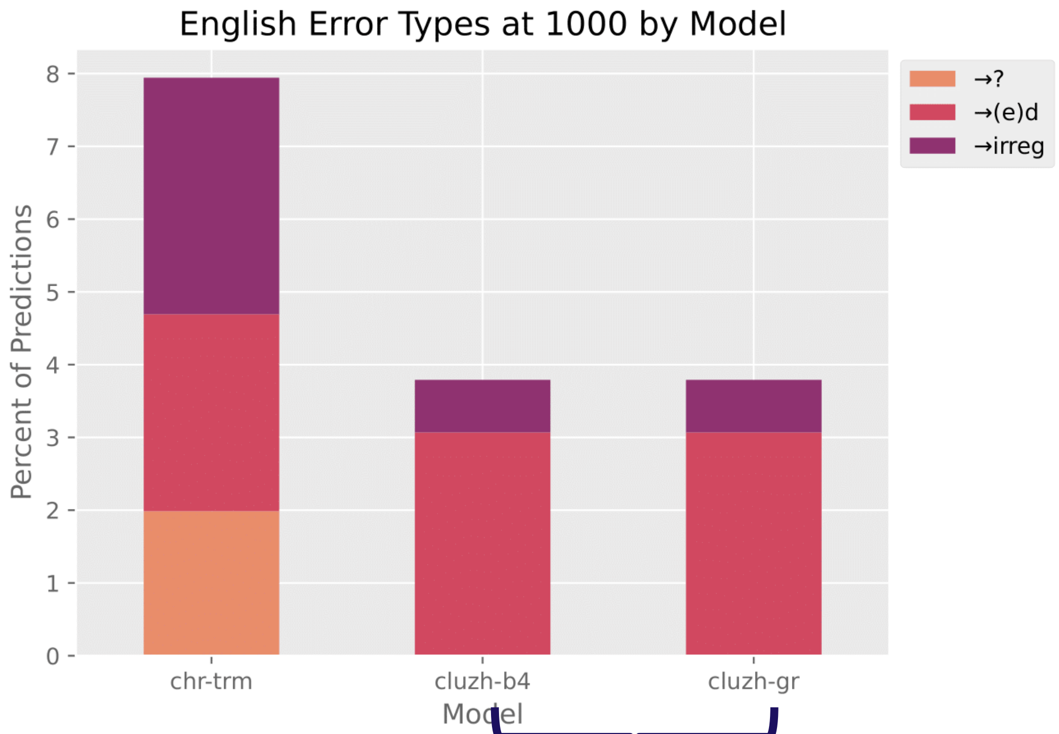
🚨 **Oscillation** in distribution of errors

🚨 **Oscillation** is **not** developmental regression, contra Kirov & Cotterell

🚨 **No developmental regression:** spike in error rate caused by **over-irregularization**

Model Results: English Past Tense





✅ More **over-regularization** than over-irregularization

🚨 Still proportionally **more over-irregularization** than expected (e.g., ***correspond-correspood***)

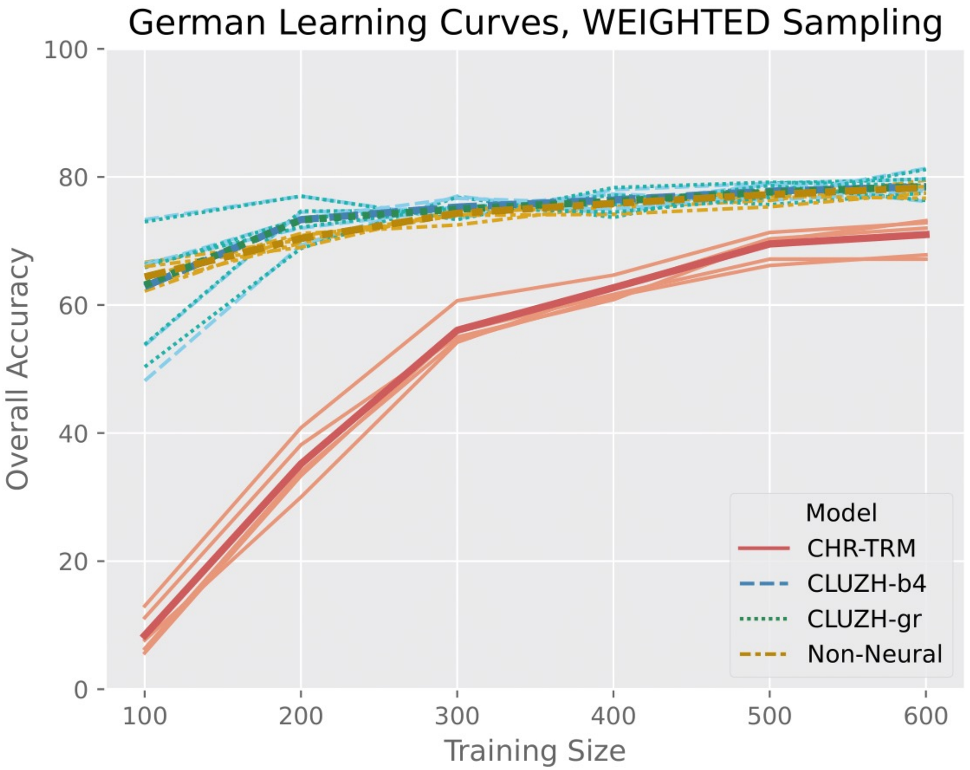
Acquisition Patterns: German Noun Plurals

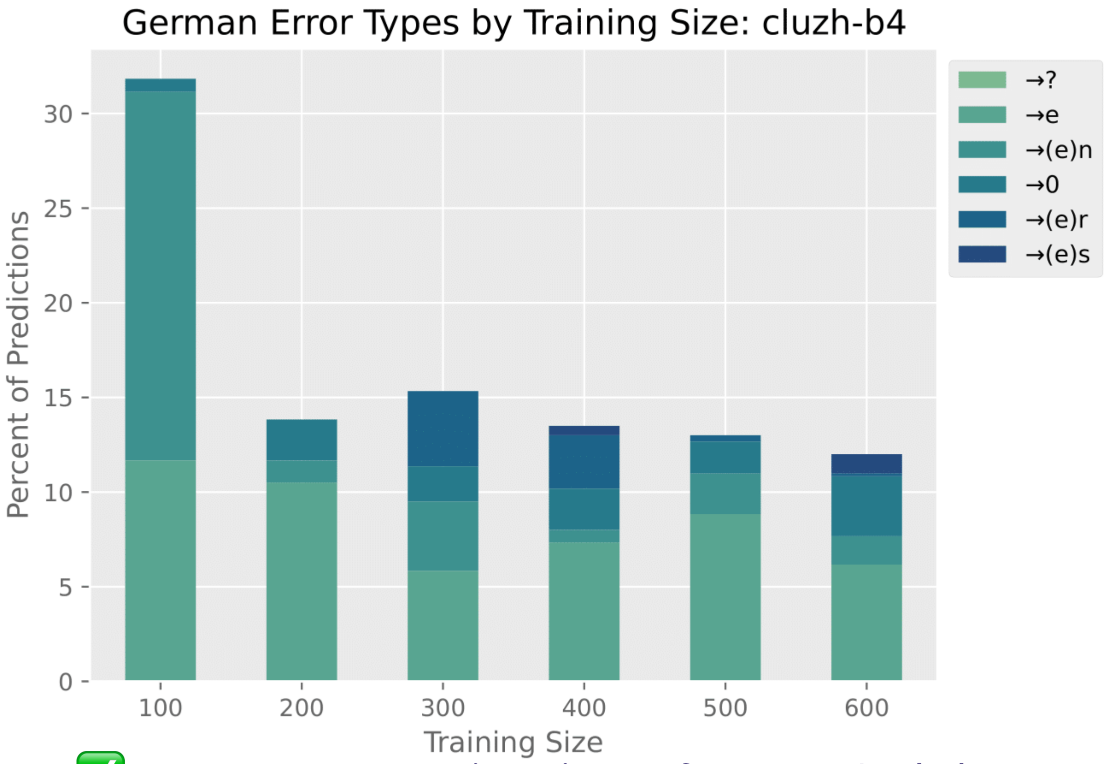
- **Confound** in English verbs:
  - Productive **-ed** is by far the **most frequent**
- German nouns take one of **5 endings**
  - **Gender & stem-final segments** condition affix
  - Interacts with **Umlaut**
  - Apparent default **–s** is the **least frequent**
- Productive use of **–s** appears late

Suffix	Percent
-(e)n	37.3%
-e	34.4%
-∅	19.2%
-er	2.0%
-s	4.0%
other	2.1%

Kopcke (1998); Marcus et al. (1995); Szagun (2001); Elsen (2002); Sonnenstuhl & Huth (2002); Corkerey et al. (2019)

Model Results: German Noun Plurals





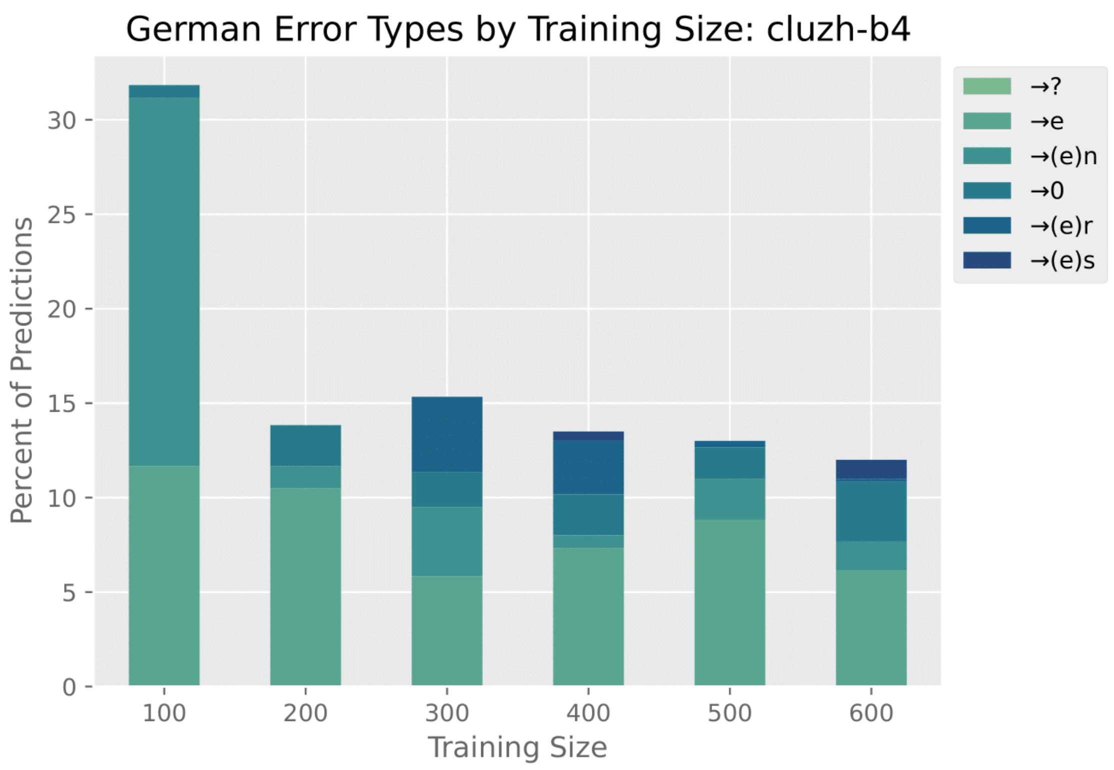
✅ Early overapplication of –e and –*(e)n* fits well with developmental findings

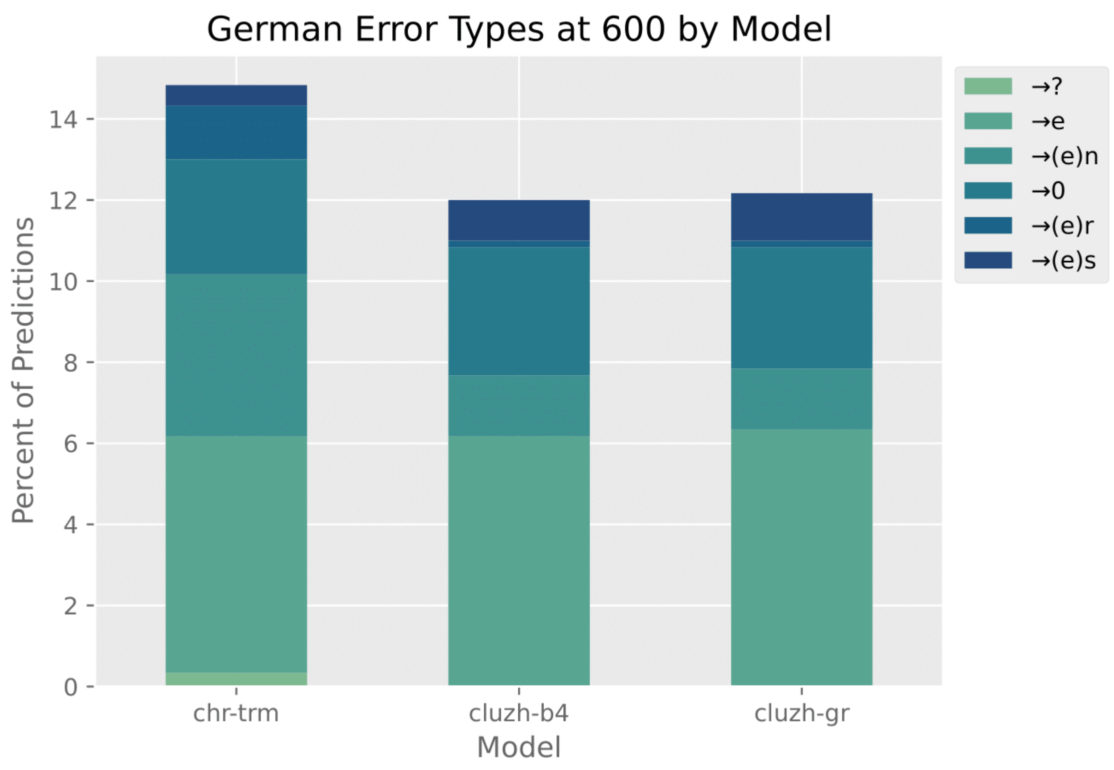
🚨 High error rates

Lower accuracy than English

Gawlitzek-Maiwald (1994); Elsen (2002)

Model Results: German Noun Plurals



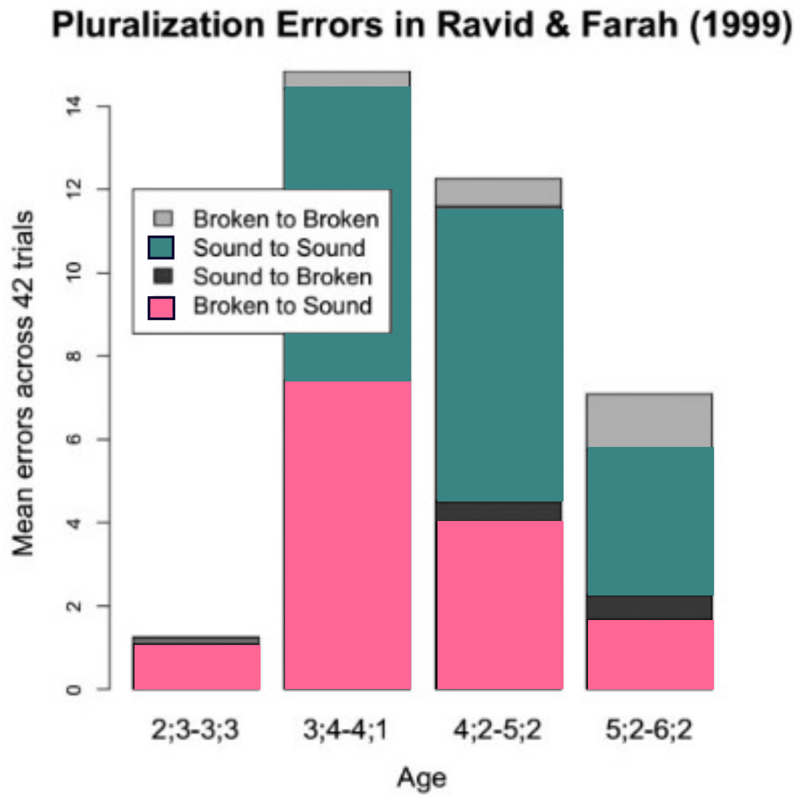


✅ Some **over-application of –s** is present for all systems on full train

✅ Learning trajectories roughly as expected

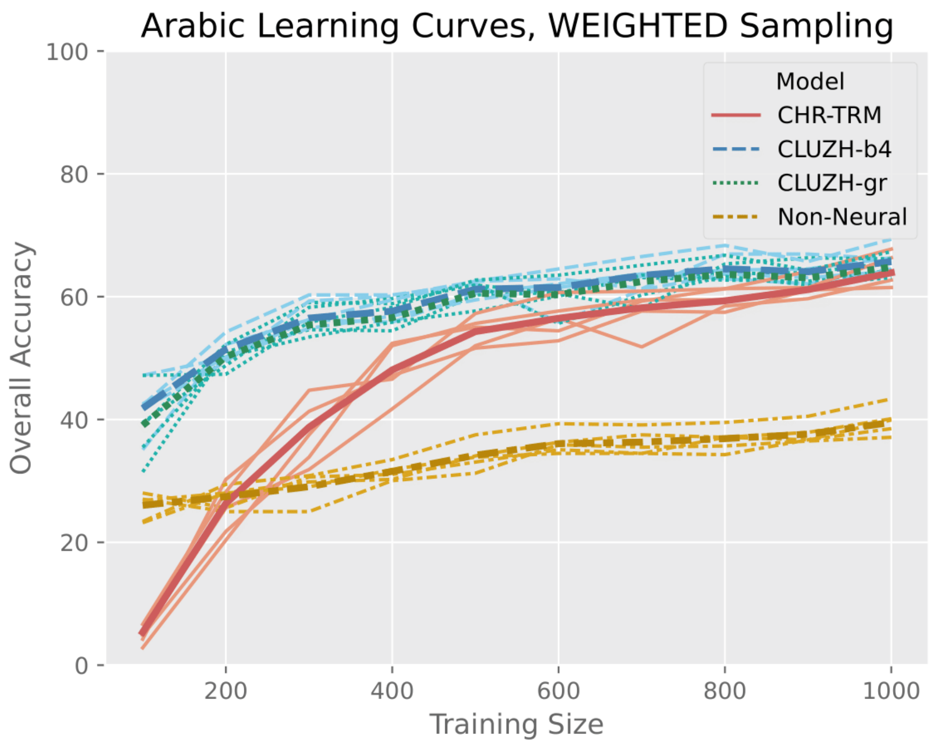
Acquisition Patterns: Arabic Noun Plurals

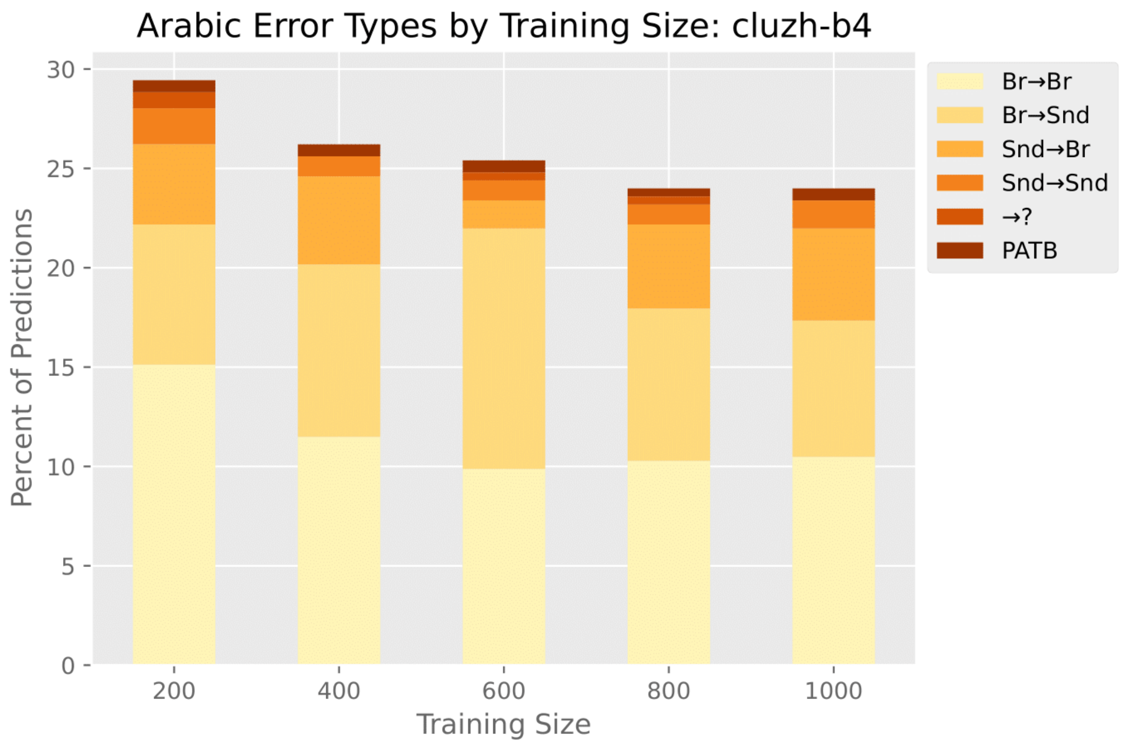
- Two plural types:
  - **Sound plurals** take a suffix
    MASC → *–ūn*, FEM → *–āt*
    some non-human **MASC** nouns take ***–āt***
  - **Broken plurals** undergo a stem change
    ~30 patterns
- Two kinds of **developmental regression:**
  - **MASC sound → FEM sound**
  - **Broken → FEM sound**



Ravid & Farrah (1999); Dawdy-Hesterberg and Pierrehumbert (2014)

Model Results: **Arabic Noun Plurals**

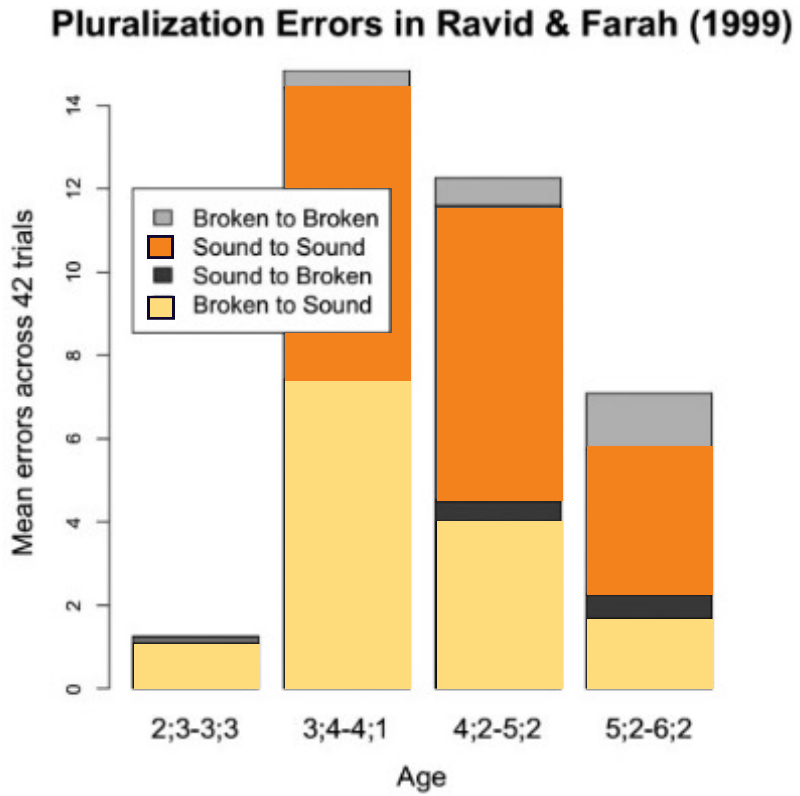


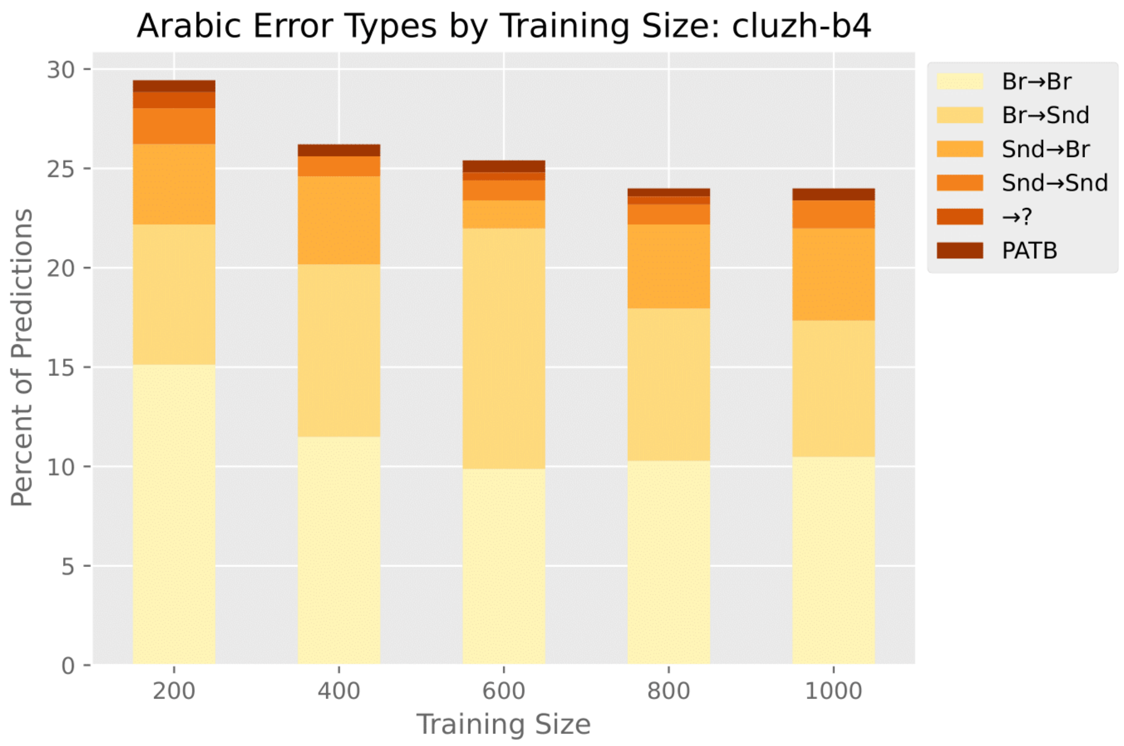


Lower accuracy than English/German

🚨**Learning is monotonic:** neither type of developmental regression is observed

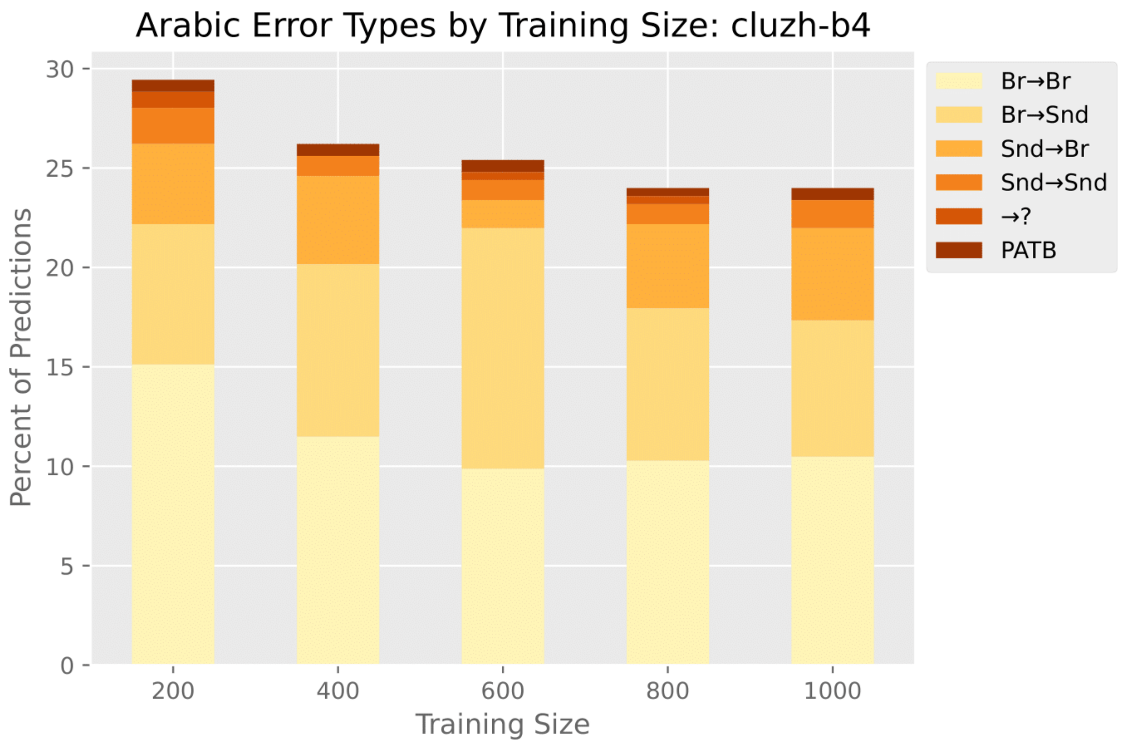
Model Results: **Arabic Noun Plurals**

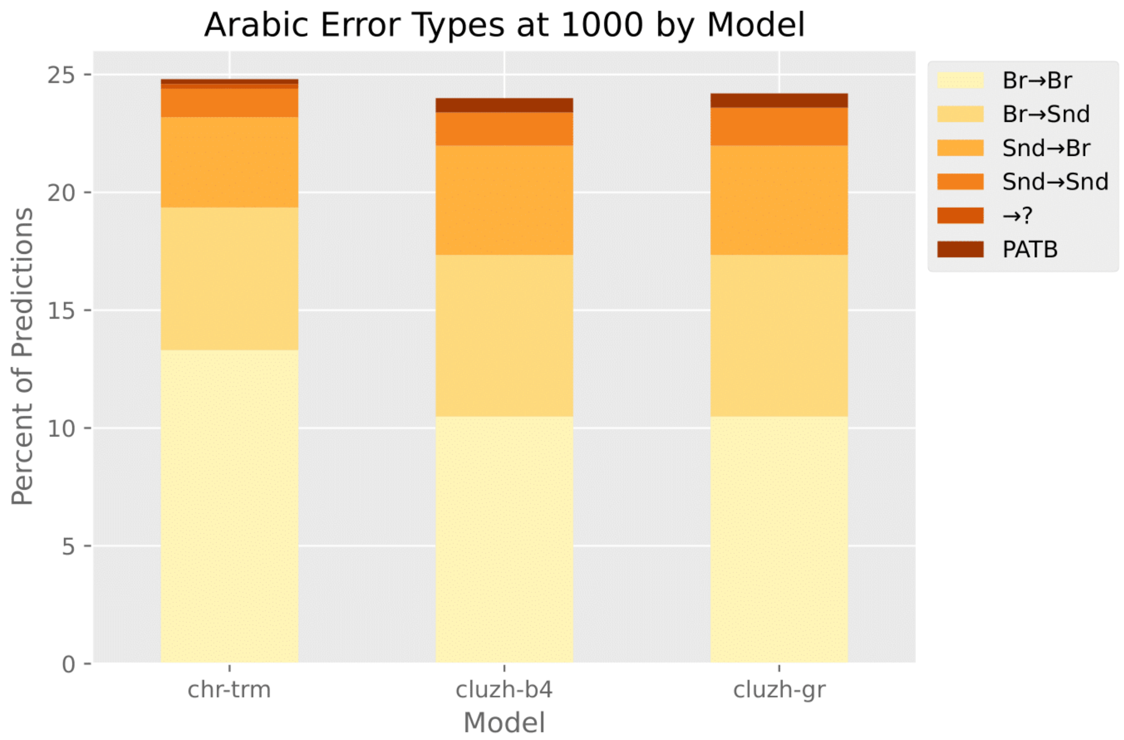




🚨 **Learning is monotonic:** neither type of developmental regression is observed

Model Results: Arabic Noun Plurals





✅ **Broken → Sound** errors relatively common

🚨 **Sound → Sound** errors are rare even though they dominate developmentally

🚨 Most errors are **over-irregularizations:**
Broken → Broken, Sound → Broken

🚨 **FEM → MASC** errors are proportionally much more common than they are developmentally

Interim Summary

- Performance on **English > German > Arabic** reflects pattern complexity
- **Good accuracy overall**, especially considering small training
- But **error patterns are not human-like**
  - Far too much **over-irregularization**
  - **No developmental regression** in English or Arabic
- Current ANNs are clearly **not learning morphology in the same way as humans**

Outline

- **Background**
  - Defining the task
  - Input sparsity
  - Developmental trajectories & error patterns
- **Developmentally-grounded** evaluation
- **Another approach: Abduction of Tolerable Productivity**
- Background
- Revisiting the train-test overlap
- Probing feature-based generalization
- Conclusions

ATP: Making Sense of Production Errors

Caleb Belth

Deniz Beser

Jordan Kodner

Charles Yang

- Children **over-regularize** & don't **over-irregularize**
  - Account for this with **rule-based mappings**:
    - Apply rule when no exception known
      - **Over-regularization** when exception not yet learned
      - **Developmental regression** when rule first learned

Preliminaries: The Tolerance Principle

Intuitions: given a set of *N* items:

- If ***most*** do X, then all do X **(generalization)**
- If ***few*** do X, memorize those that do **(lexicalization)**

Tolerance of exceptions

Generalize a rule applying to **N items** with **e exceptions** iff:

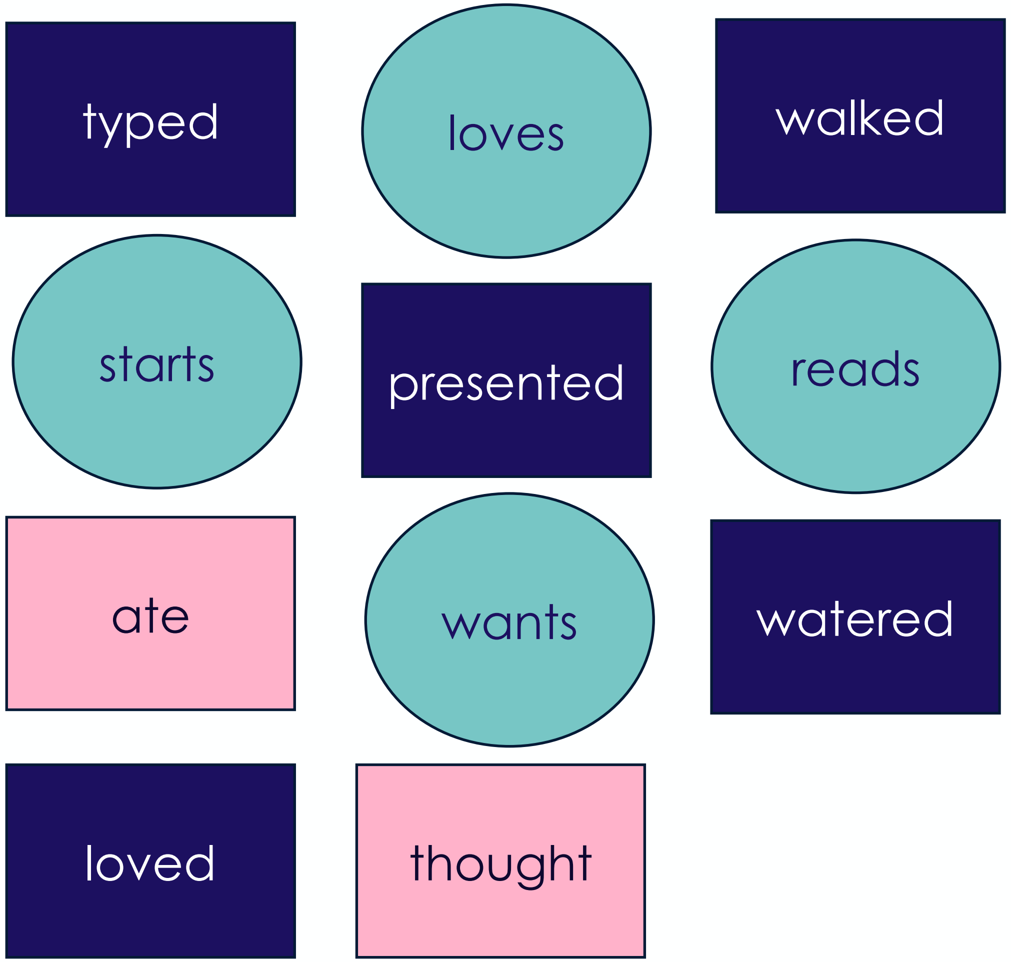
$$e \leq \theta_N = \frac{N}{\ln N}$$

(Yang 2016)

ATP Model: Recursive Subdivision

- Apply TP **recursively**
  - Given ***N*** items, do **enough** of them take **-x affix**?
    - If yes, **productive rule learnt!**
    - If not, **subdivide** into disjoint subsets & **recurse**
- **Terminate** when:
  - Productive rule found **(generalization)**
  - No more subdivisions possible **(lexicalization)**
- Apply to **English past tense** and **German noun plurals**

ATP Model: Toy Example



- 11 items: 4 **-s**, 5 **-ed**, 2 **other**
- **Generalize** most frequent?
  $\boldsymbol{N - M = 11 - 5 = 6 > \theta_{11} = 4.5}$
- **Subdivide!** Hypothesize a rule:

ATP Model: Toy Example

typed

walked

presented

ate

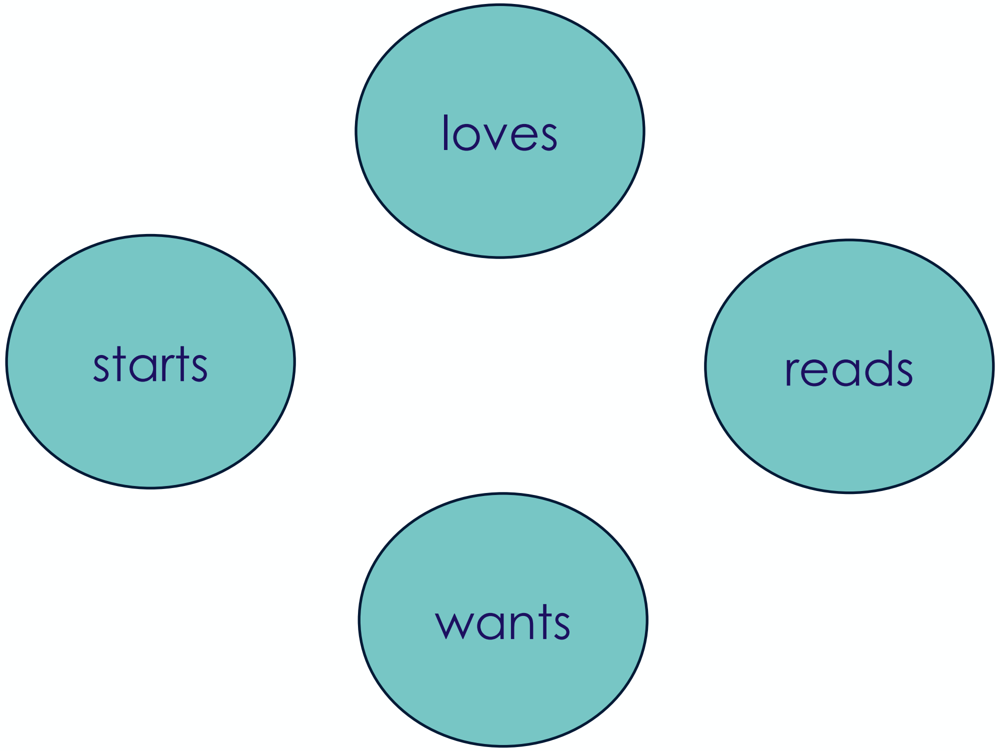
watered

loved

thought

- 11 items: 4 **-s**, 5 **-ed**, 2 **other**
- **Generalize** most frequent?
  🚨 $N - M = 11 - 5 = 6 > \theta_{11} = 4.5$
- **Subdivide!** Hypothesize a rule:
  - PAST → ***-ed***
- **Test** the rule:
  - $N - M = 2 < \theta_7 = 3.5$ ✅
- R1 productive! PAST → ***-ed***
  - Memorize ***ate*** and ***thought***

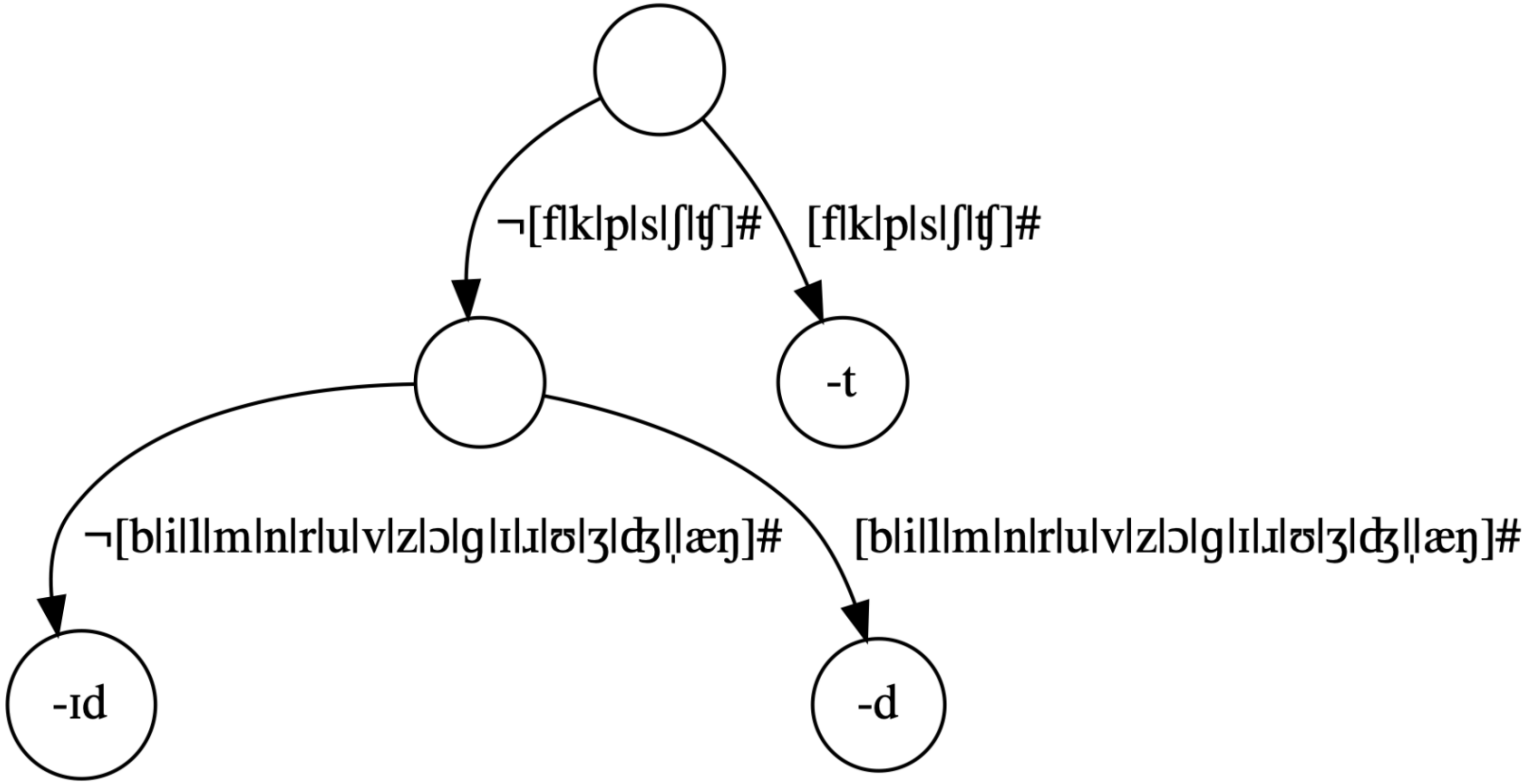
ATP Model: Toy Example



- 11 items: 4 **-s**, 5 ***-ed***, 2 **other**
- **Generalize** most frequent?
  🚨 $\boldsymbol{N - M = 11 - 5 = 6 > \theta_{11} = 4.5}$
- **Subdivide!** Hypothesize a rule:
  - PAST → ***-ed***
- **Test** the rule:
  - $\boldsymbol{N - M = 2 < \theta_7 = 3.5}$ ✅
- R1 productive! PAST → ***-ed***
  - Memorize ***ate*** and ***thought***
- **Recurse:** PRES,3,SG → **-s**

ATP Model: Sample learning trace

English past tense: morphophonological conditioning



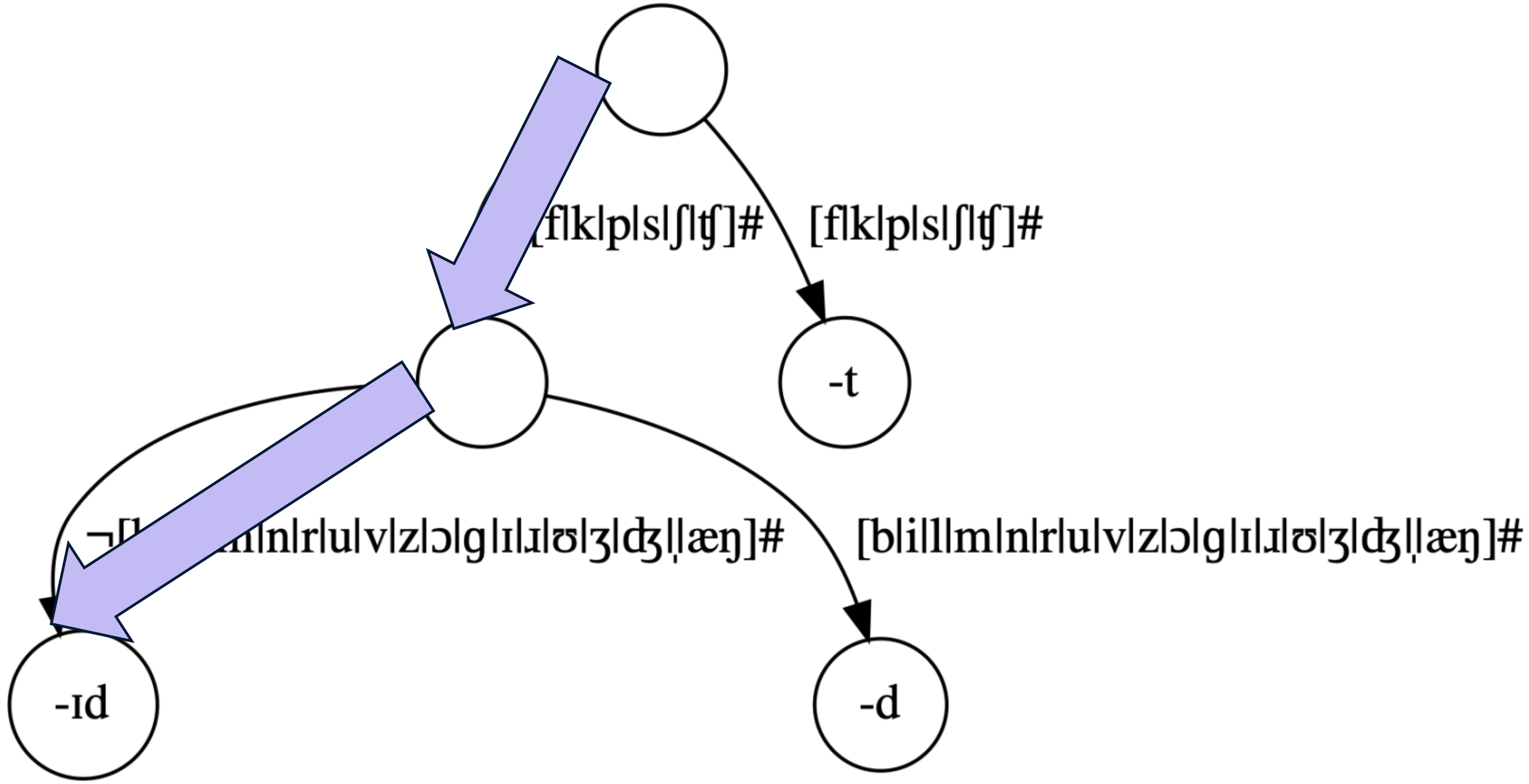
ATP Model: Inflection and Generation

- During test, given **novel forms & features** to inflect
- Traverse decision tree to correct node
  - If node has ***productive rule***, apply the rule
  - If no ***productive rule***, either:
    - Produce unmarked form
    - **Analogize** to a known form at this node

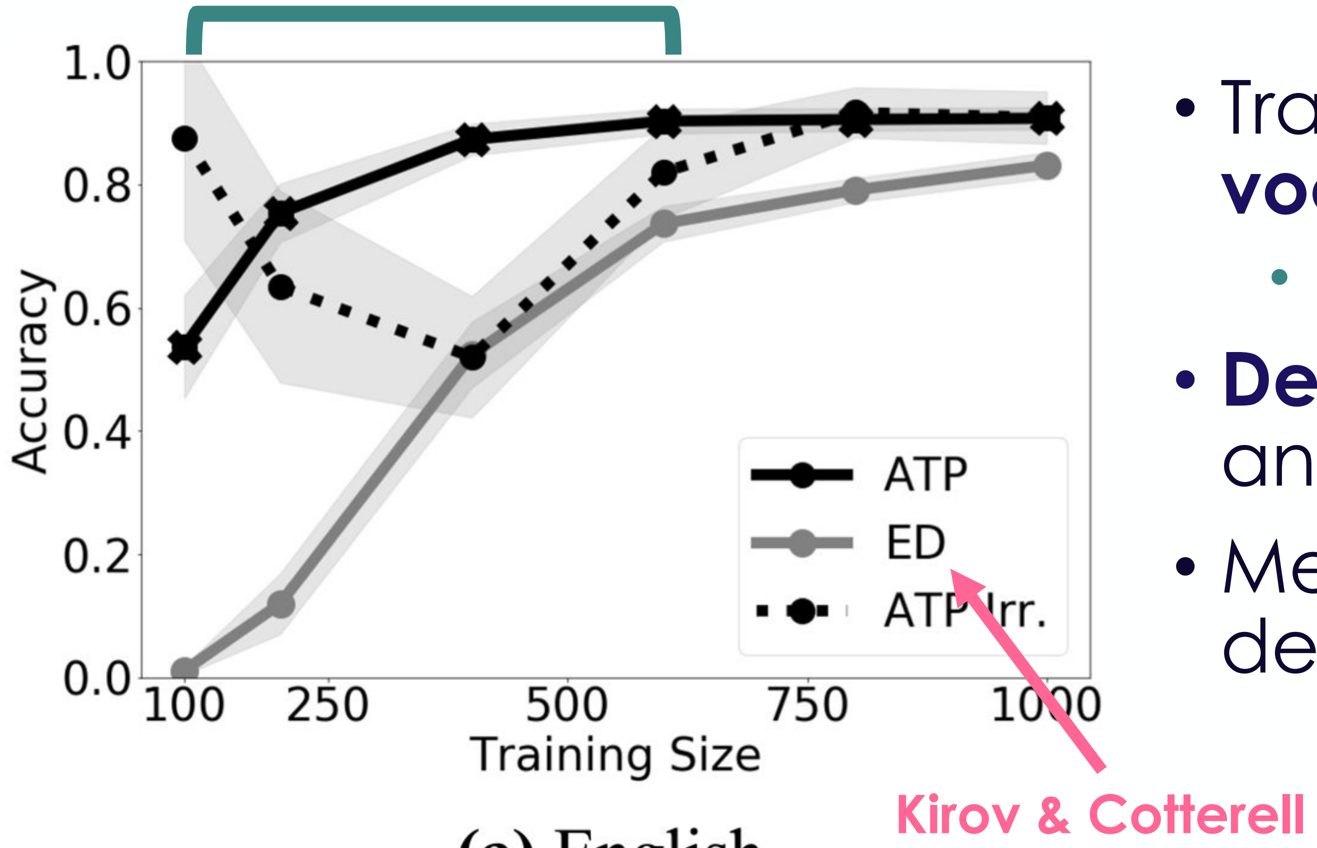
ATP Model: Sample learning trace

English past tense: inflect /want/

/wantɪd/



ATP: English Results

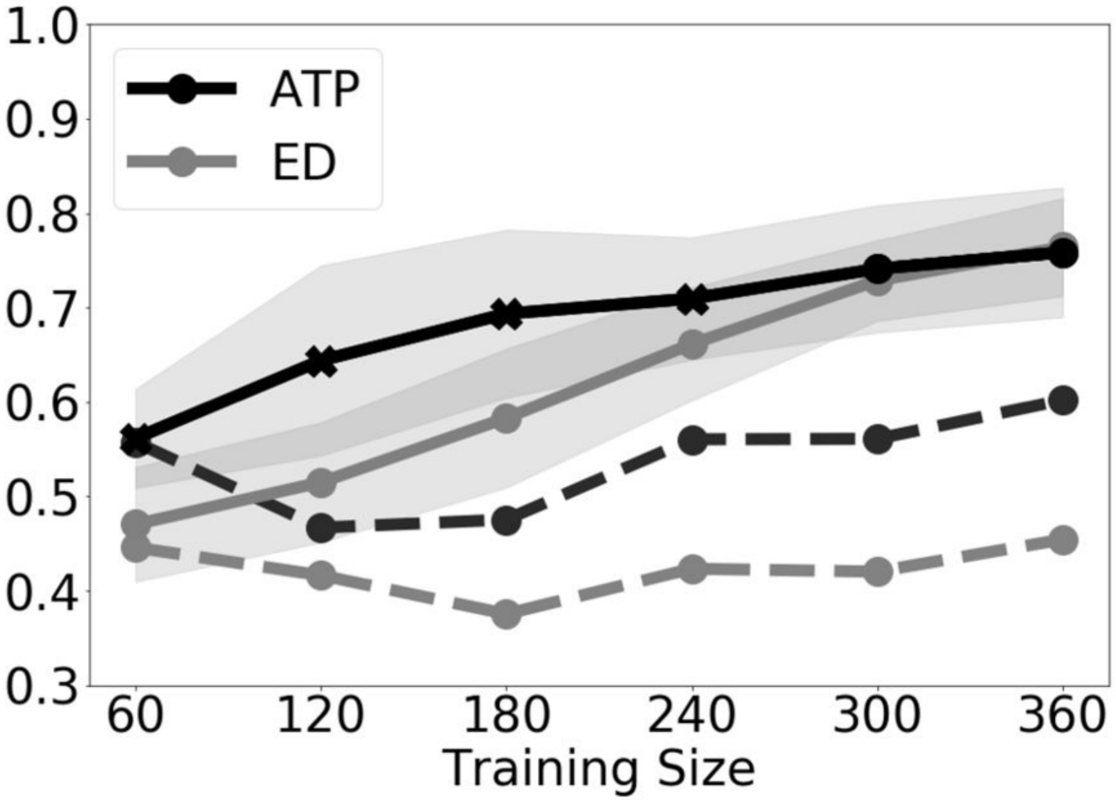


(a) English

- Trained on **plausible vocabulary**
  - **1000** inflected forms
- **Developmental regression** and **overregularization**
- Mechanistic account of developmental regression

ATP: German Results

- Trained on **plausible vocabulary**
  - **400** inflected forms
- Relies less on gender than K&C ⇒ more human-like
  - **Solid lines** = gender info given at test
  - **Dashed lines** = gender info not given at test



(b) German

ATP: Summary

- Evaluated on **sparse, skewed input**
- Evaluation conducted over **multiple splits** and averaged
- **Human-like error patterns**
  - **Over-regularization**
  - **Developmental regression**

ATP gives a **mechanistic account** of *why* these errors occur and **how the morphological grammar is acquired from sparse input**

ATP: Future Work

- **Currently:**
  - **Incremental, online** implementation
  - **Evaluation on more languages:** Chinese, Northern East Cree, Icelandic
- **Future work:**
  - **Feature-based generalization** in ATP
  - **Payne et al (2021):** Spanish feature-based generalization in a similar model

Outline

- **Background**
  - Defining the task
  - Input sparsity
  - Developmental trajectories & error patterns
- **Developmentally-grounded** evaluation
- Another approach: **Abduction of Tolerable Productivity**
- **Revisiting the train-test overlap**
- Probing feature-based generalization
- Conclusions

Kodner, Payne, Khalifa & Liu (2023, ACL)

Jordan Kodner

Salam Khalifa

Zoey Liu

- **Three shortcomings** of previous evaluation practices:
  - **Uniform** sampling & **large training** sets
  - **Uncontrolled overlap** between train & test components
  - Evaluation on **single splits**

Revisiting Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?
- Four possible relationships between train & test triples:

Illustrative Train Set

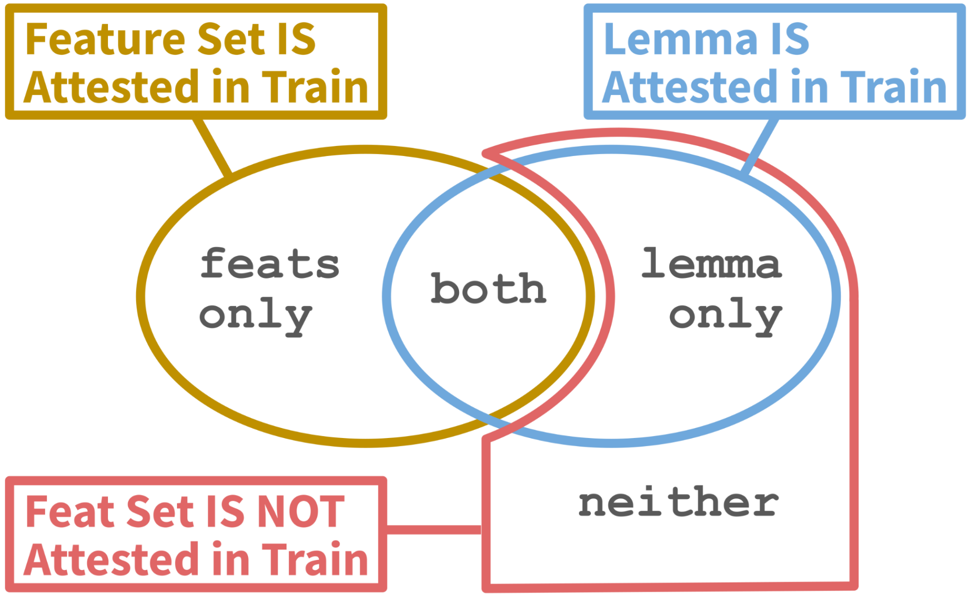
eat	eating	V;V.PTCP;PRS
run	ran	V;PST

Illustrative Test Set

eat	V;PST	← **No OOV**, not attested together
run	V;NFIN	← Only **feature set** is OOV
see	V;PST	← Only **lemma** is OOV
go	V;PRS;3;SG	← **Lemma** and **feature set** are OOV

Revisiting Train-Test Overlap

- No **train triples** appear in test
  - But what about **lemmas** or **feature sets** individually?
- Four possible relationships between train & test triples:



Do lemma and/or feature set overlap predict performance?

Kodner, Payne, Khalifa & Liu (2023, ACL)

- **5 Languages:** German, English, Spanish, Swahili, Turkish
  - **UniMorph 3 + 4** intersected with frequency info for weighted sampling
  - **CHILDES** for German, English, Spanish
  - **Wikipedia** for Swahili & Turkish

Kodner, Payne, Khalifa & Liu (2023, ACL)

- **5 Languages:** German, English, Spanish, Swahili, Turkish
- **3 Split Types:**
  - **UNIFORM:** partition UniMorph **uniformly at random**
  - **WEIGHTED:** partition at random weighted by **type frequency**
  - **OVERLAPAWARE:** enforce a maximum **50% proportion of FEATSATTESTED**

Kodner, Payne, Khalifa & Liu (2023, ACL)

- **5 Languages:** German, English, Spanish, Swahili, Turkish
- **3 Split Types:** UNIFORM, WEIGHTED, OVERLAPAWARE
- **4 Systems:**
  - **CLUZH-B4:** character-level **transducer** that significantly outperformed the 2022 SIGMORPHON baseline, with **beam decoding**
  - **CLUZH-GR:** character-level **transducer** with **greedy decoding**
  - **CHR-TRM:** character-level **transformer** that was used as a baseline in 2021 and 2022 SIGMORPHON shared tasks
  - **NONNEUR:** non-neural baseline using a **majority classifier**

Wehrli et al. (2022); Wu et al. (2021); Cotterell et al. (2017)

Kodner, Payne, Khalifa & Liu (2023, ACL)

- **5 Languages:** German, English, Spanish, Swahili, Turkish
- **3 Split Types:** UNIFORM, WEIGHTED, OVERLAPAWARE
- **4 Systems:** CLUZH-B4, CLUZH-GR, CHR-TRM, NONNEUR
- Re-splitting/re-evaluation on **5 random seeds**

Feature Overlap in Training

400 train
100 ftune
1000 test

SmallTrain	featsAttested	featsNovel	σ
Uniform	80.33	19.67	19.5
Weighted	90.44	9.56	11.1
OverlapAware	48.81	51.19	0.98

1600 train
400 ftune
1000 test

LargeTrain	featsAttested	featsNovel	σ
Uniform	96.17	3.83	5.55
Weighted	95.36	4.64	7.28
OverlapAware	49.92	50.08	0.17

Feature Overlap in Training

400 train
100 ftune
1000 test

SmallTrain	featsAttested	featsNovel	σ
Uniform	80.33	19.67	19.5
Weighted	90.44	9.56	11.1
OverlapAware	48.81	51.19	0.98

1600 train
400 ftune
1000 test

LargeTrain	featsAttested	featsNovel	σ
Uniform	96.17	3.83	5.55
Weighted	95.36	4.64	7.28
OverlapAware	49.92	50.08	0.17

Overlap rate is high but not 100% when not controlled for
UNIFORM & WEIGHTED are similar for large training size

Feature Overlap in Training

400 train
100 ftune
1000 test

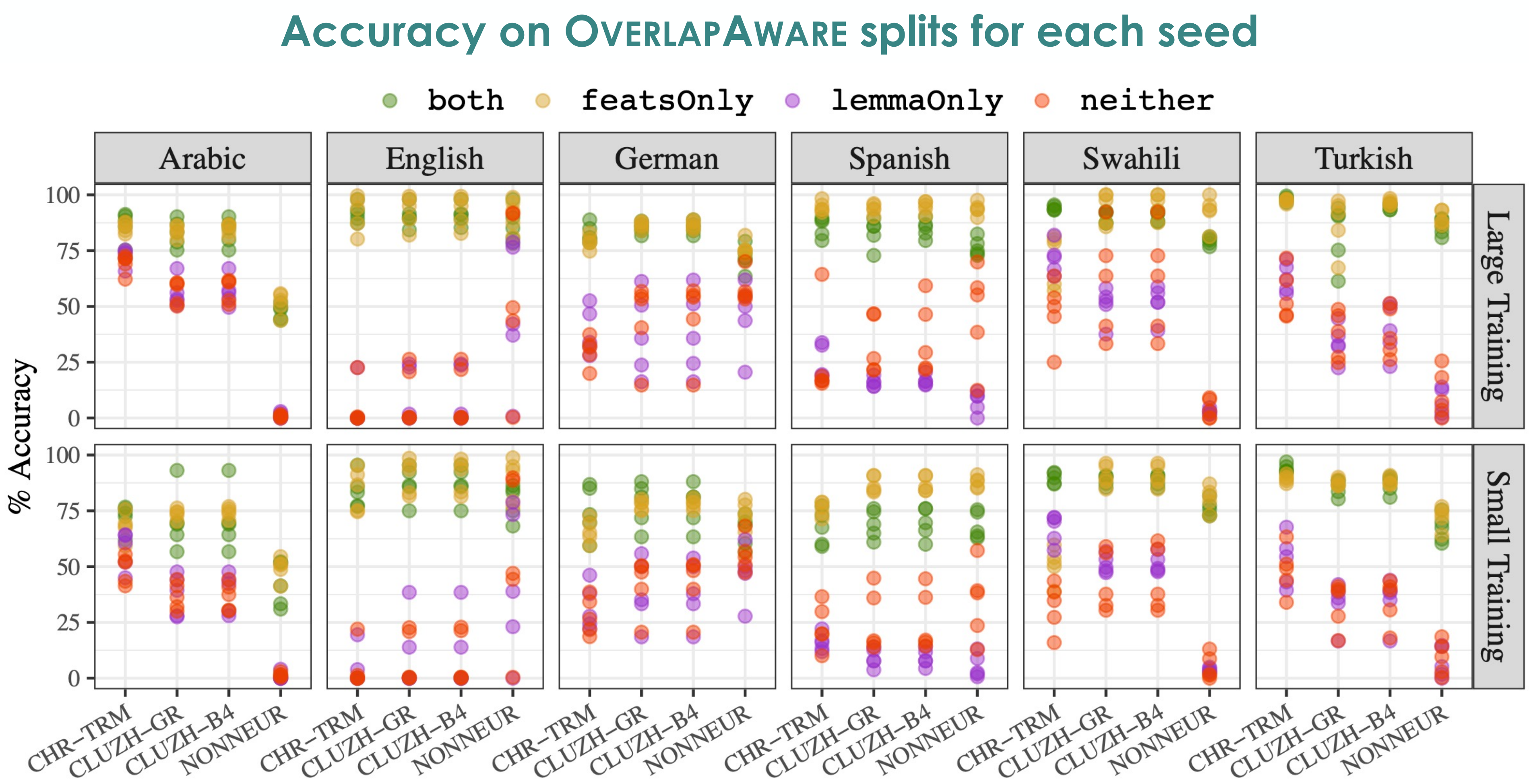
SmallTrain	featsAttested	featsNovel	σ
Uniform	80.33	19.67	19.5
Weighted	90.44	9.56	11.1
OverlapAware	48.81	51.19	0.98

1600 train
400 ftune
1000 test

LargeTrain	featsAttested	featsNovel	σ
Uniform	96.17	3.83	5.55
Weighted	95.36	4.64	7.28
OverlapAware	49.92	50.08	0.17

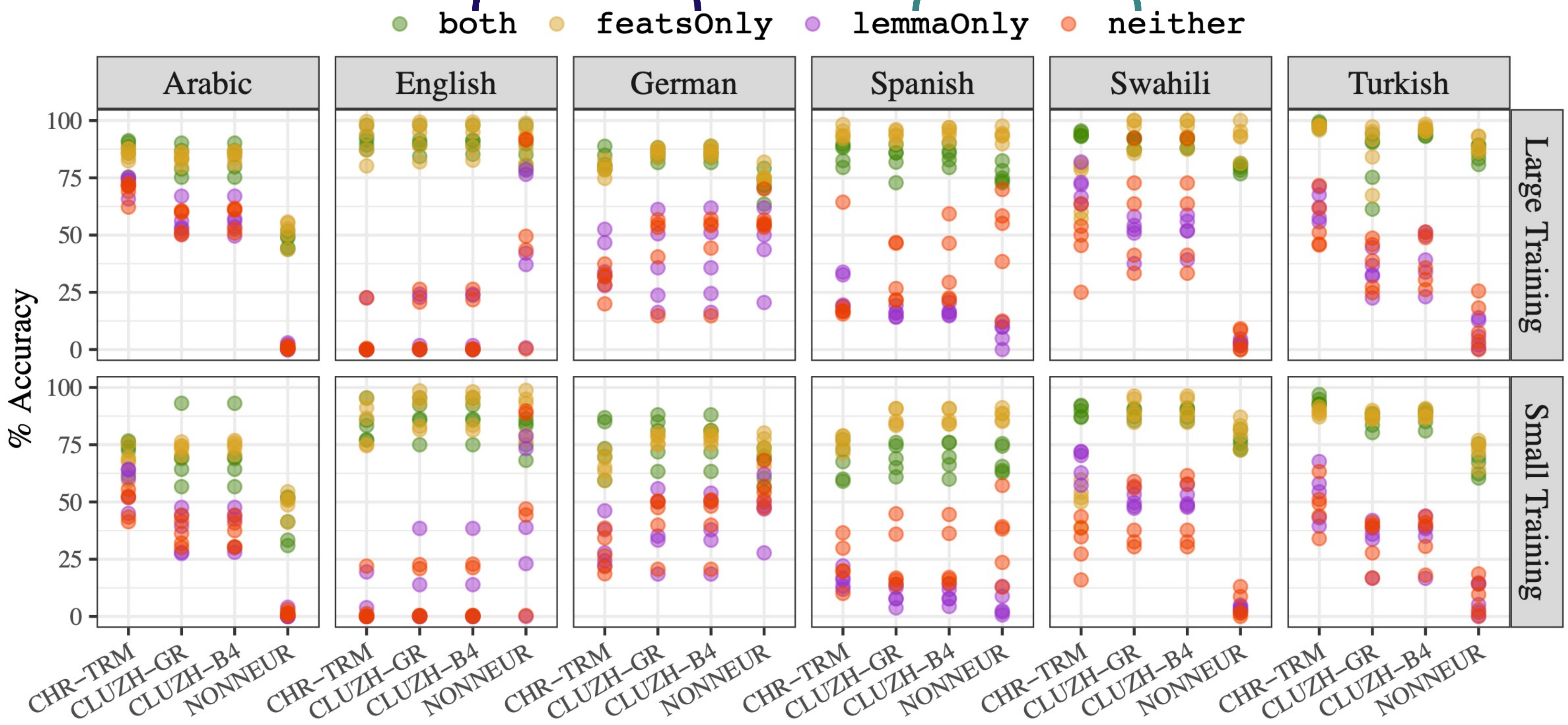
Overlap rate is highly variable across seed/language when not controlled for

Results: Effect of Overlap



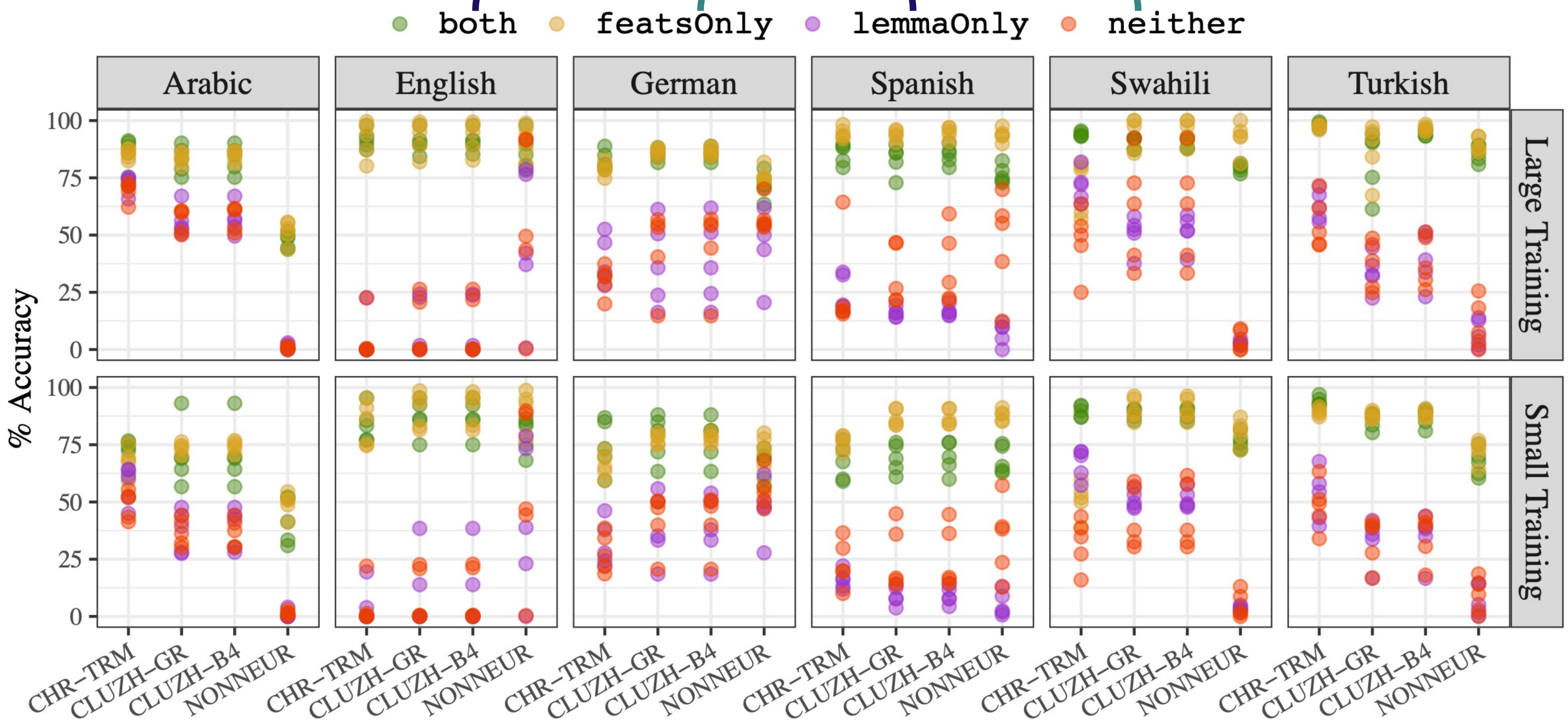
Results: Effect of Feature Overlap

FEATSATTESTED - FEATSNOVEL = **49.75%** (large) **48.02%** (small)



Results: Effect of Lemma Overlap

LEMMAATTESTED - LEMMANOVEL < 2% (small & large)

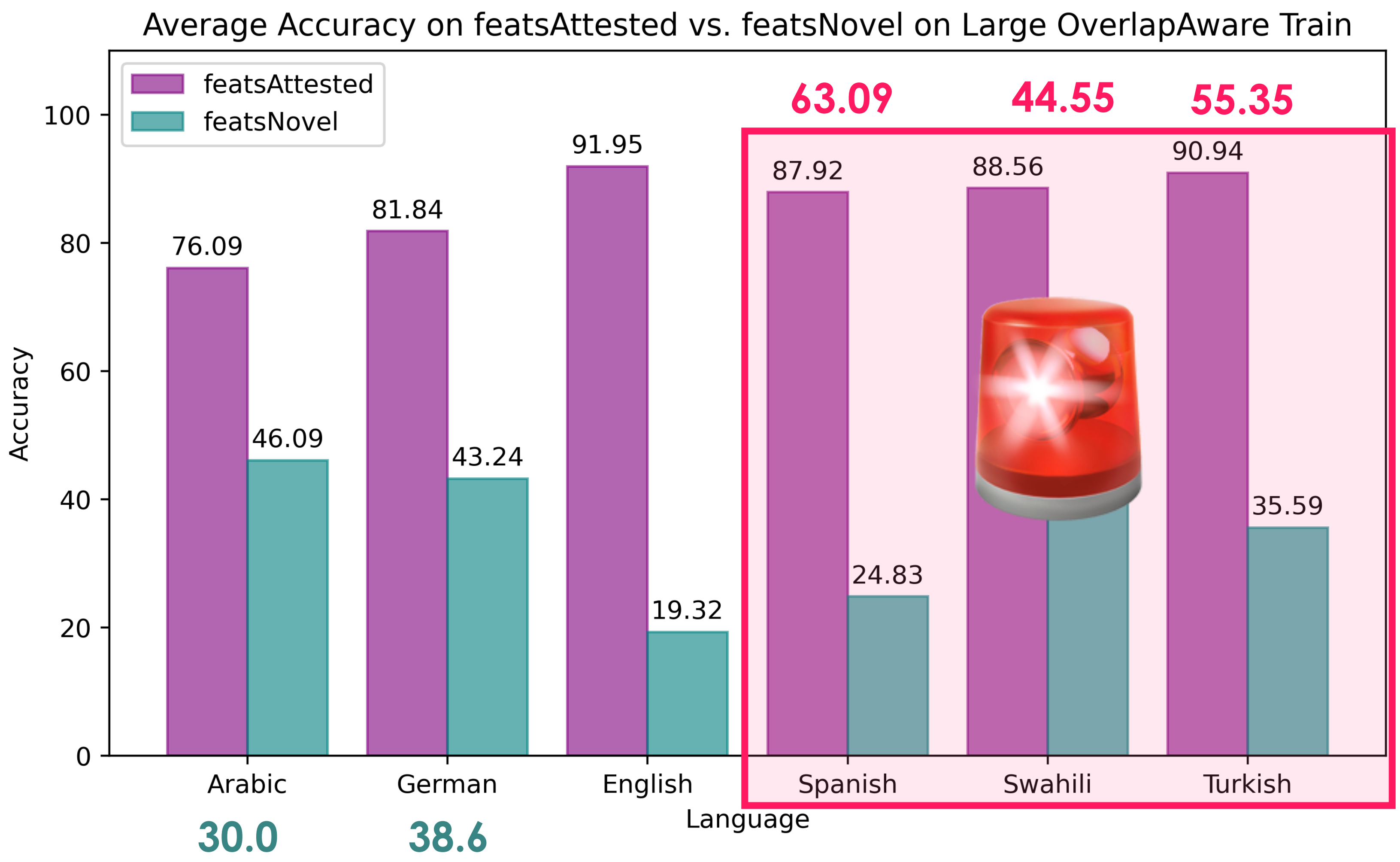


Is feature generalization realistic?

- Two factors at play: **paradigm size** and **agglutinativity**
  - **Large paradigm** → yes　　**small paradigm** → maybe not
  - **Highly agglutinative** → yes　　**highly fusional** → no

Swahili & Turkish
some Spanish

Is feature generalization realistic?



Outline

- **Background**
  - Defining the task
  - Input sparsity
  - Developmental trajectories & error patterns
- **Developmentally-grounded** evaluation
- Another approach: **Abduction of Tolerable Productivity**
- Revisiting the train-test overlap
- **Probing feature-based generalization**
- Conclusions

Kodner, Khalifa, & Payne (2023, EMNLP)

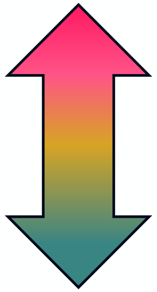
- Data splits to test specific components of **feature-based generalization in ANNs**
  - **Language–specific probes** for **feature-based generalizations that *should* be possible**
  - And some that shouldn't for comparison
  - Designing these probes requires **linguistic expertise**

Jordan Kodner

Salam Khalifa

Kodner, Khalifa, & Payne (2023, EMNLP)

- Data splits to test specific components of **feature-based generalization in ANNs**
- **3 languages:**
  - **English** (fusional)
  - **Spanish** (mixed)
  - **Swahili** (agglutinative)
  - **Orthography** & phonological **transcription**



nearly impossible

very possible

Jordan Kodner

Salam Khalifa

Kodner, Khalifa, & Payne (2023, EMNLP)

- Data splits to test specific components of **feature-based generalization in ANNs**
- **3 languages:** English, Spanish, Swahili
- **3 models:**
  - **CLUZH:** **character-level transducer** with **beam decoding**
  - **CHR-TRM:** **character-level transformer**
  - **ENC-DEC:** Kirov & Cotterell (2018) **encoder-decoder**

Jordan Kodner

Salam Khalifa

Wehrli et al. (2022); Wu et al. (2021); Cotterell et al. (2018)

Language-Specific Probes

- **BLIND:** language-independent **OVERLAPAWARE** sampling
  Verbs: **English** (en, fusional) - **Spanish** (es) - **Swahili** (sw, agglutinative)
- **PROBE:** random sampling testing specific **feature-based generalizations**

Agglutinative Feature Generalization Probes

es-FUT	suffixation
es-AGGL	suffixation (harder)
sw-1PL	prefixation
sw-NON3	prefixation (harder)
sw-FUT	string infixation
sw-PST	infixation w/ distractor

Conjugational class generalization probes

es-IR	suffixation
es-IRAR	suffixation (harder)

Fusional Feature Generalization Probes

en-NFIN	suffixation
en-PRS	suffixation
en-PRS3SG	suffixation
es-PSTPFV	suffixation
sw-PSTPFV	infix w/ distractor

Example Probe: es-FUT

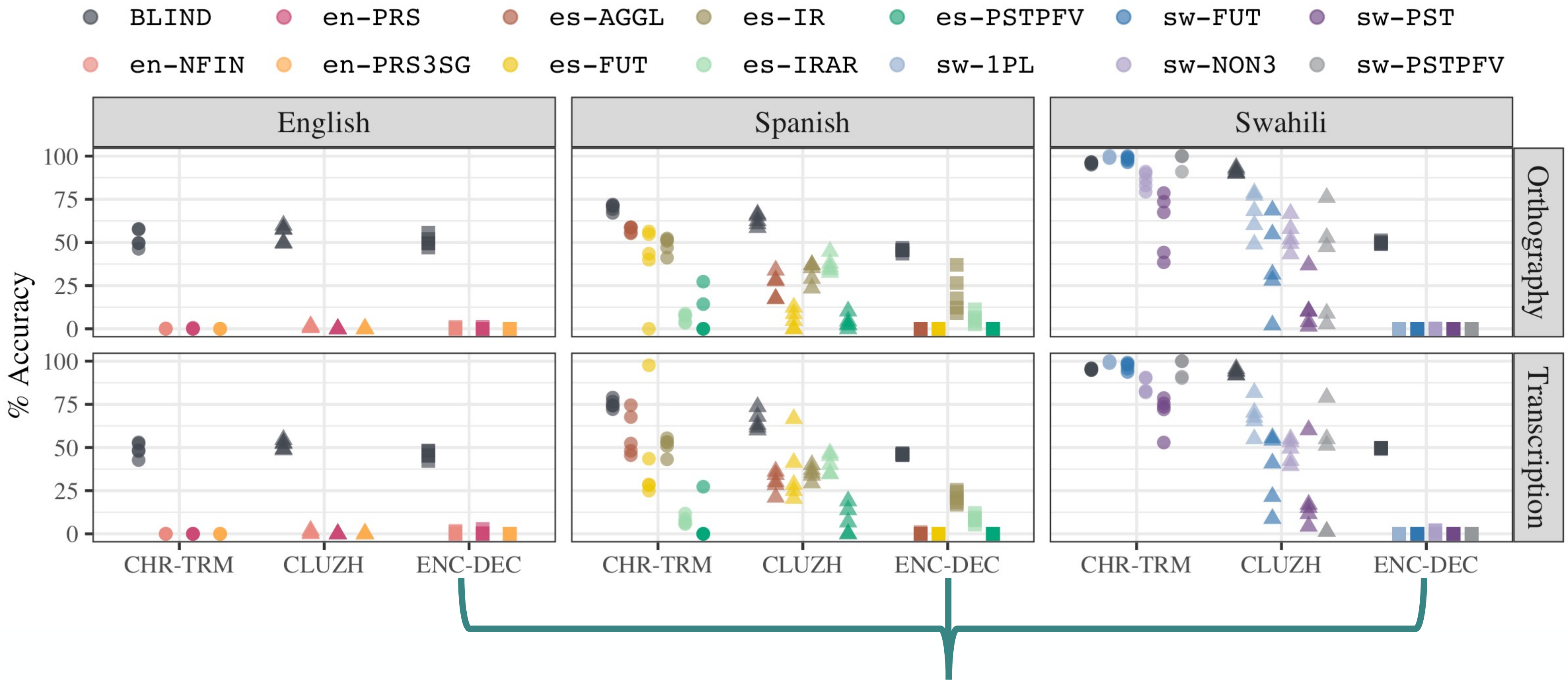
- The Spanish future tense is **agglutinative**:
  - Infinitive + **person-number marking**
  - **Person-number marking** matches most other tenses/moods

	SG	PL
1	INF+é	INF+á-mos
2;INFM	INF+á-s	INF+á-is
2;FORM	INF+á	
3	INF+á	INF+á-n

Example Probe: es-FUT

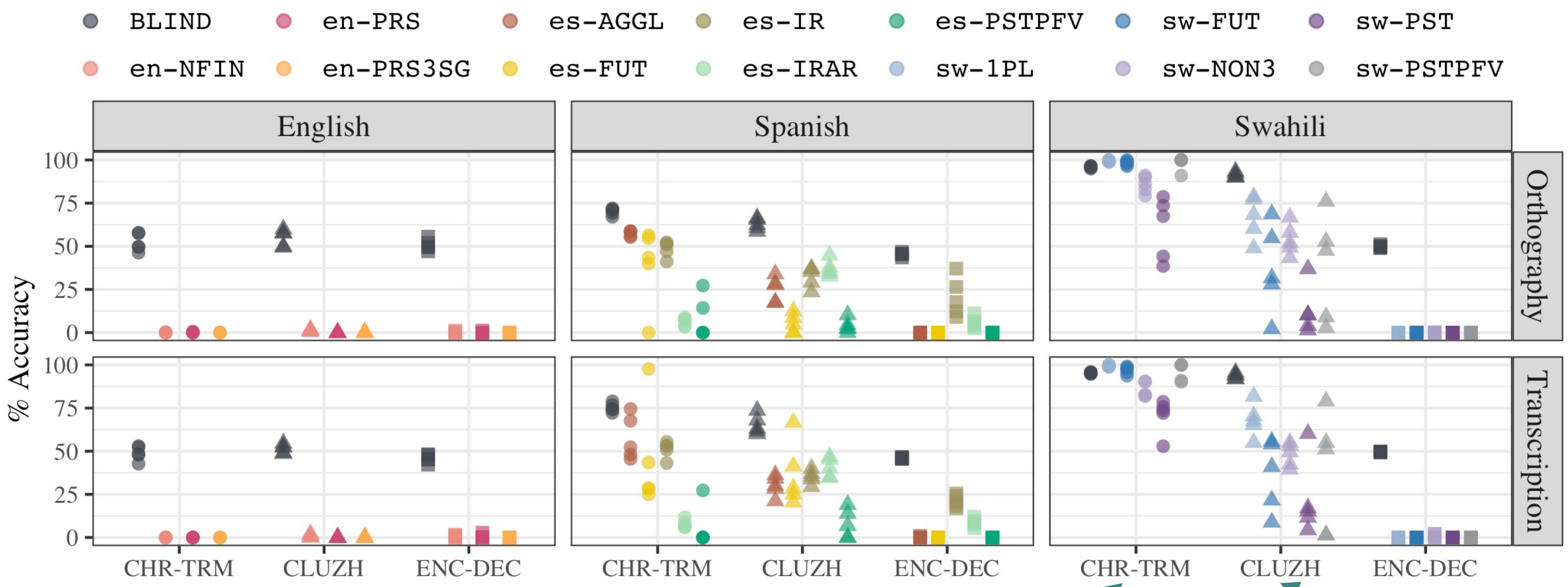
- The Spanish future tense is **agglutinative**
- For **5 random seeds**:
  - **5 of 7** person-number combinations containing **V;IND;FUT** are randomly withheld for test
  - **Train** sampling proceeds as normal **except for these features**
    - 1600 training + 400 ftune
  - **Test** sampling proceeds as normal
    - All triples that aren't relevant are **discarded from test**

Results: Language-Specific Probes



ENC-DEC only achieves meaningful performance on **es-IR** and **es-IRAR** generalize across conjugation classes but **not feature sets**

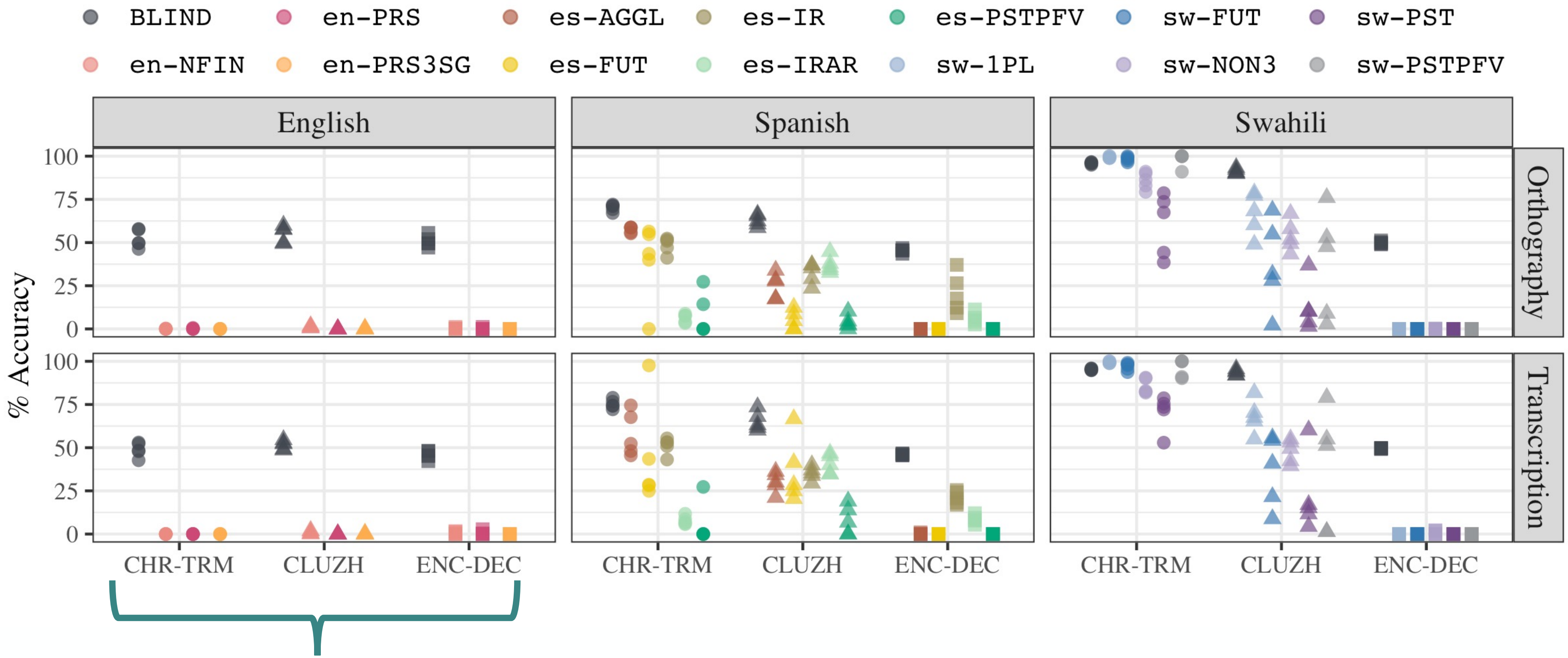
Results: Language-Specific Probes



CHR-TRM performs well on Swahili probes

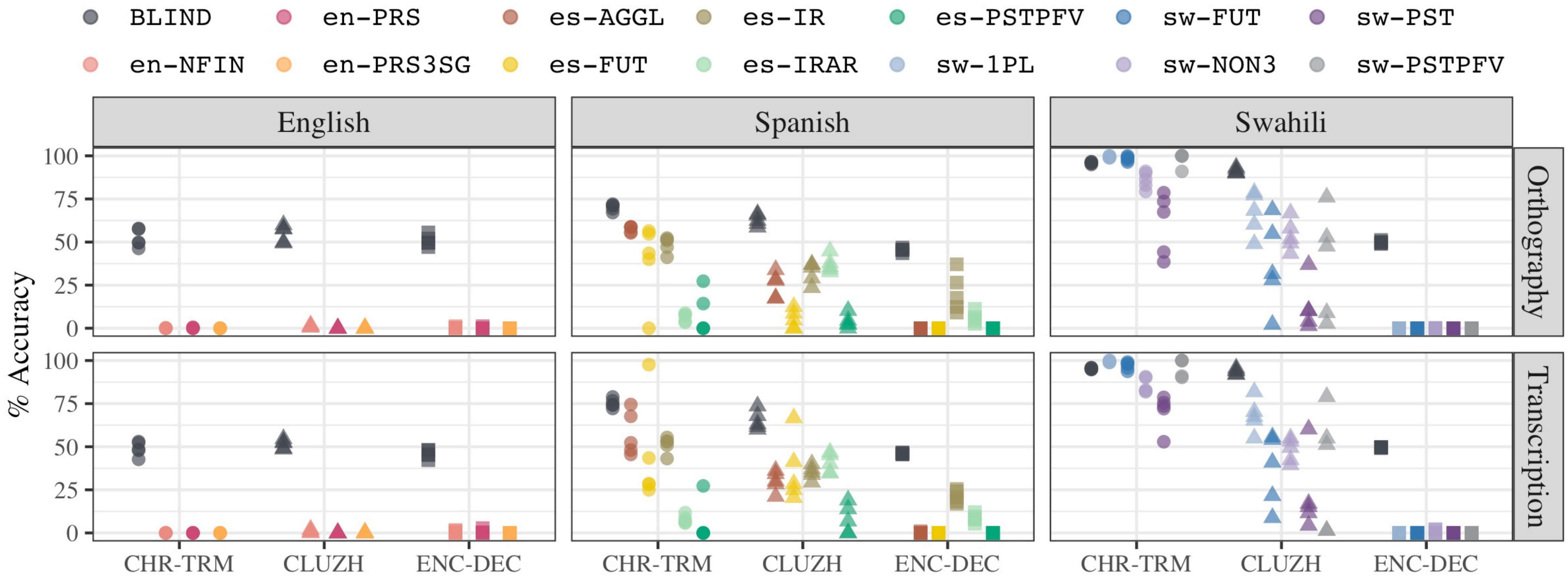
CLUZH shows high variability across seeds on Swahili probe splits

Results: Language-Specific Probes



English probe splits are **intentionally impossible**
Errors are insightful: no model output the bare lemma
All output primarily ***–ing, -ed, -es*** forms

Results: Language-Specific Probes



Systems succeed and fail on different probes and the types of errors they make reveal differing generalization strategies

Interim Summary

- **UNIFORM** and **WEIGHTED** sampling yield similar results
  - **WEIGHTED** is more cognitively-plausible
- Models tend to **generalize poorly to unseen feature sets**
  - Even when this **should be possible** in principle
  - Language-specific probes reveal **systems generalize differently**
- **Score ranges are high** across random seeds
  - Highlights importance of **evaluating on multiple seeds**

Outline

- **Background**
  - Defining the task
  - Input sparsity
  - Developmental trajectories & error patterns
- **Developmentally-grounded** evaluation
- Another approach: **Abduction of Tolerable Productivity**
- Revisiting the train-test overlap
- Probing feature-based generalization
- **Conclusions**

Conclusions

- **Morphological learning models** should be evaluated:
  - On realistically **sparse, skewed, input**
    Children learn from only a few hundred types!
  - On multiple **random splits**
    Performance varies greatly across splits!
  - On **language-specific probes** for **feature set overlap**
    These give specific, detailed insights into how models generalize!
  - Against **learning trajectories** and **error patterns**
    Should match with children's developmental patterns!

Conclusions

- When evaluated this way, **current ANNs fall short**
  - **Do not generalize to new feature sets** when it should be possible
  - **Error patterns** and **learning trajectories** don't match children's
- **BUT:** more thorough evaluation **helps us understand why!**
  - ANNs are prone to **over-irregularization**
  - Current ANNs **struggle to generalize across feature sets**
- **Rule-based models** may not have these shortcomings
  - ATP makes **human-like errors** and **exhibits developmental regression**
  - When trained on **plausible data** over **multiple splits**

Thank you!!

Caleb Belth
U. of Utah

Deniz Beser

Salam Khalifa

Jordan Kodner

Zoey Liu
U. of FL

Charles Yang
UPenn

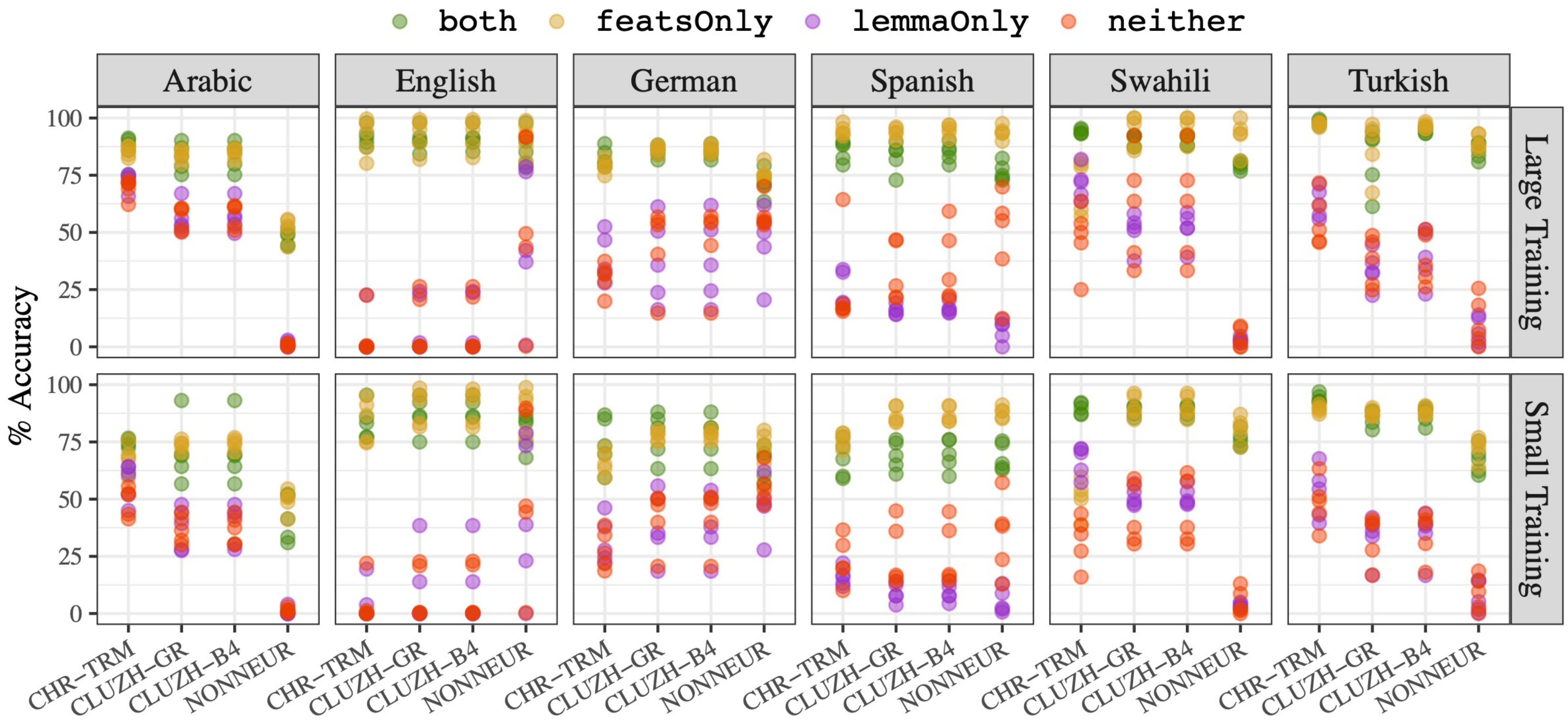
Special thanks to **Jeff Heinz, Scott Nelson, Mark Aronoff**, and **Bob Berwick.**

This work was supported by the **Institute for Advanced Computational Science Graduate Research Fellowship** and **NSF Graduate Research Fellowship**.

Extra Slides

Results: Effect of Feature Overlap

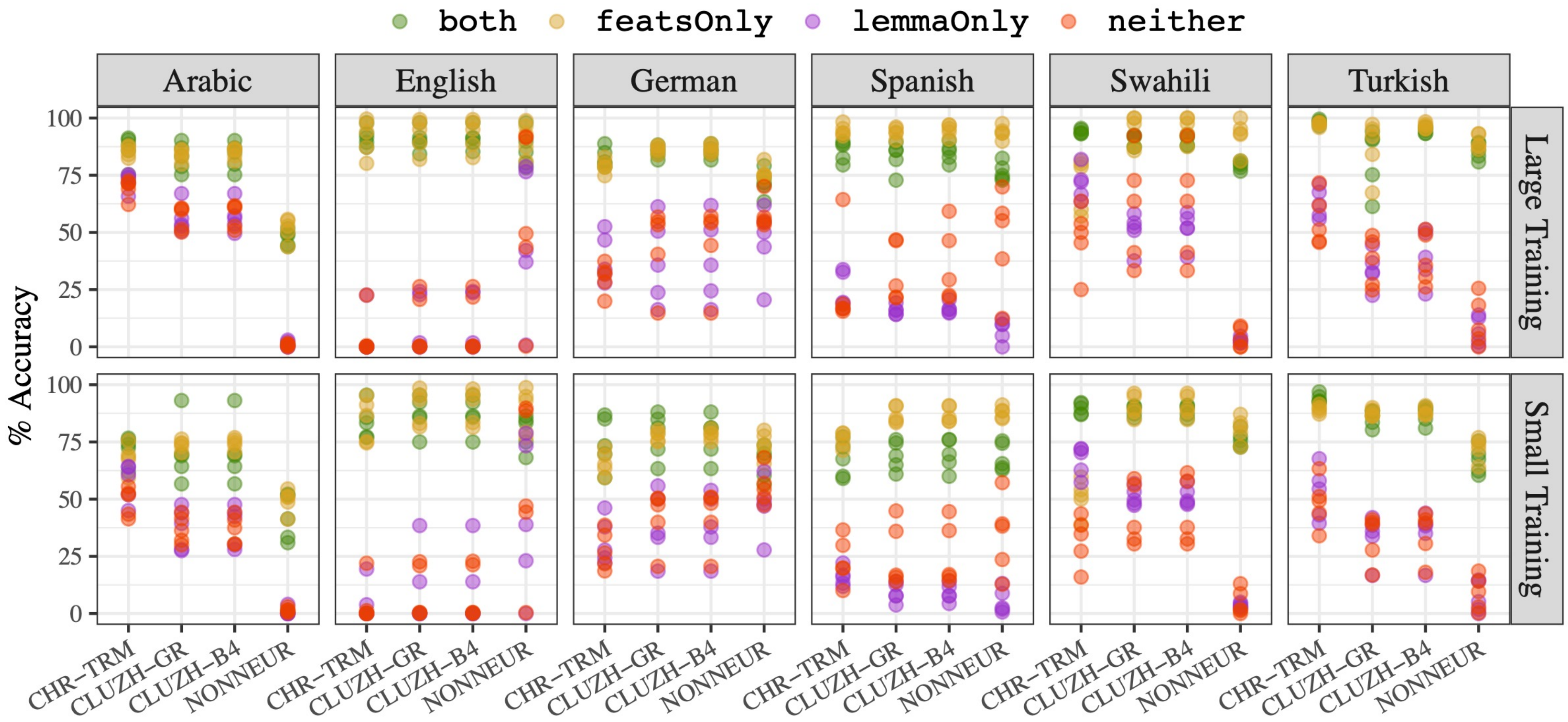
Correlation between proportion **FEATSATTESTED** & **accuracy**: **ρ = 0.68** (large) **ρ = 0.69** (small)



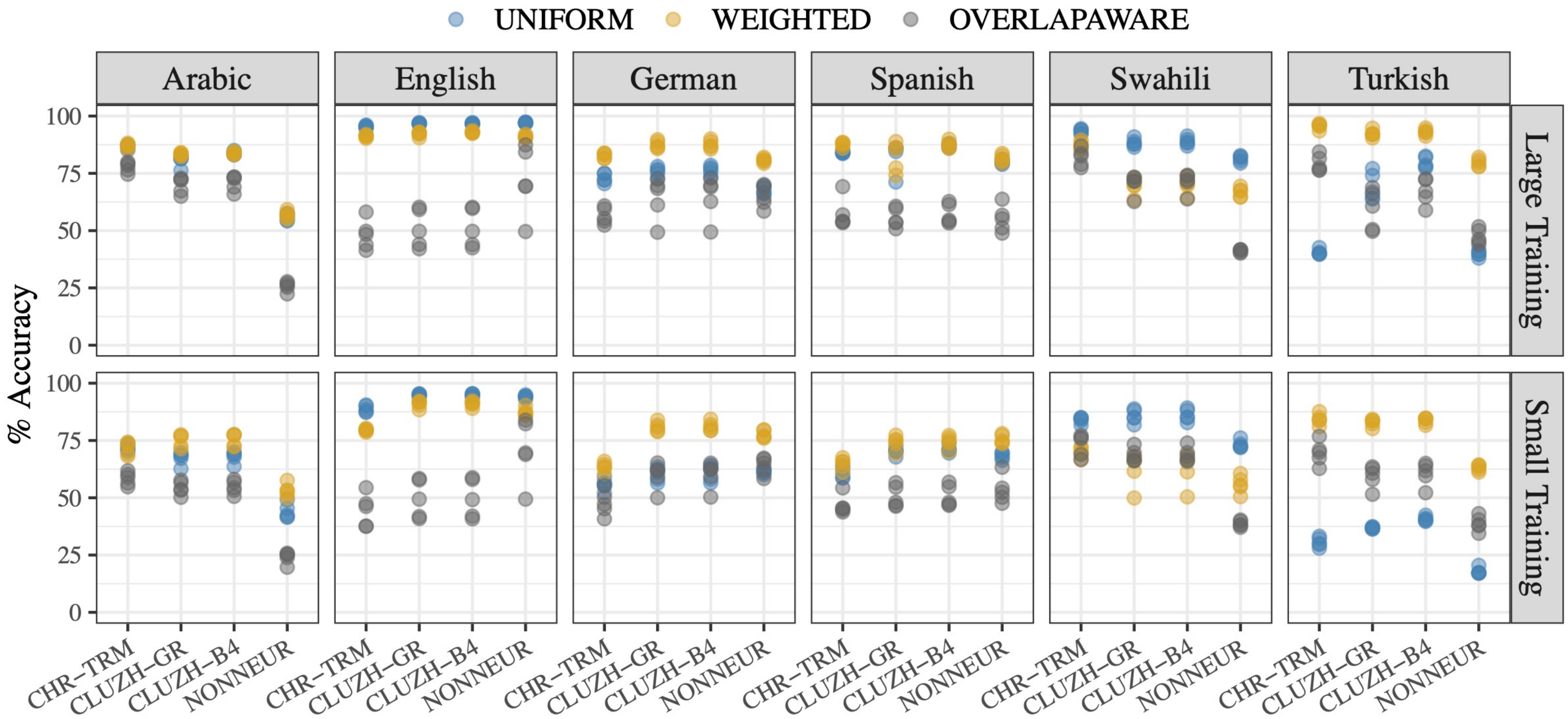
Results: Effect of Lemma Overlap

Correlation between proportion **LEMMAATTESTED** & **accuracy**:

ρ = -0.10 (large)
ρ = 0.10 (small)



Results: Effect of Sampling Strategy



WEIGHTED (83.75%, 74.22%) > **UNIFORM** (79.20%, 66.16%) > **OVERLAPAWARE** (60.86%, 55.37%)

Variability Across Random Seeds

- **Score range:** highest – lowest overall accuracy
- **Random seed variability:** standard deviation across seeds
- **OverlapAware** has **highest variability** despite consistent overlap
  - Not just feature set attestation, but ***which* feature sets are attested**

SmallTrain	Score Range	Random Seed Variability
Uniform	4.51	1.84
Weighted	6.33	2.57
OverlapAware	12.13	5.01
LargeTrain	**Score Range**	**Random Seed Variability**
Uniform	3.99	1.68
Weighted	4.08	1.66
OverlapAware	13.06	5.5

Is feature generalization realistic?

- Two factors at play: **paradigm size** and **agglutinativity**
  - **Large paradigm** → yes — **small paradigm** → maybe not
  - **Highly agglutinative** → yes — **highly fusional** → no

Swahili & Turkish
some Spanish

Turkish *guakamole* 🥑

Feature Set	Inflected Form
N;ACC;SG	?
N;ACC;PL	guakamoleleri
N;DAT;SG	guakamoleye
N;DAT;PL	?
N;ACC;PL;PSS3SG	guakamolelerini
N;DAT;PL;PSS3SG	guakamolelerine

Is feature generalization realistic?

- Two factors at play: **paradigm size** and **agglutinativity**
  - **Large paradigm** → yes   **small paradigm** → maybe not
  - **Highly agglutinative** → yes   **highly fusional** → no

Swahili & Turkish
some Spanish

Turkish *guakamole* 🥑

Feature Set	**Inflected Form**
N;Acc;Sg	?
N;Acc;**Pl**	guakamole**leri**
N;Dat;Sg	guakamoleye
N;Dat;**Pl**	?
N;Acc;**Pl**;Pss3Sg	guakamole**ler**ini
N;Dat;**Pl**;Pss3Sg	guakamole**ler**ine

Is feature generalization realistic?

- Two factors at play: **paradigm size** and **agglutinativity**
  - **Large paradigm** → yes — **small paradigm** → maybe not
  - **Highly agglutinative** → yes — **highly fusional** → no

Swahili & Turkish
some Spanish

Turkish *guakamole*

Feature Set	Inflected Form
N;**Acc**;Sg	?
N;**Acc**;**Pl**	guakamole**leri**
N;Dat;Sg	guakamoleye
N;Dat;**Pl**	?
N;**Acc**;**Pl**;Pss3Sg	guakamole**lerini**
N;Dat;**Pl**;Pss3Sg	guakamole**ler**ine

Is feature generalization realistic?

- Two factors at play: **paradigm size** and **agglutinativity**
  - **Large paradigm** → yes | **small paradigm** → maybe not
  - **Highly agglutinative** → yes | **highly fusional** → no

$\underbrace{\hspace{6cm}}_{}$

Swahili & Turkish
some Spanish

Turkish *guakamole* 🥑

Feature Set	Inflected Form
N;Acc;Sg	?
N;Acc;Pl	guakamoleleri
N;Dat;Sg	guakamoleye
N;Dat;Pl	?
N;Acc;Pl;Pss3Sg	guakamolelerini
N;Dat;Pl;Pss3Sg	guakamolelerine

Is feature generalization realistic?

- Two factors at play: **paradigm size** and **agglutinativity**
  - **Large paradigm** → yes — **small paradigm** → maybe not
  - **Highly agglutinative** → yes — **highly fusional** → no

Swahili & Turkish
some Spanish

Turkish *guakamole* 🥑

Feature Set	Inflected Form
N;Acc;Sg	?
N;Acc;Pl	guakamoleleri
N;Dat;Sg	guakamoleye
N;Dat;Pl	?
N;Acc;Pl;Pss3Sg	guakamolelerini
N;Dat;Pl;Pss3Sg	guakamolelerine

Is feature generalization realistic?

- Two factors at play: **paradigm size** and **agglutinativity**
  - **Large paradigm** → yes
  - **Highly agglutinative** → yes

$\underbrace{\qquad\qquad\qquad\qquad}_{}$

Swahili & Turkish
some Spanish

small paradigm → maybe not
highly fusional → no

Turkish *guakamole* 🥑

Feature Set	Inflected Form
N;ACC;SG	guakamole**yi**
N;ACC;PL	uakamole**leri**
N;DAT;SG	akamole**ye**
N;DAT;PL	kamole**lere**
N;ACC;PL;PSS3SG	guakamole**lerini**
N;DAT;PL;PSS3SG	guakamole**lerine**

🥳

The Past Tense Debate

- **Rumelhart & McClelland (1986):** ***single-route, connectionist*** model can:
  - Exhibit ***developmental regression***
  - Exhibit ***overregularization***

  ∴ **Rule-like behavior**

- **Pinker & Prince (1988):** actually…

  🚨 ***Developmental regression*** = artifact of training data
  - First trained on ***80% irregulars***
  - Then trained on ***80% regulars***

  🚨 Exhibits ***over-irregularization***
  - ***sip-sept, type-typeded, mail-membled***

  ∴ **No rule-like behavior**

Background: The Past Tense Debate Revisited

- **Kirov & Cotterell (2018):** encoder-decoder RNNs can overcome empirical limitations
  - Near ***100% test accuracy***
  - Learn ***several classes at once***
  - Trained on ***developmentally-representative*** data
  - Main errors = ***overregularizations***

- **Corkery et al (2019):** ED model **still fails empirically!**

  🚨 Predictions **don't match well with humans** on nonce English past tense forms
  - **Still over-irregularizes!**

  🚨 Massive **variability in model rankings** between seeds
  - **Correlation with human ratings** also varies massively

Background: The Past Tense Debate Revisited

- **Kirov & Cotterell (2018):** encoder-decoder RNNs can overcome empirical limitations
  - Near ***100% test accuracy***
  - Learn ***several classes at once***
  - Trained on ***developmentally-representative*** data
  - Main errors = ***overregularizations***

No ***developmental regression!***

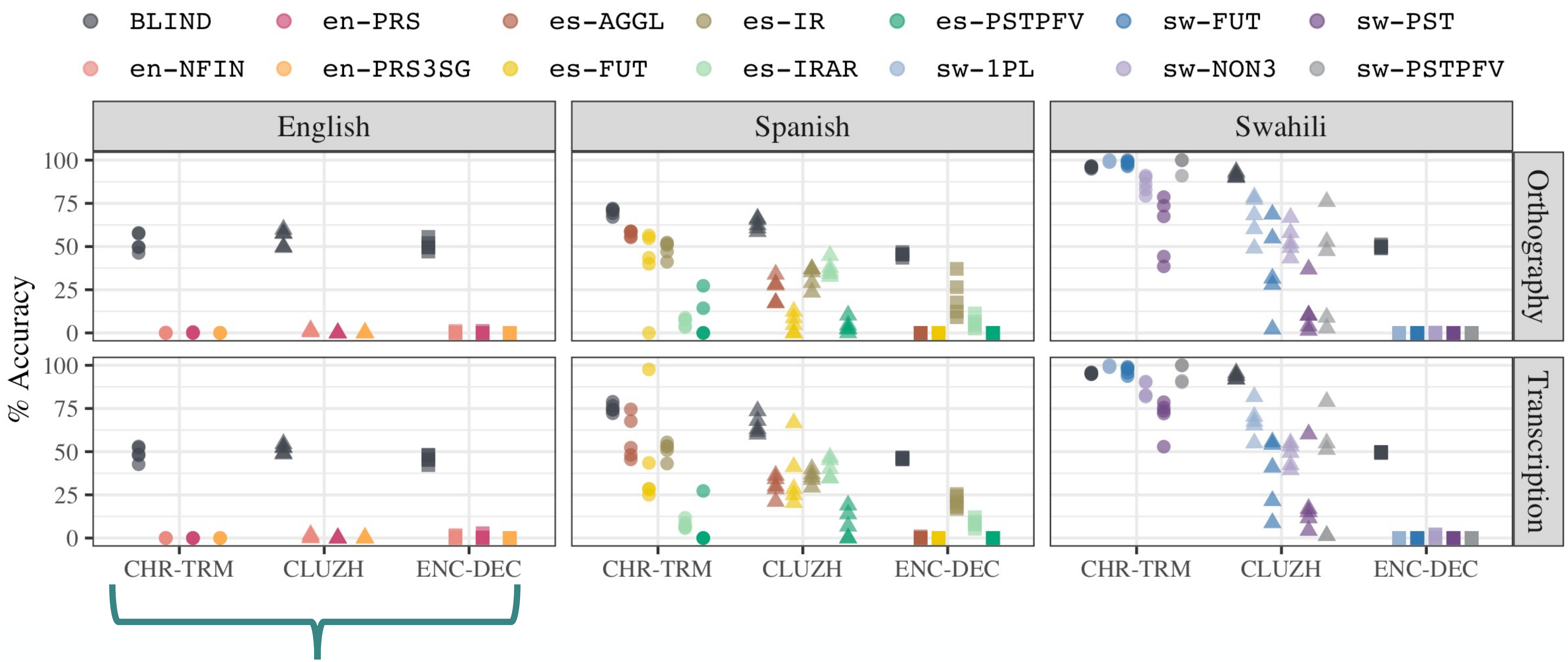
Trained on ***>3500 verbs in their full paradigm***

- Children know **< 350 verbs** at 3;0
- Would need to see **> 15k lemmas** to see 3,500 in complete paradigm

German Noun Plurals: We *really* aren't there

- **Marcus et al (1995):** NNs overapply the ***most common process*** rather than the ***default***
  - **German:** most common ≠ default
- **McCurdy et al (2020a):** Train on German noun plurals & **test on nonce words**
  - Model predictions ***don't match well with human predictions***
  - ***Overproduction of frequent*** affixes rather than default
- **McCurdy et al (2020b):** Model **uses gender** as main cue, humans **use phonology**

Results: Language-Specific Probes



On **en-PRS**, **CHR-TRM** and **CLUZH** both output primarily *–ing* or **–es**, showing generalization of **PRS** from **PRS;3SG** and **PRS;PRS.PTCP**