

Beyond Accuracy: Evaluating the Cognitive Plausibility of Neural Models of Morphological Inflection

Sarah Brogden Payne

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



Stony Brook
University



University of Florida Computational Linguistics Lab

October 13th, 2026

Morphological Inflection

- Patterns of **word formation** to express **grammatical categories**
 - **English:** *walk*+**PAST** → *walked*
 - **Mandarin:** **3**+**PL** → *tāmen* ‘they’
 - **Hebrew:** √*ḥtl*+**DIM**+**SG**+**DEF** → *haḥataltul* ‘the kitty’
 - **Latin:** *amic*+**FEM**+**SG**+**GEN** → *amicae* ‘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 as an NLP Task

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

swim	V;Pst	swam
eat	V;PRS;3;SG	eats
cat	N;PL	cats
...	...	...

- 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, feature set, inflected form)

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

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

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

Modeling Morphological Inflection

- Modeling gives insights into the **cognitive computations** underlying morphological learning
 - Major class: **Neural Models of Morphological Inflection (NNMIs)**
- **Past Tense Debate**
 - Early **connectionist** account (Rumelhart & McClelland 1986)
 - Several shortcomings
 - Competing **rule-based** accounts
- Recent advances in **ANN architectures**
 - Renewed interest in the **cognitive plausibility of NNMIs**

Modeling Morphological Learning: Solved?

- **Kirov & Cotterell (2018)**: encoder-decoder network can **overcome practical limitations** of older NNMs

- 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

Modeling Morphological Learning: Solved?

High accuracy doesn't necessarily entail **cognitive plausibility**

- Two main questions to look further:
 - Do models **generalize** like humans?
 - **Feature-based** generalization
 - **Lemma-based** generalization
 - Do models **make errors** like humans?
 - Overregularization
 - German noun plurals
 - Developmental regression

When we look at NNMs through these questions, we see that they **aren't cognitively plausible models of acquisition!**

Outline

High accuracy doesn't necessarily entail **cognitive plausibility**

- Two main questions to look further:

- **Do models generalize like humans?**

- **Feature-based** generalization

- **Lemma-based** generalization

- Do models **make errors** like humans?

- Overregularization

- German noun plurals

- Developmental regression

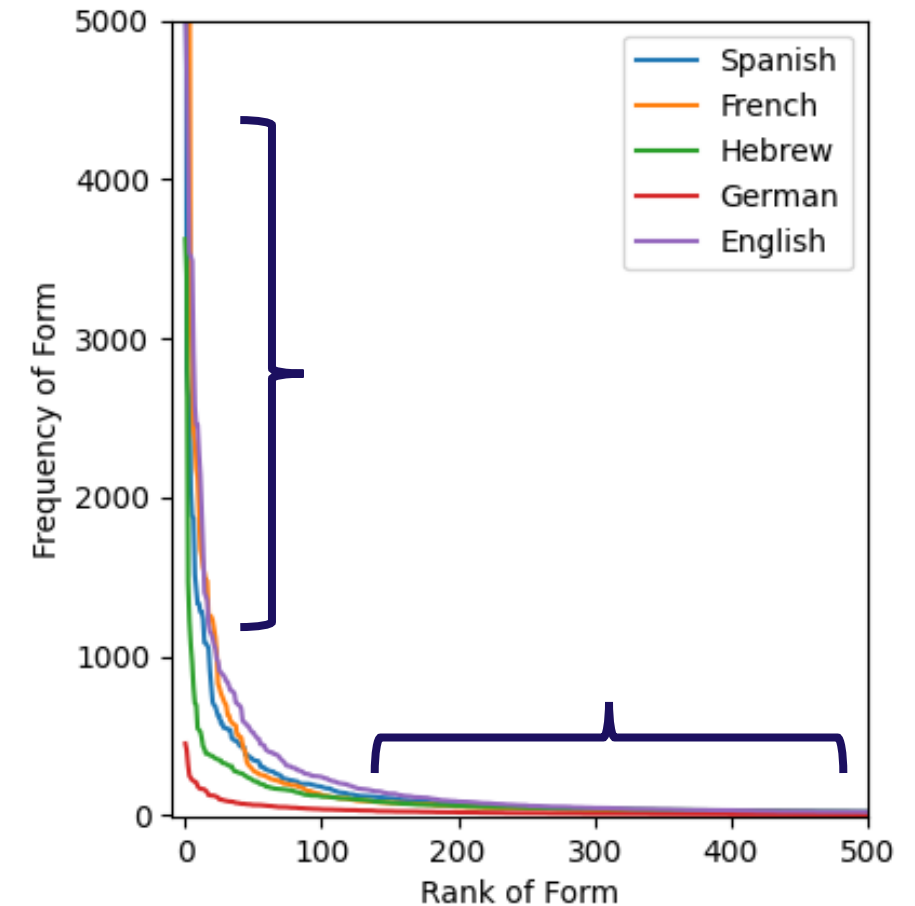
When we look at NNMs through these questions, we see that they **aren't cognitively plausible models of acquisition!**

Generalization: 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)

Generalization: Paradigm Saturation

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

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

	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)

Generalization: Paradigm Saturation

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

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

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

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

(Chan 2008, Lignos & Yang 2016)

Generalization: Paradigm Saturation

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

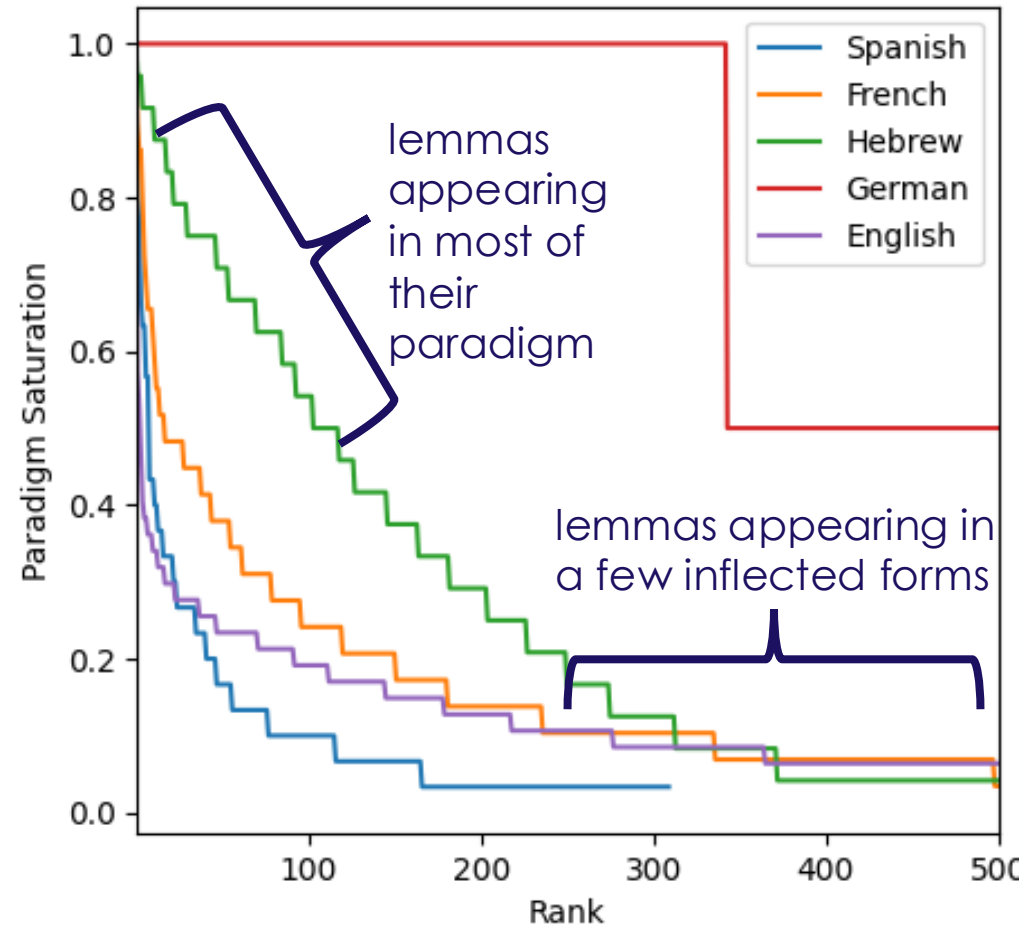
$$\text{saturation} = \frac{\# \text{ seen}}{\# \text{ possible}}$$

$$= \frac{7}{\# \text{ possible}}$$
$$= \frac{7}{26} \approx 27\%$$

	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

(Chan 2008, Lignos & Yang 2016)

Generalization: Paradigm Saturation



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

Generalization: Interim Summary

- Children must **generalize from small, sparse input**
 - From a **few hundred** of the **most-frequent** forms
 - To unseen **feature sets**, especially in **highly-inflected languages**
 - To **lemmas** not seen in that part of the paradigm
- Can NNMs generalize similarly:
 - To unseen **feature sets?**
Kodner et al. 2023 (ACL); Kodner et al. 2023 (EMNLP)
 - To unseen **lemmas?**
Payne and Kodner (2026)

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

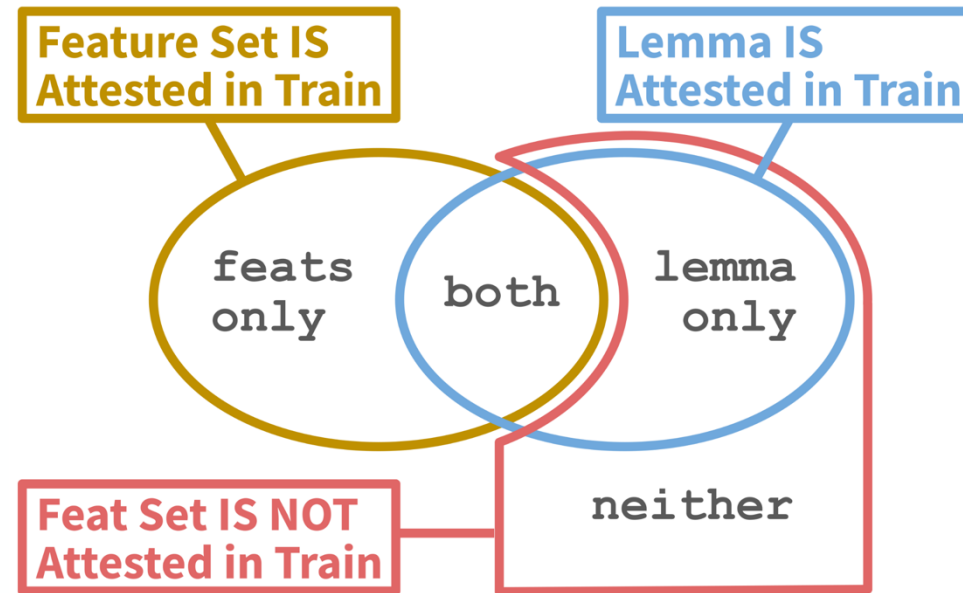
eat	eating	V;V.PTCP;PRS
run	ran	V;PST

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?

Feature-based Generalization

- **5 Languages:** German, English, Spanish, Swahili, Turkish
- **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**

Kodner et al. (2023, ACL)

Feature-based Generalization

- **5 Languages:** German, English, Spanish, Swahili, Turkish
- **4 Systems:** CLUZH-B4, CLUZH-GR, CHR-TRM, NONNEUR
- **3 Split Types:**
 - **UNIFORM:** partition UniMorph **uniformly at random**
 - **WEIGHTED:** partition at random weighted by **type frequency**
 - **OVERLAP AWARE:** enforce a maximum **50% proportion of FEATS ATTESTED**

Kodner et al. (2023, ACL)

Feature Overlap in Training

		SmallTrain	featsAttested	featsNovel	σ
400 train 100 ftune 1000 test	Uniform		80.33	19.67	19.5
	Weighted		90.44	9.56	11.1
	OverlapAware		48.81	51.19	0.98
		LargeTrain	featsAttested	featsNovel	σ
1600 train 400 ftune 1000 test	Uniform		96.17	3.83	5.55
	Weighted		95.36	4.64	7.28
	OverlapAware		49.92	50.08	0.17

Kodner et al. (2023, ACL)

Feature Overlap in Training

		SmallTrain	featsAttested	featsNovel	σ
400 train 100 ftune 1000 test	Uniform		80.33	19.67	19.5
	Weighted		90.44	9.56	11.1
	OverlapAware		48.81	51.19	0.98
		LargeTrain	featsAttested	featsNovel	σ
1600 train 400 ftune 1000 test	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

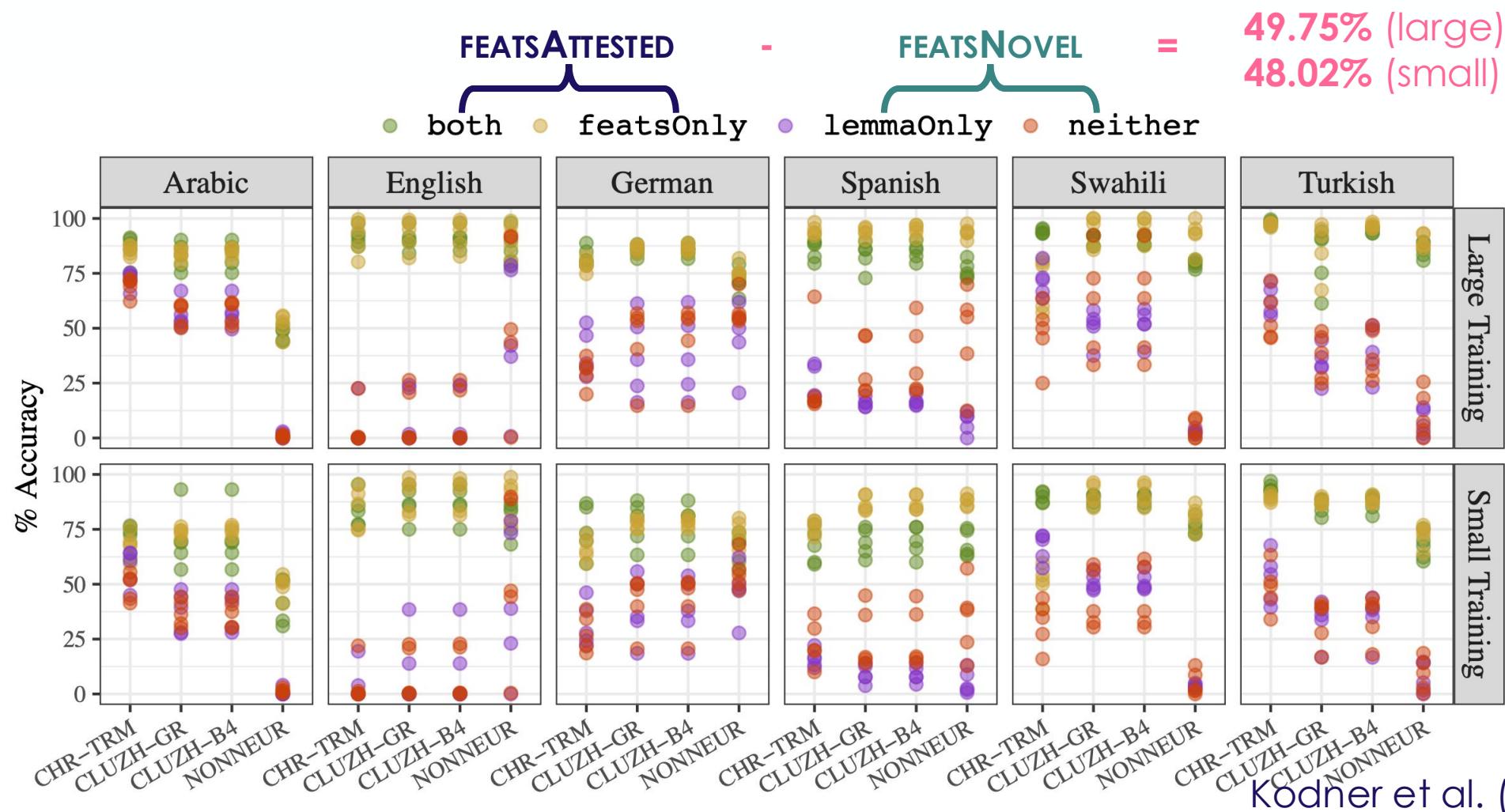
Kodner et al. (2023, ACL)

Feature Overlap in Training

		SmallTrain	featsAttested	featsNovel	σ
400 train 100 ftune 1000 test	Uniform		80.33	19.67	19.5
	Weighted		90.44	9.56	11.1
	OverlapAware		48.81	51.19	0.98
		LargeTrain	featsAttested	featsNovel	σ
1600 train 400 ftune 1000 test	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
Kodner et al. (2023, ACL)

Results: Effect of Feature Overlap

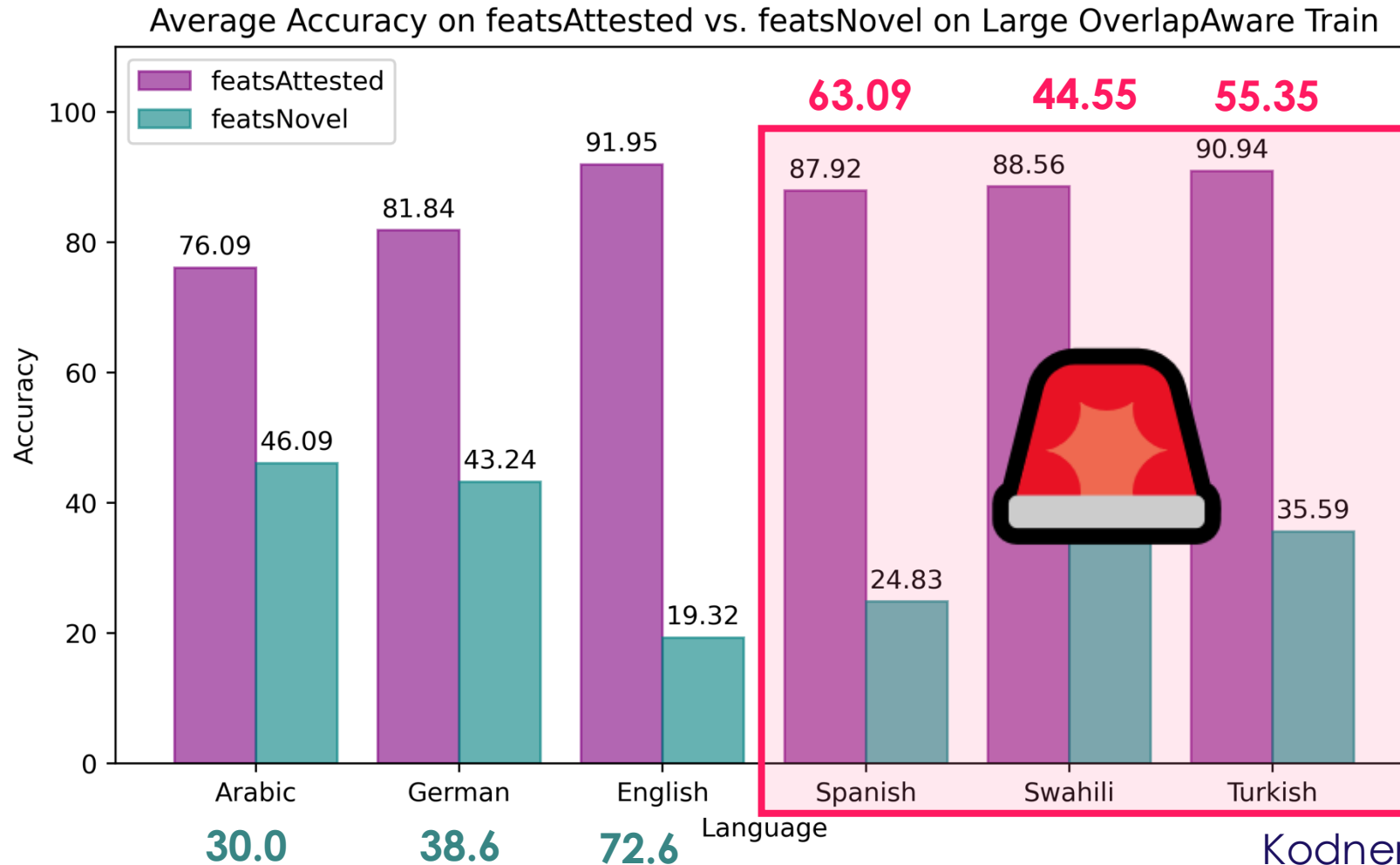


Is feature generalization realistic?

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

Kodner et al. (2023, ACL)

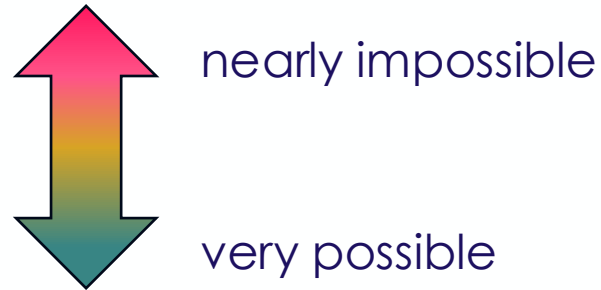
Is feature generalization realistic?



Kodner et al. (2023, ACL)

Language-Specific Probes

- Data splits to test specific components of **feature-based generalization in NNMs**
 - **Language-specific probes** for **feature-based generalizations that should be possible**
 - And some that shouldn't for comparison
- **3 languages:**
 - **English** (fusional)
 - **Spanish** (mixed)
 - **Swahili** (agglutinative)
 - **Orthography** & phonological **transcription**



Kodner et al. (2023, EMNLP)

Kodner, Khalifa, & Payne (2023, EMNLP)

- Data splits to test specific components of **feature-based generalization in NNMs**
- **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

Kodner et al. (2023, EMNLP)

Example Probe: es-FUT

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

present		SG	PL
	1	amo	ama-mos
	2;INFM	ama-s	amá-is
	2;FORM	ama	ama-n
	3	ama	ama-n

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

Kodner et al. (2023, EMNLP)

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**
 - Same person-number combinations can occur in other tenses/moods
 - 1600 training + 400 ftune
 - **Test** sampling proceeds as normal
 - All triples that aren't relevant are **discarded from test**

Kodner et al. (2023, EMNLP)

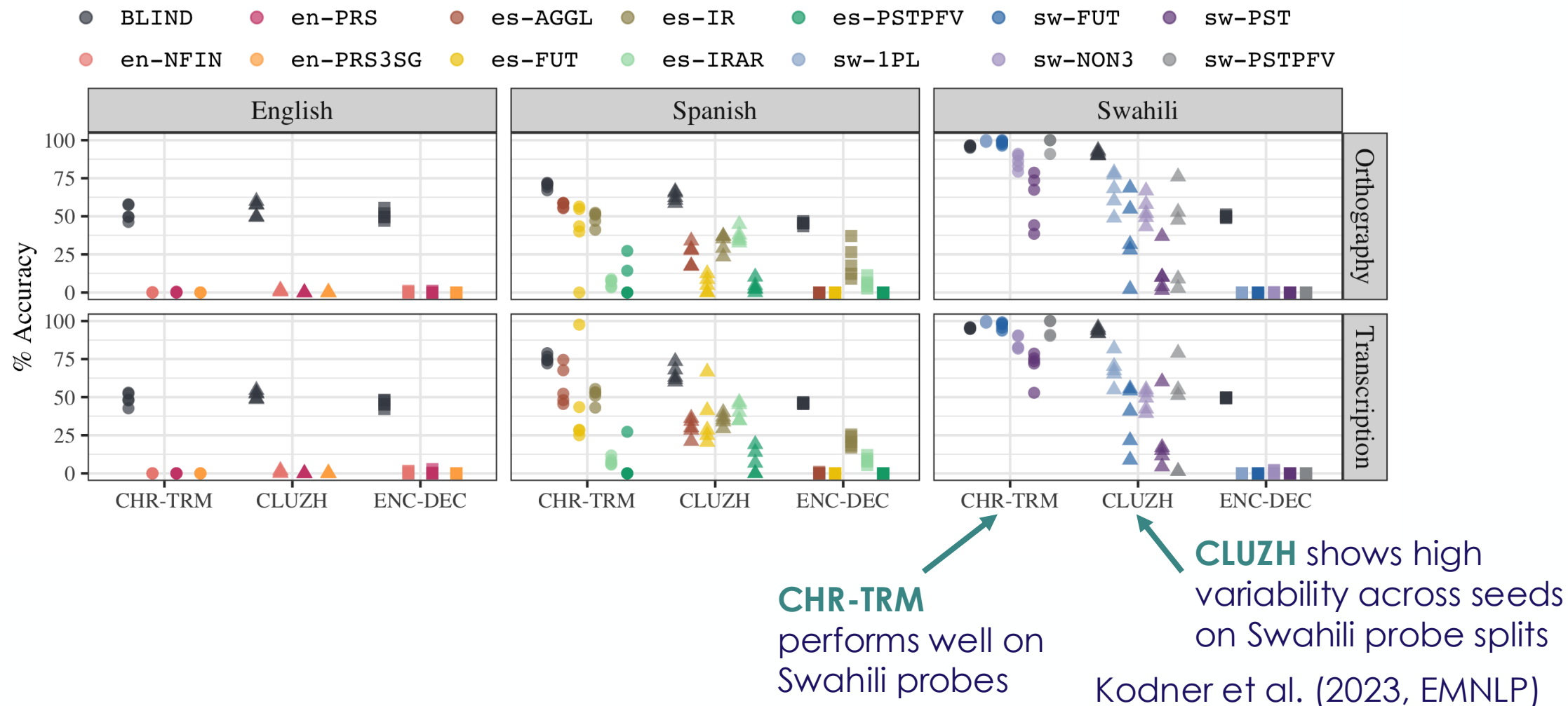
Results: Language-Specific Probes



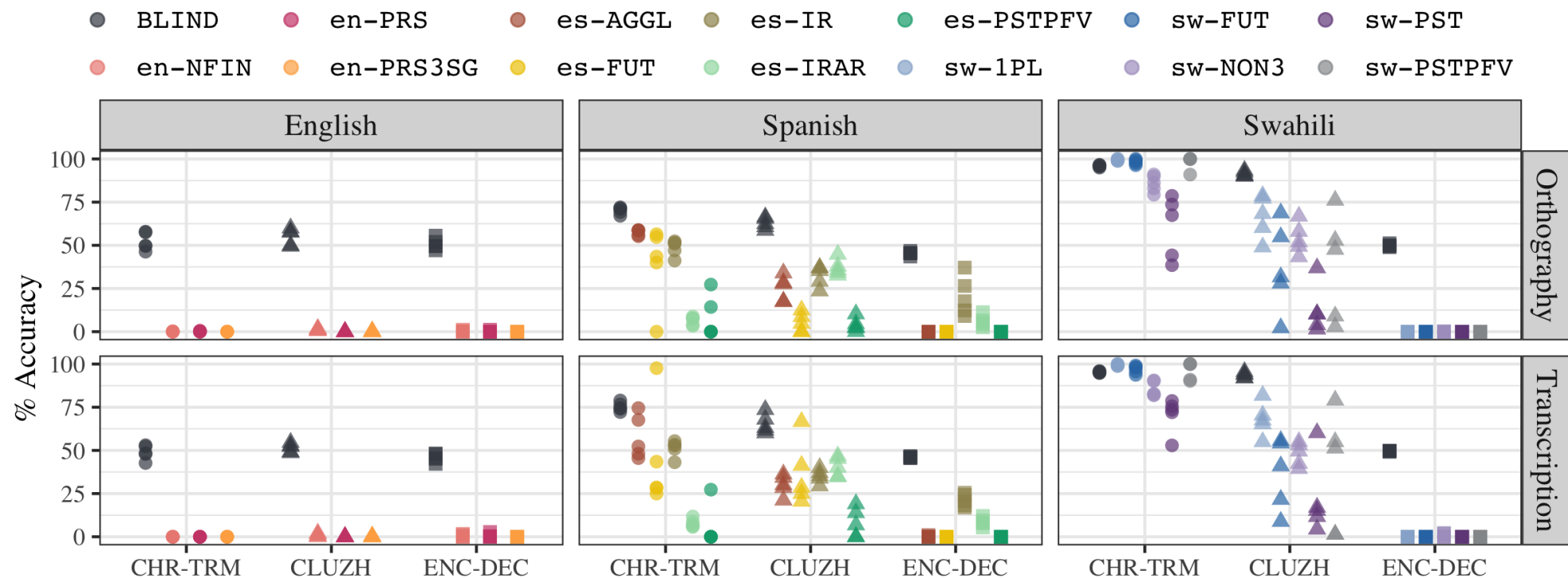
ENC-DEC only achieves meaningful performance on generalizations across conjugation classes but **not feature sets**

Kodner et al. (2023, EMNLP)

Results: Language-Specific Probes



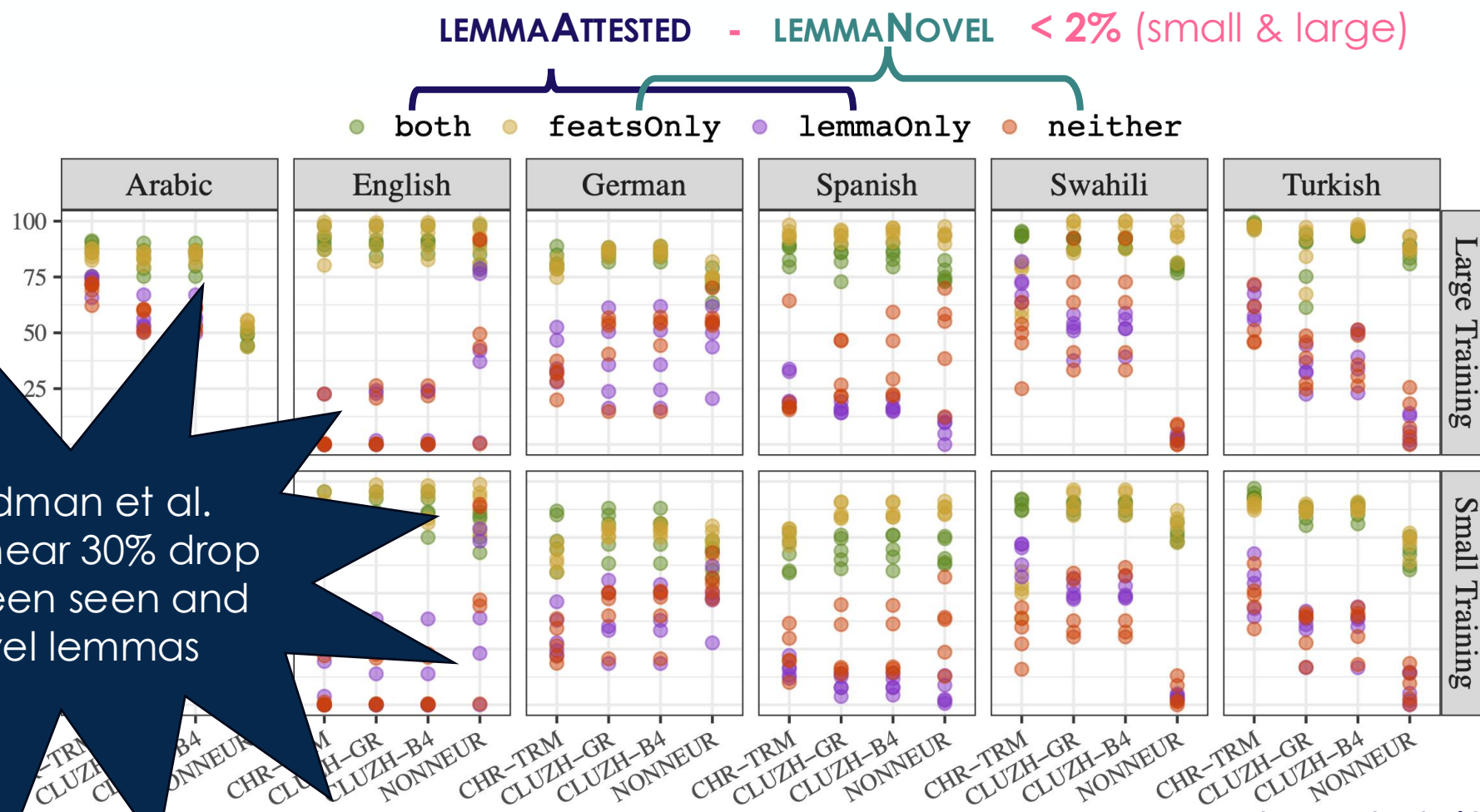
Results: Language-Specific Probes



Systems succeed and fail on different probes, but even systems that **succeed on simple ones fail on more complex but possible ones**

Kodner et al. (2023, EMNLP)

Results: Effect of Lemma Overlap



Kodner et al. (2023, EMNLP)

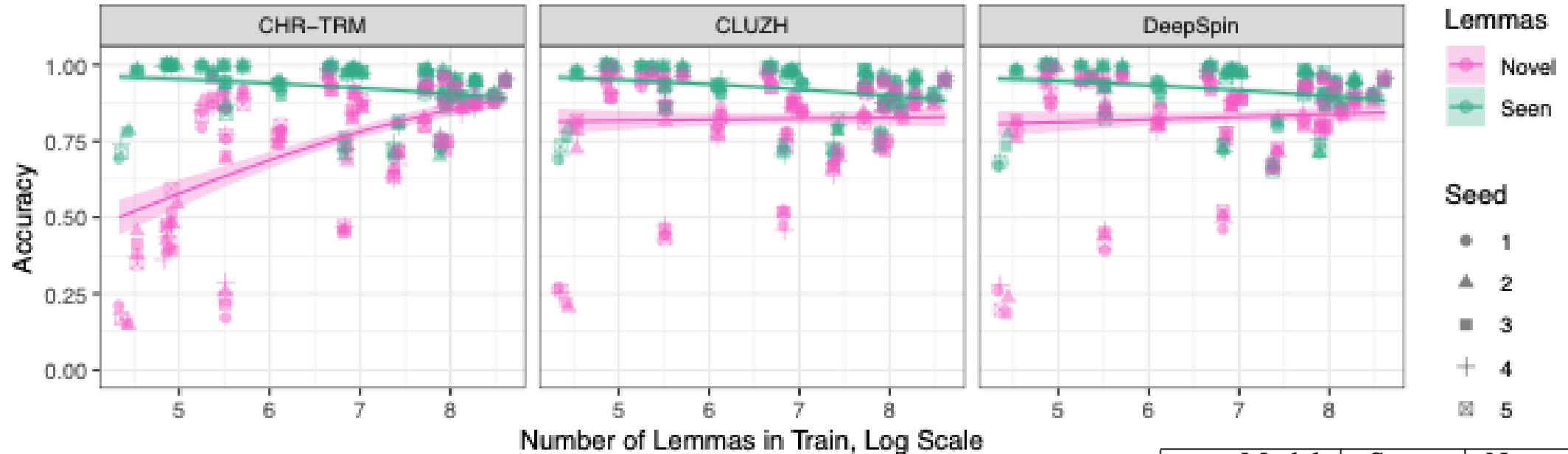
Lemma-based Generalization

- **36 Languages:** those in SIGMORPHON 2020 with at least 10k items
- **4 Systems:**
 - **CLUZH:** character-level **transducer** that significantly outperformed the 2022 SIGMORPHON baseline, with **beam decoding**
 - **CHR-TRM:** character-level **transformer** that was used as a baseline in 2021 and 2022 SIGMORPHON shared tasks
 - **DEEPSPIN:** multilingual RNN used by Goldman et al.
- **Data splits:** max 50% LEMMAATTESTED in test
 - 7,000 / 1,000 / 2,000 train/dev/test split
 - Near ceiling FEATSATTESTED

Payne & Kodner (2025, ACL)

Lemma-based Generalization

Accuracy on Seen and Novel Lemmas



Significant relationship between **number of train lemmas** and:

- **NOVEL** accuracy
 $\beta = 0.558$ for CHR-TRM, 0.148 for CLUZH, 0.190 for DEEPSPIN
- The **DROP** in accuracy from seen to novel lemmas
 $\beta = -0.524$ for CHR-TRM, -0.174 for CLUZH, -0.194 for DEEPSPIN

Model	SEEN	NOVEL	DROP
CHR-TRM	92.594	75.786	-16.808
CLUZH	92.138	83.221	-8.917
DEEPSPIN	91.946	83.828	-8.118
Macro-Avg	92.226	80.945	-11.281

Payne & Kodner (2025, ACL)

Lemma-based Generalization

- Why did we find an **11%** drop when Goldman et al. found **30%**?
 - **One issue:** we excluded data with less than 10k samples
 - **Another issue:** sampling entire paradigms



Interim Summary: Generalization

- Standard uniform sampling **ignores overlap**
- When we design splits to probe generalization, we find:
 - NNMs **struggle to generalize to novel features**, even in contexts where humans have to do so
 - NNMs **struggle to generalize to novel lemmas**, though less than we previously thought
 - Performance on novel lemmas **driven by number of train lemmas**
- Are all lemma-based generalizations **in-principle possible**?
 - No!
 - **Irregulars** – but humans make **predictable errors** on these

Outline

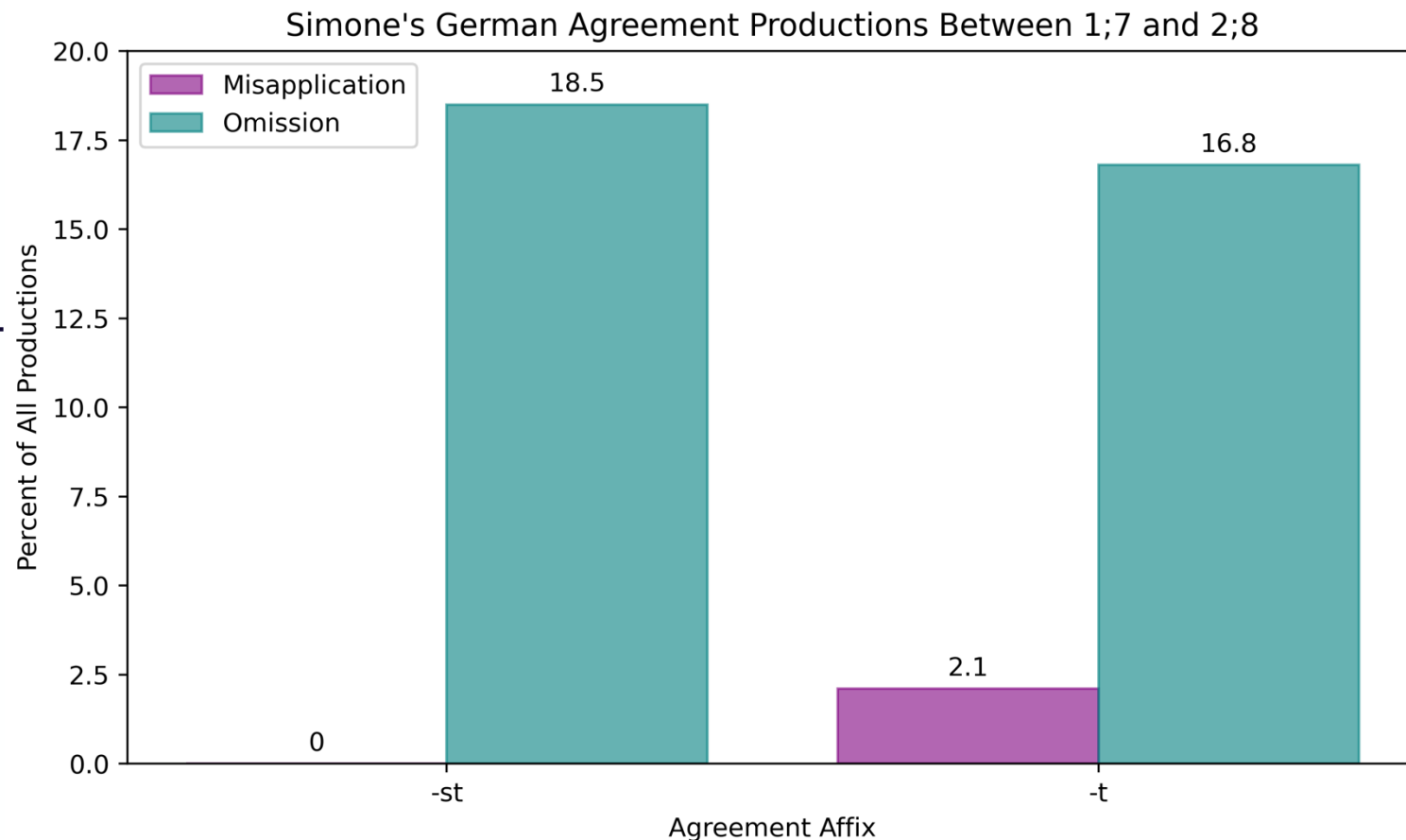
High accuracy doesn't necessarily entail **cognitive plausibility**

- Two main questions to look further:
 - Do models generalize like humans?
 - **Feature-based** generalization
 - **Lemma-based** generalization
 - **Do models make errors like humans?**
 - Overregularization
 - German noun plurals
 - Developmental regression

When we look at NNMs through these questions, we see that they **aren't cognitively plausible models of acquisition!**

Child Production Errors: Omissions

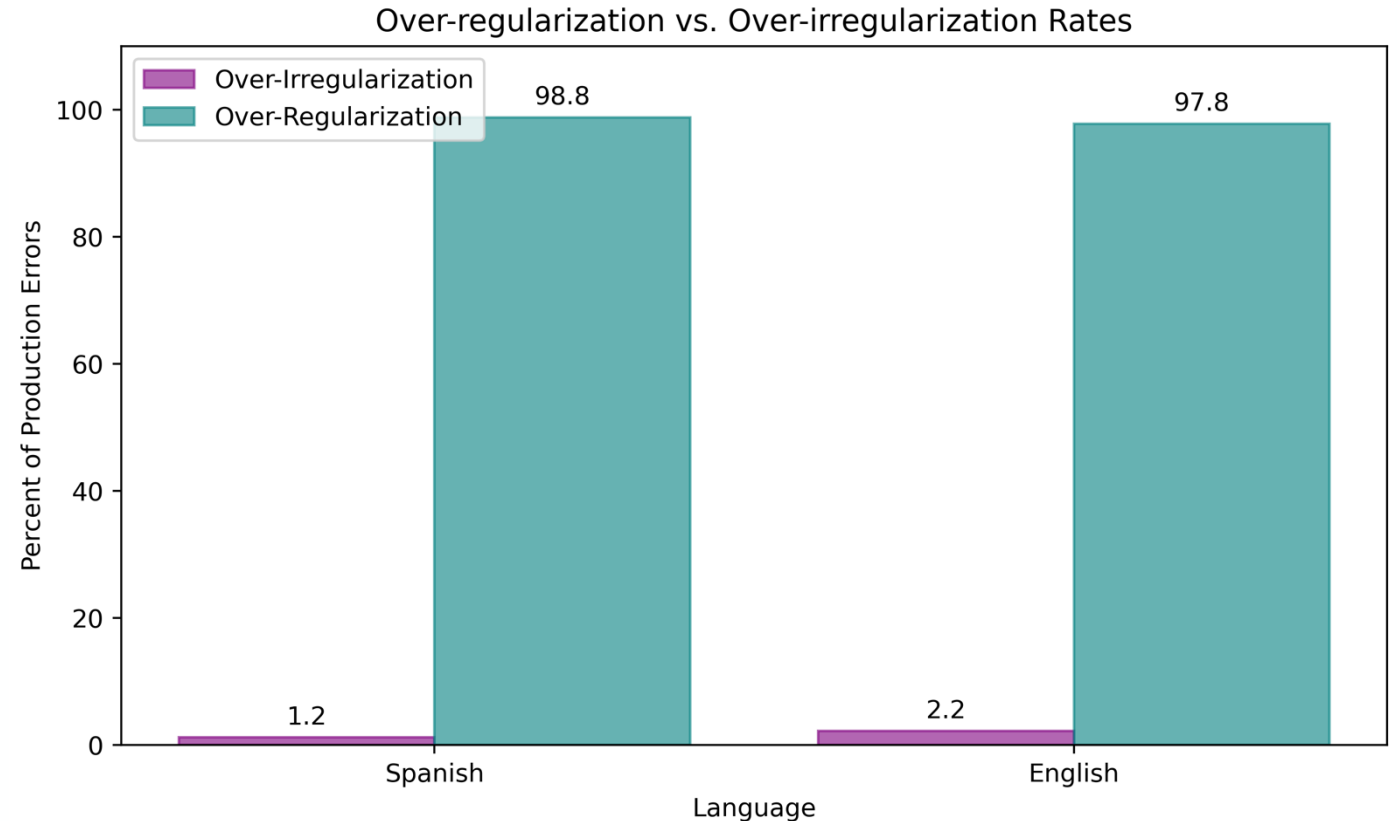
- **Omissions:** *Root Infinitives*
 - e.g. “Papa have it”
- **Substitutions:** incorrect overt affix
 - e.g. “I has it”



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

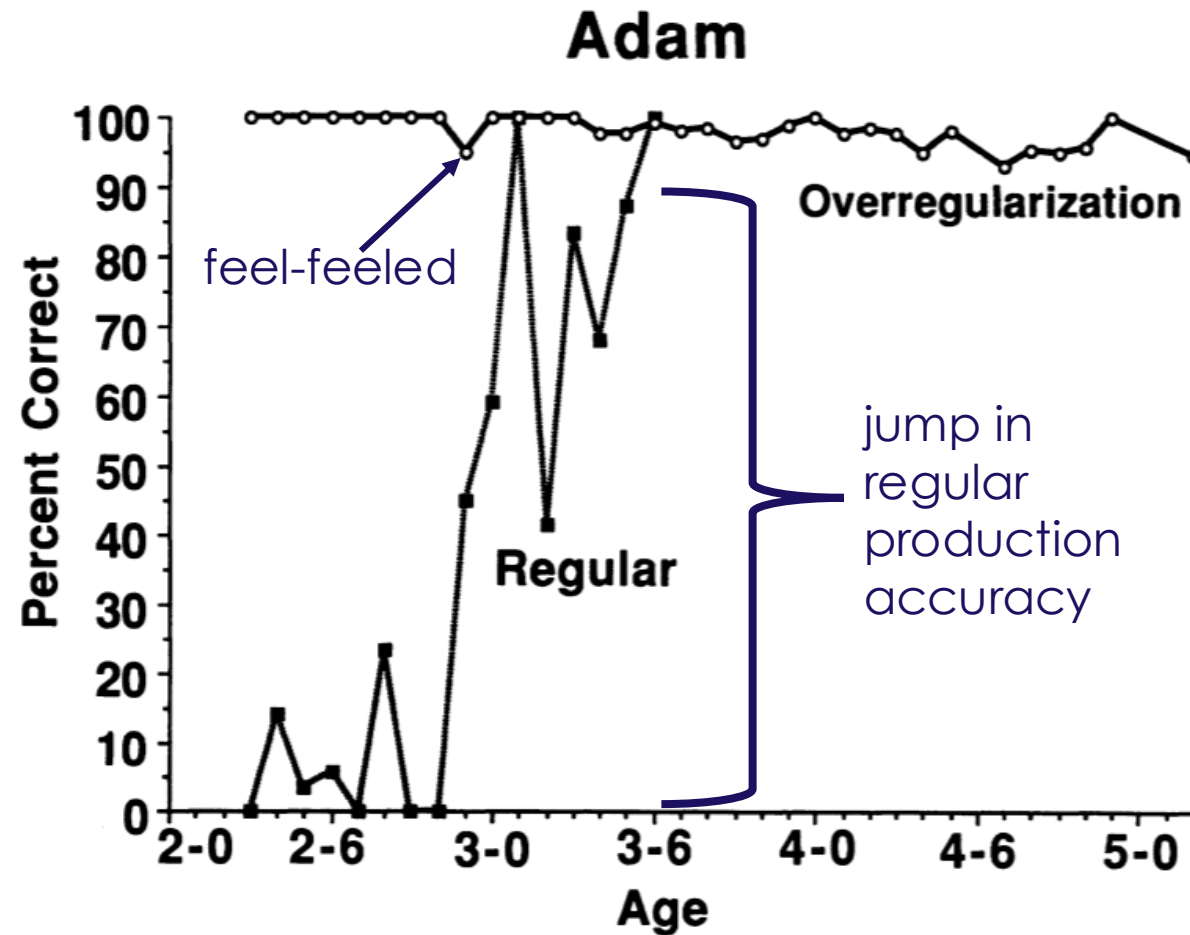
Child Production Errors: Over-regularization

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



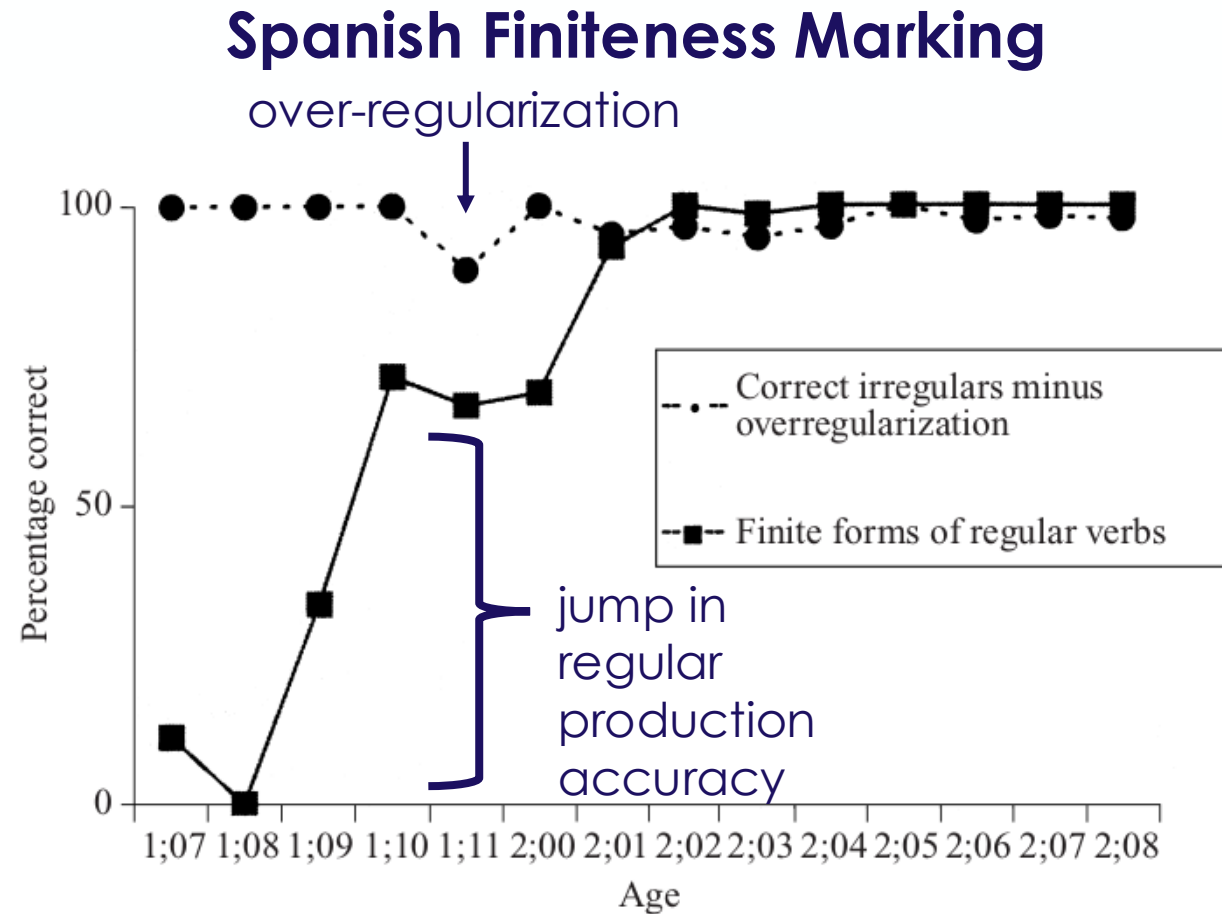
(Maslen et al 2004, Xu & Pinker 1995, Clahsen et al 2002)

Developmental Trajectories: Regression



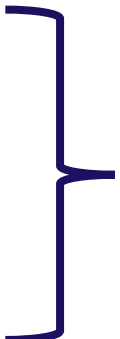
(from Marcus et al 1992)

Developmental Trajectories: Regression



(from Clahsen, Aveledo, and Roca 2002)

Probing Errors by NNMs

- Do NNMs 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**

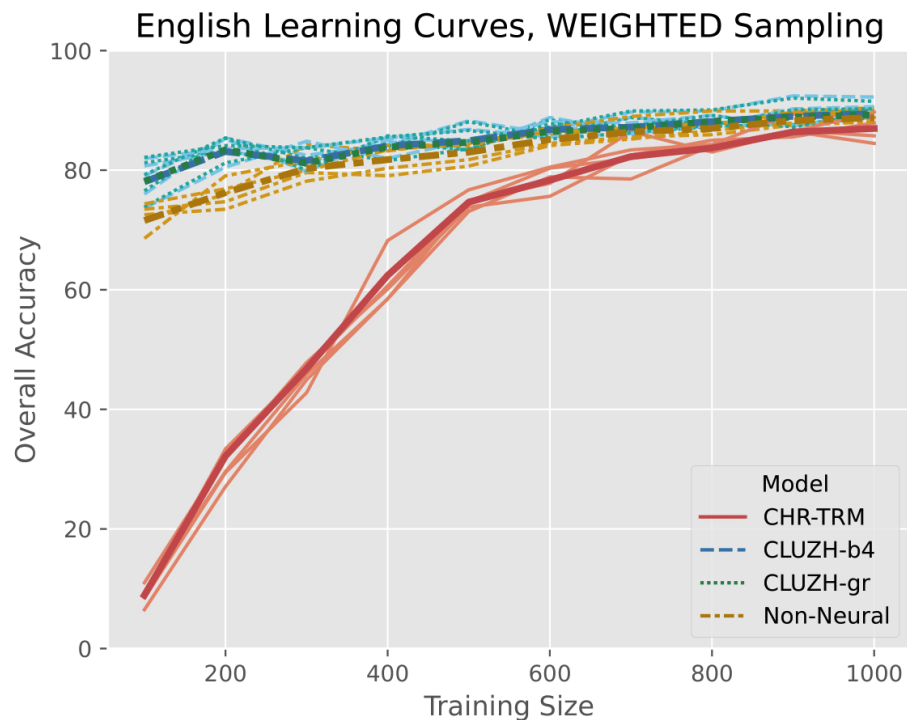
Kodner et al. (2023, CogSci)

Probing Errors by NNMs

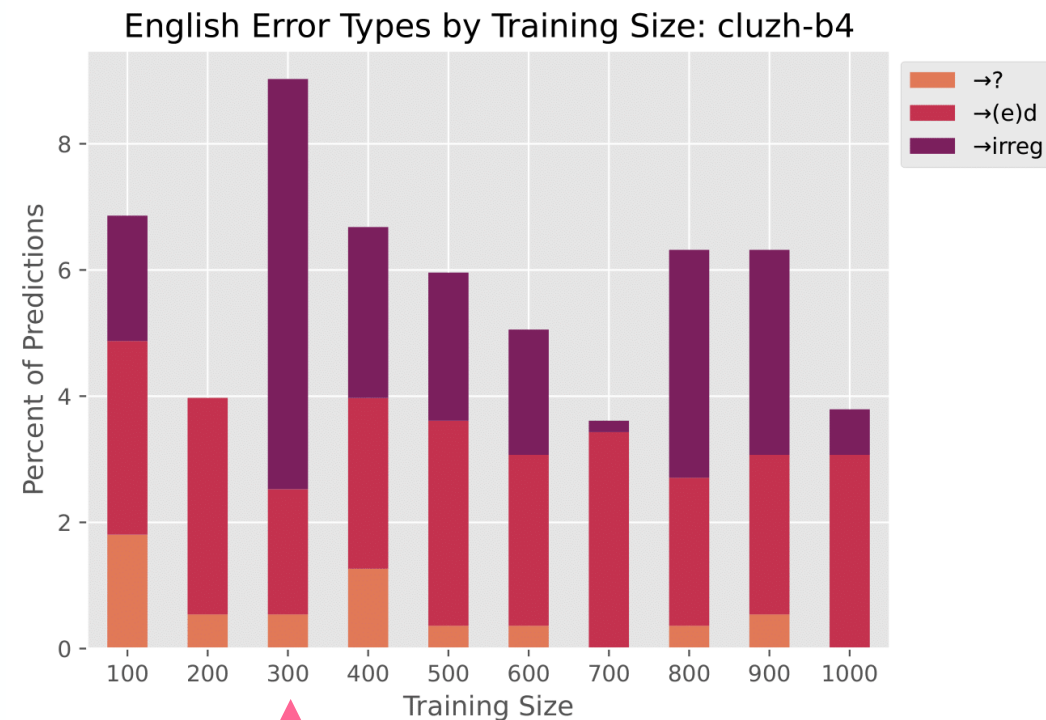
- Do NNMs match **developmental trajectories** and **error patterns** of children?
- Detailed analysis of **3 well-studied developmental phenomena**
- **4 models:**
 - **CLUZH-B4:** character-level **transducer** that significantly outperformed the 2022 SIGMORPHON baseline, with **beam decoding**
 - **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**

Kodner et al. (2023, CogSci)

Model Results: English Past Tense



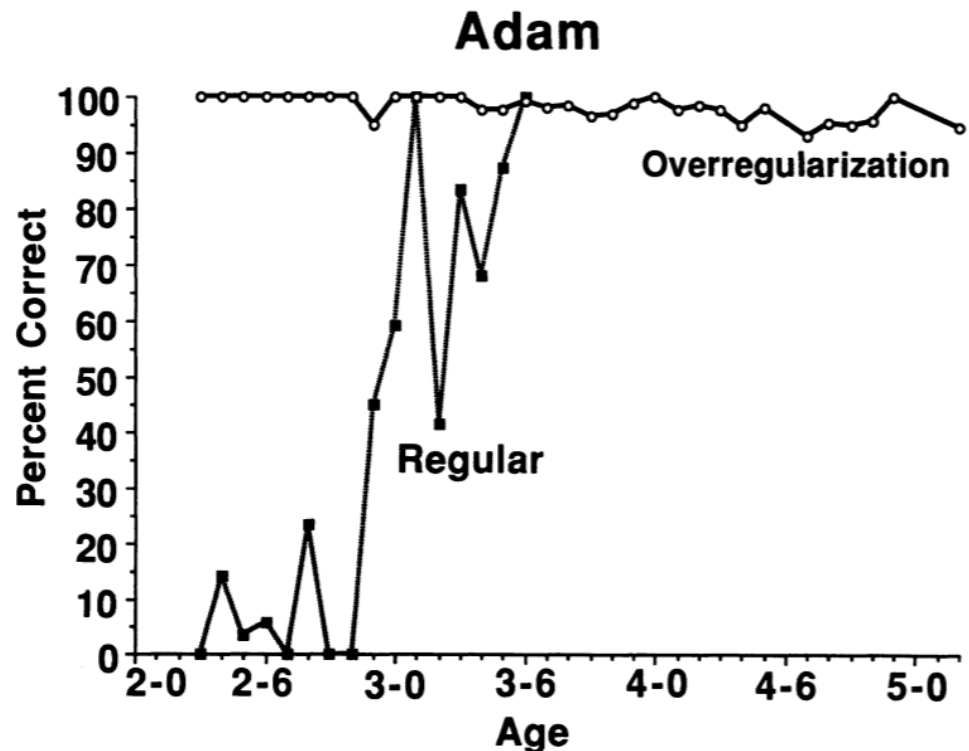
✓ High overall accuracy



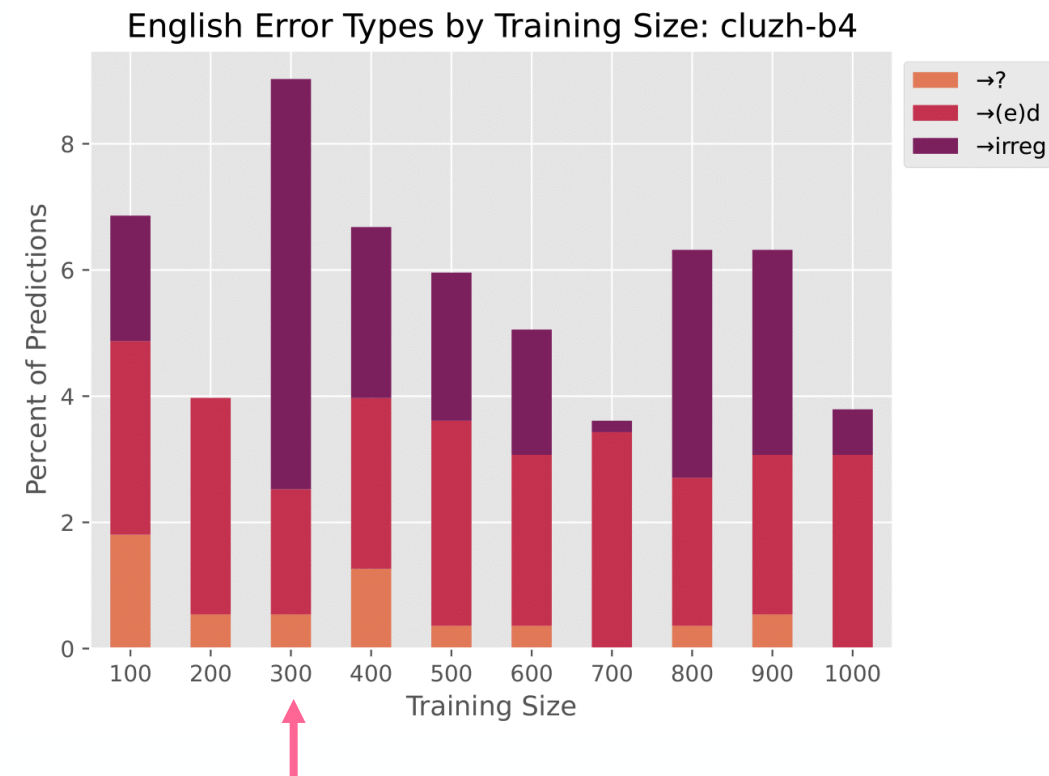
⚠ No developmental regression: spike in error rate caused by over-irregularization

Kodner et al. (2023, CogSci)

Model Results: English Past Tense



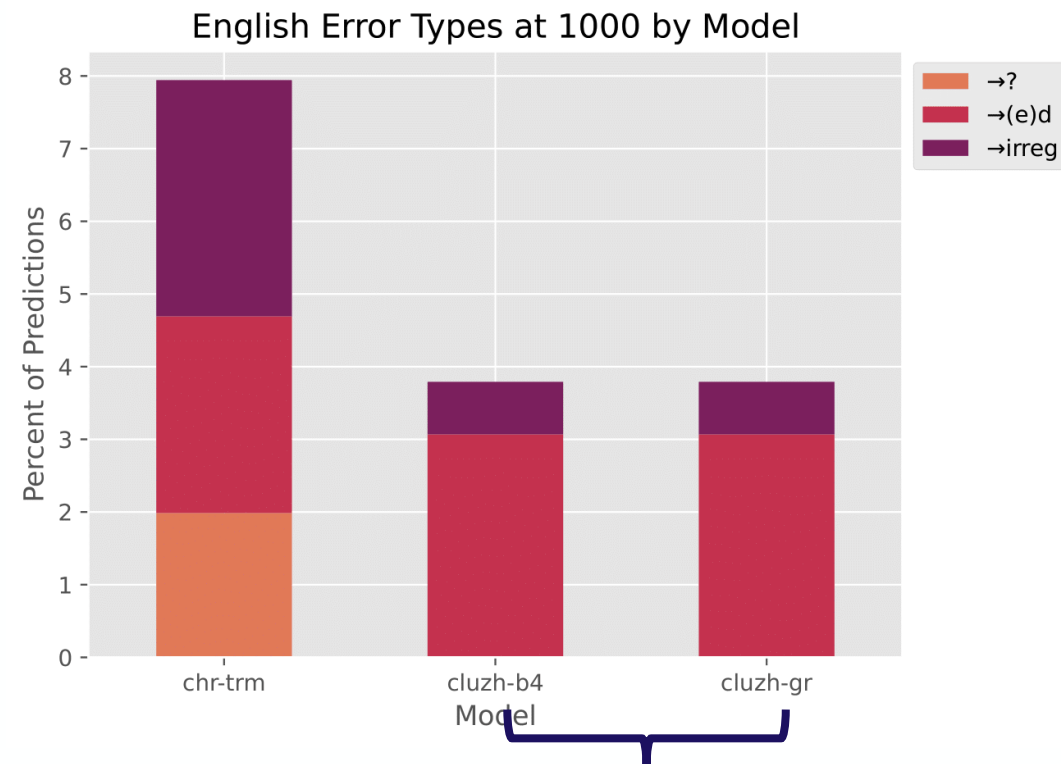
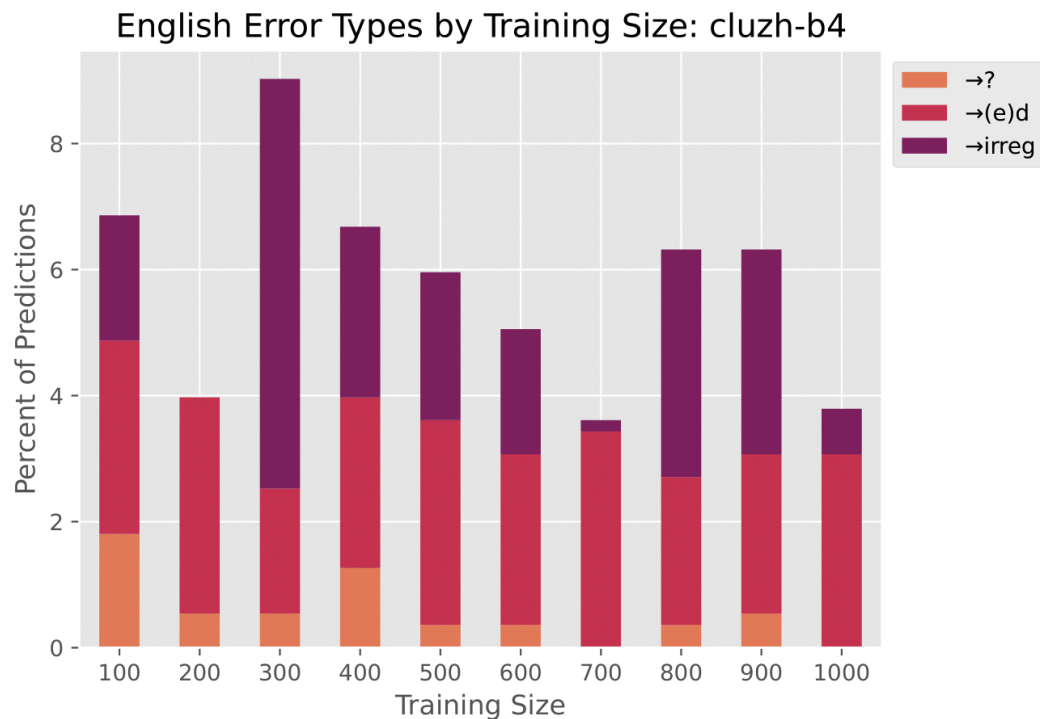
- 🚨 **Oscillation** in distribution of errors
- 🚨 **Oscillation** is **not** developmental regression, contra Kirov & Cotterell



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

Kodner et al. (2023, CogSci)

Model Results: English Past Tense



✓ More **over-regularization** than over-irregularization

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

Kodner et al. (2023, CogSci)

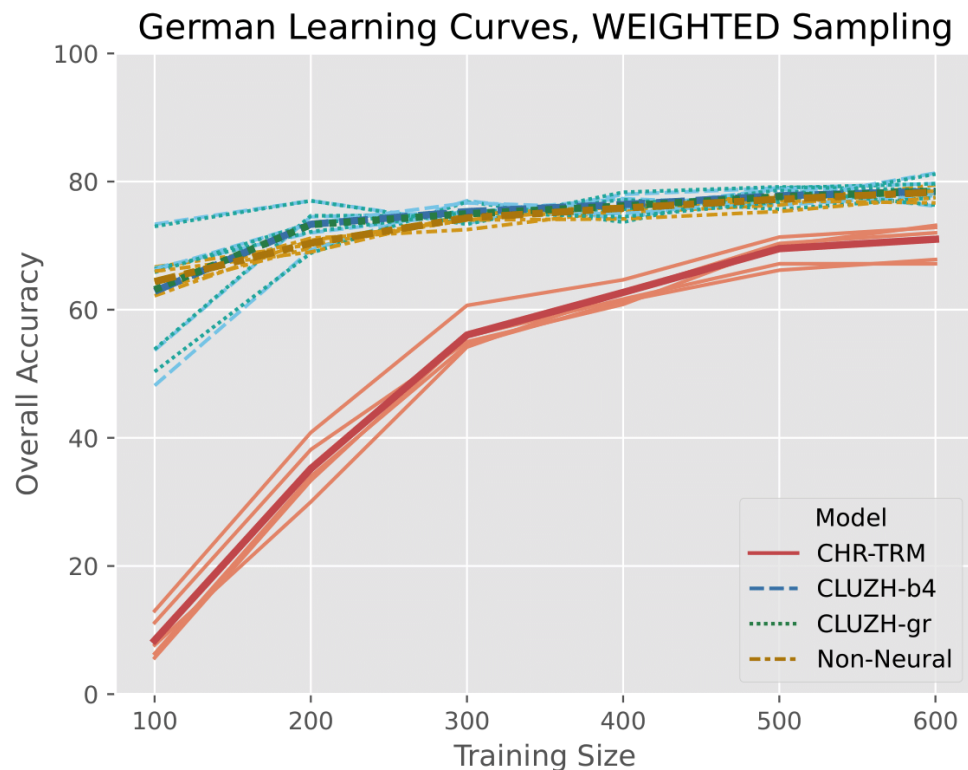
Acquisition Patterns: German Noun Plurals

- **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%

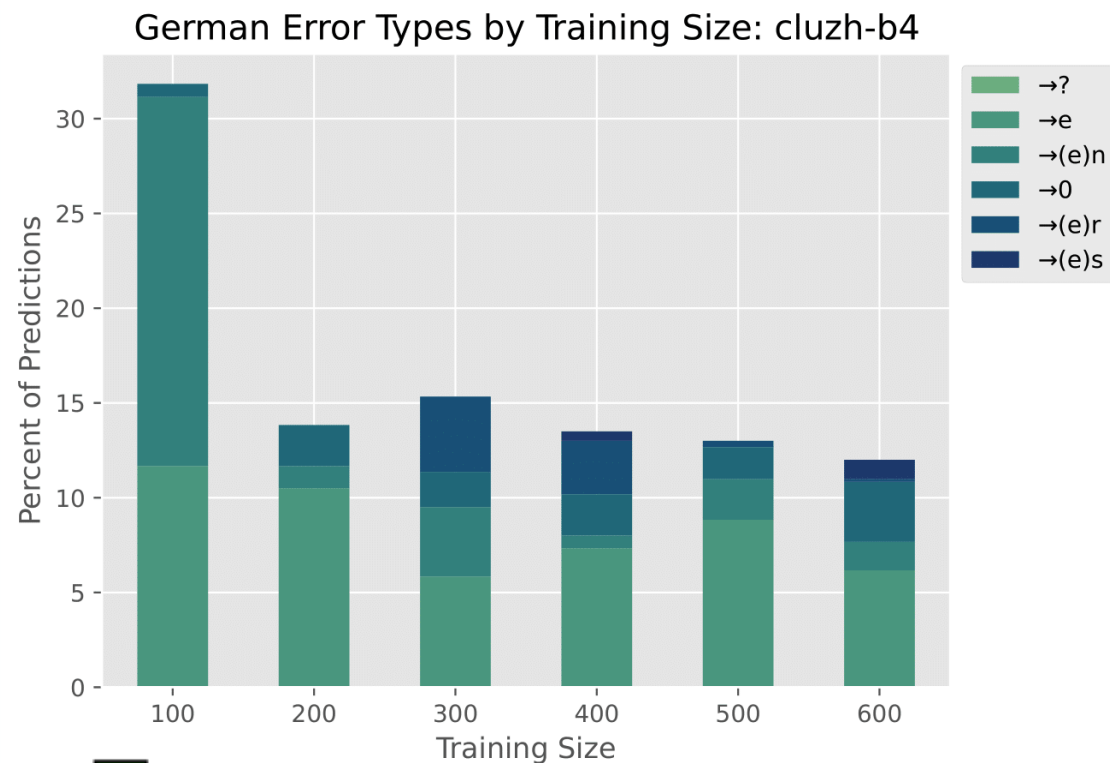
Kopcke (1998); Marcus et al. (1995);
Szagun (2001); Elsen (2002);
Sonnenstuhl & Huth (2002); Corkerey et al. (2019)

Model Results: German Noun Plurals



Lower accuracy than English

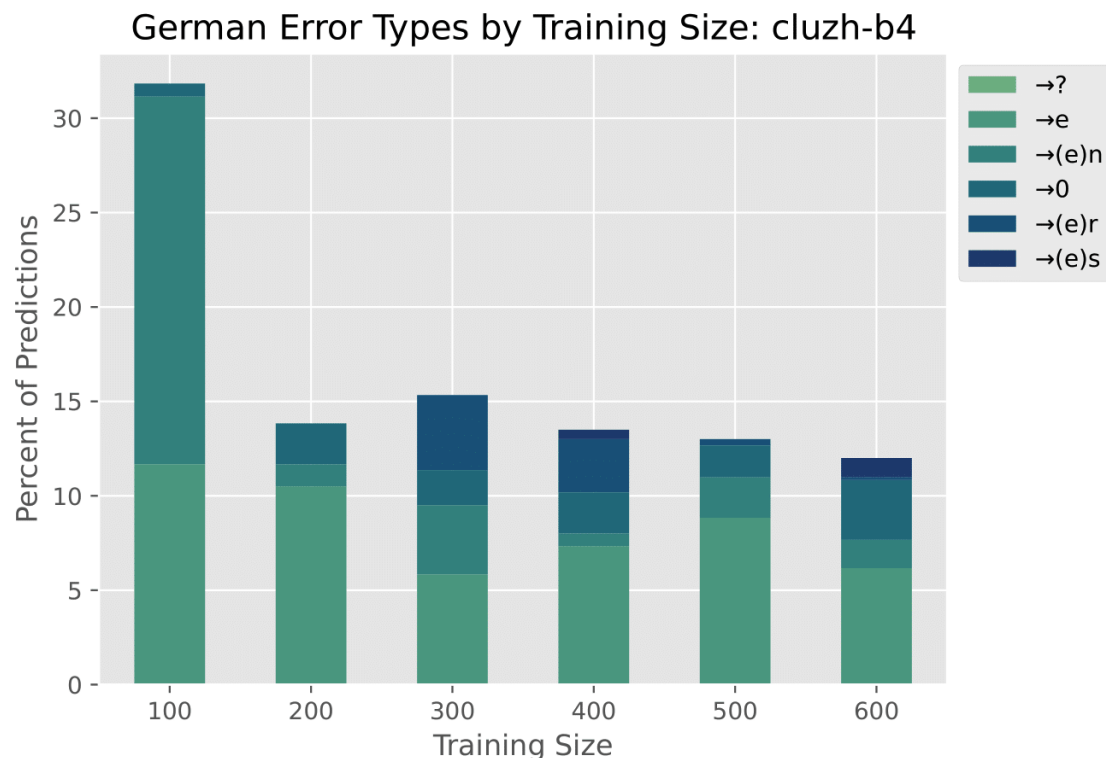
Gawlitze-Maiwald (1994); Elsen (2002)



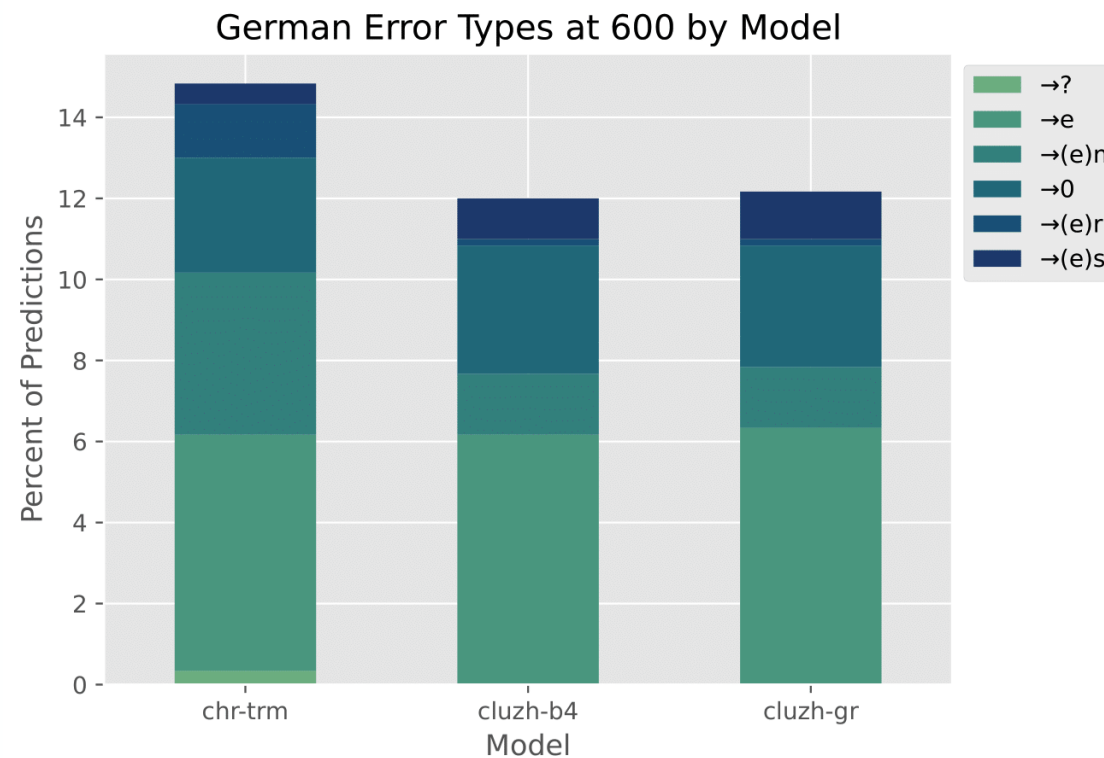
✓ Early overapplication of $-e$ and $-(e)n$ fits well with developmental findings

🚧 High error rates

Model Results: German Noun Plurals



🚨 Error rates still **far too high!**

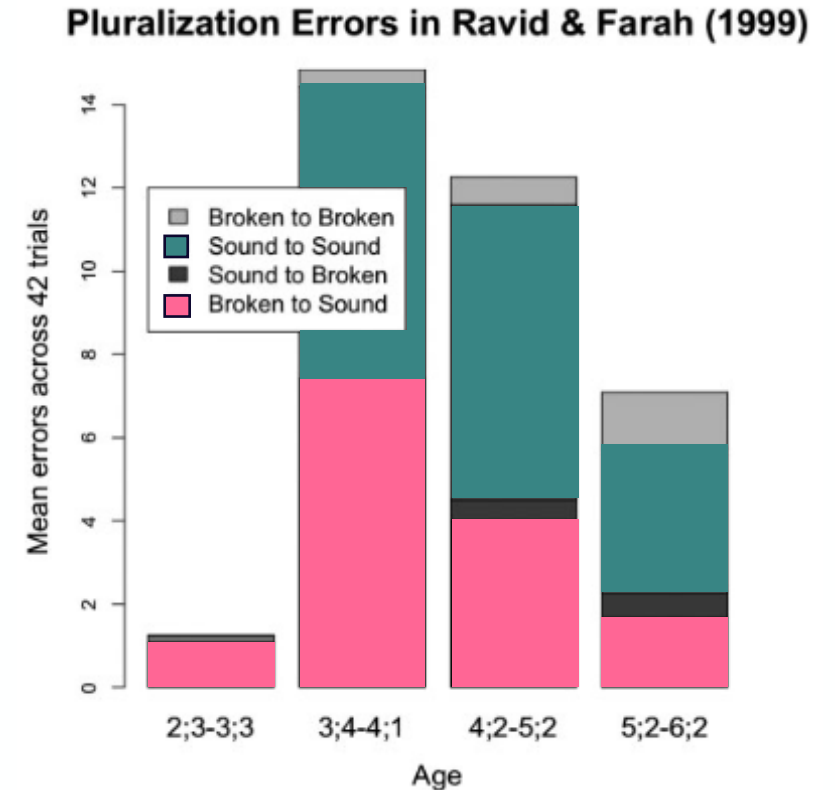


✅ Some **over-application of -s** is present for all systems on full train

✅ Learning trajectories roughly as expected

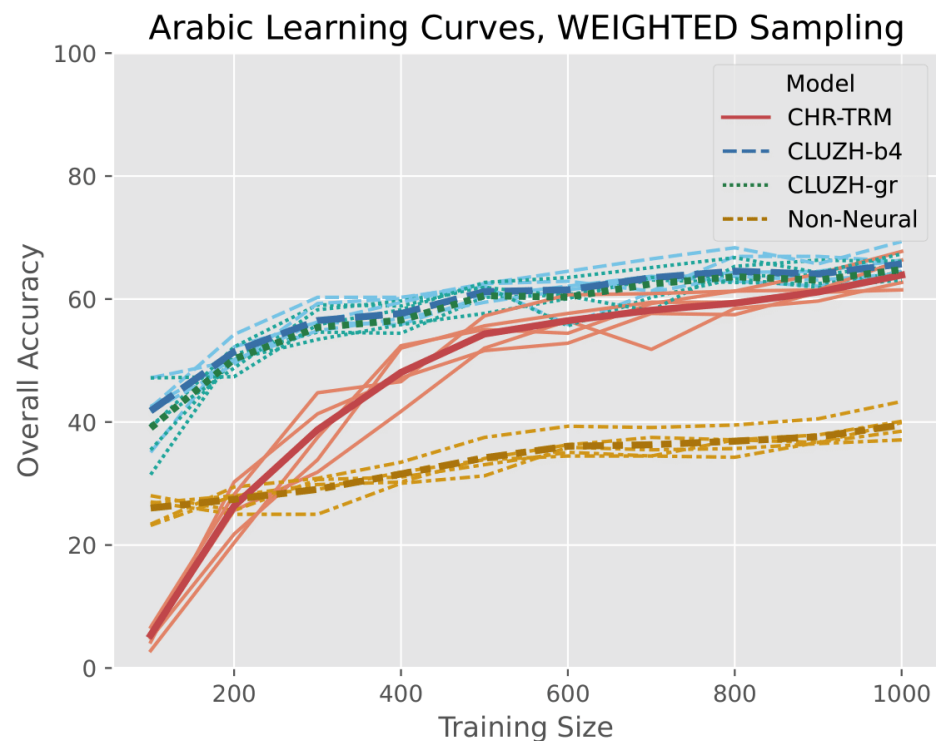
Acquisition Patterns: Arabic Noun Plurals

- 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**

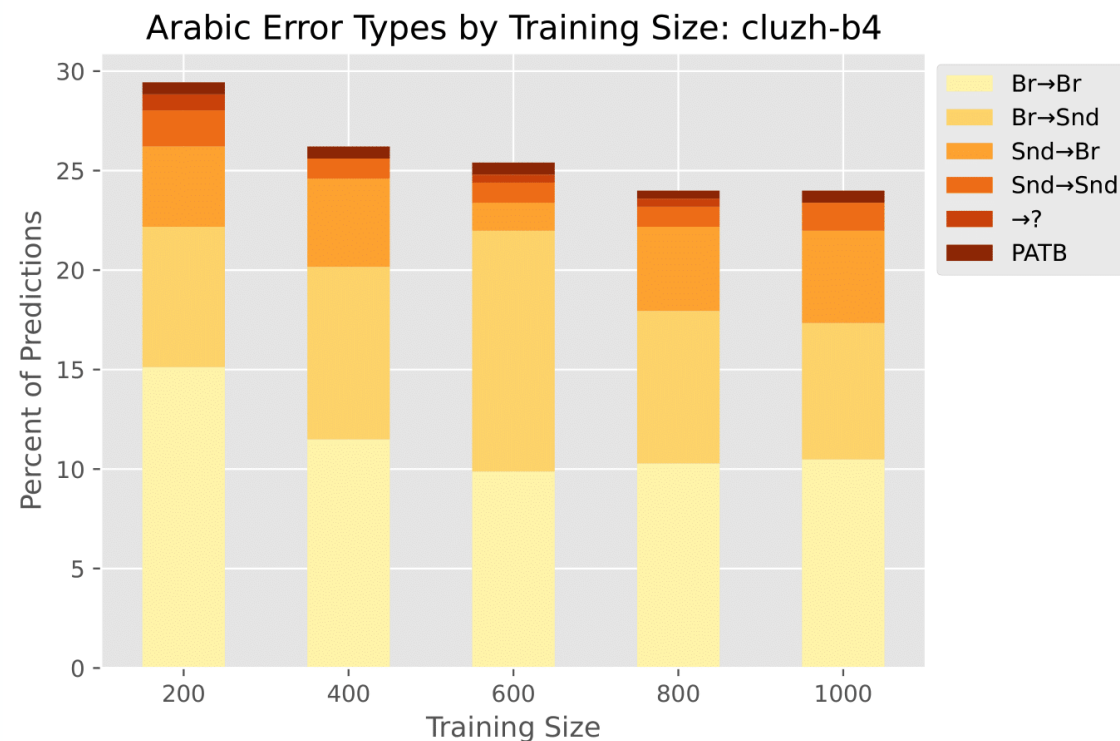


Ravid & Farrah (1999); Dawdy-Hesterberg and Pierrehumbert (2014)

Model Results: Arabic Noun Plurals



Lower accuracy than English/German

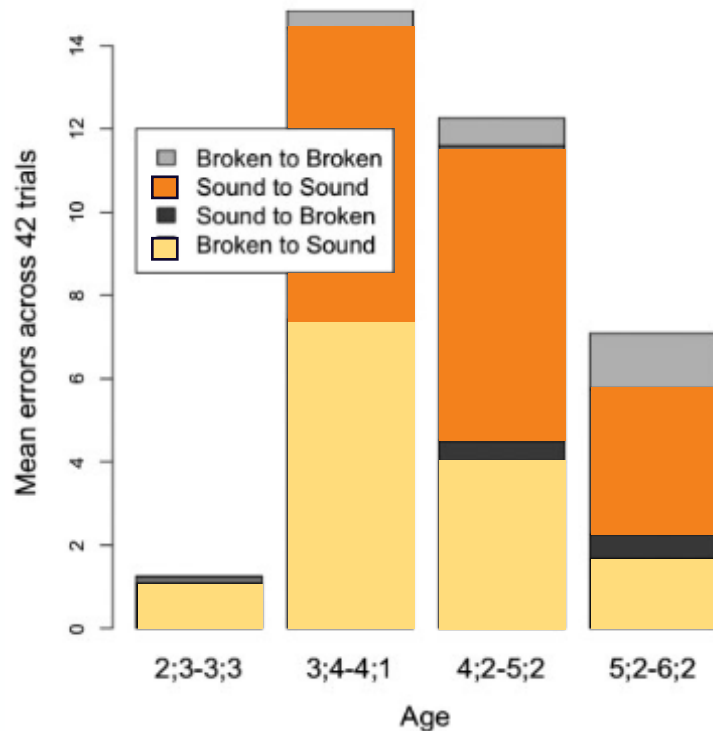


 **Learning is monotonic:** neither type of developmental regression is observed

Kodner et al. (2023, CogSci)

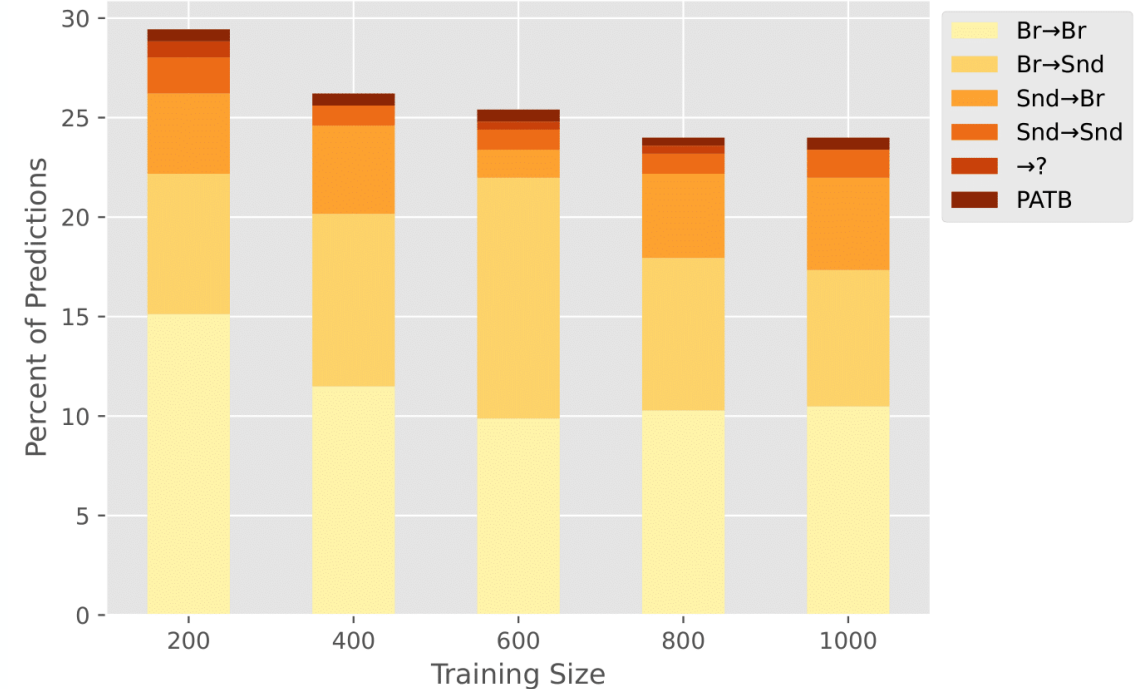
Model Results: Arabic Noun Plurals

Pluralization Errors in Ravid & Farah (1999)



🚨 **Overpredicts** Broken-Broken and Sound-Broken errors, **underpredicts child errors**

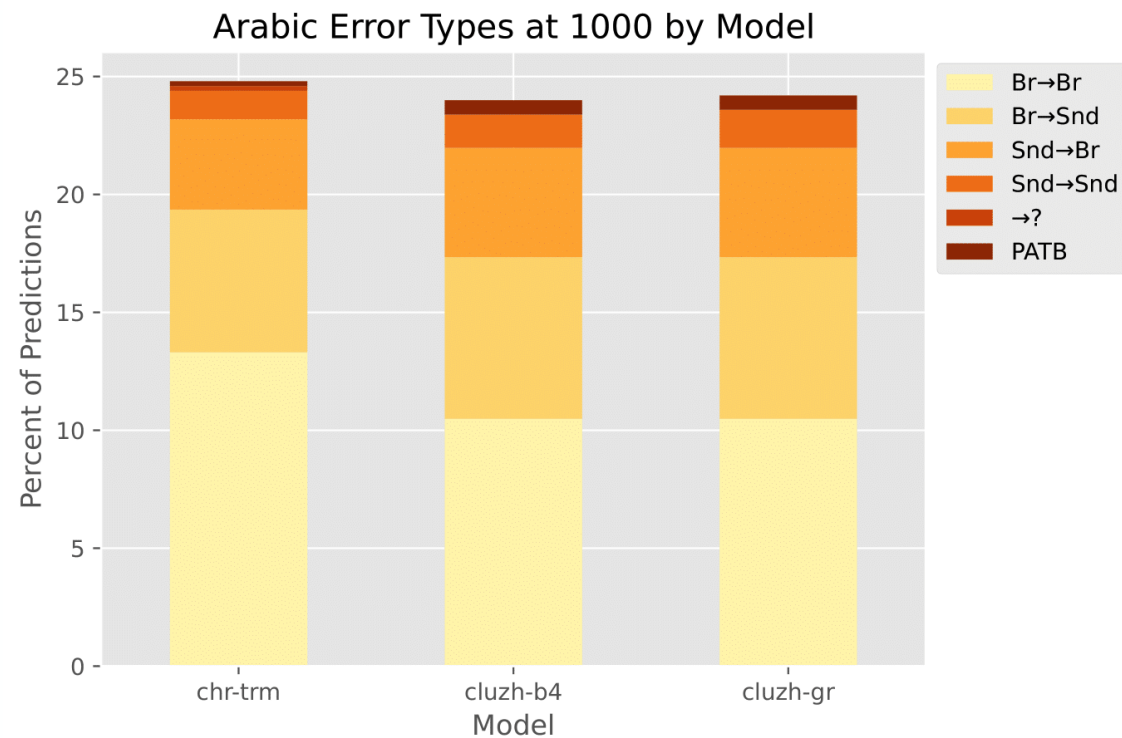
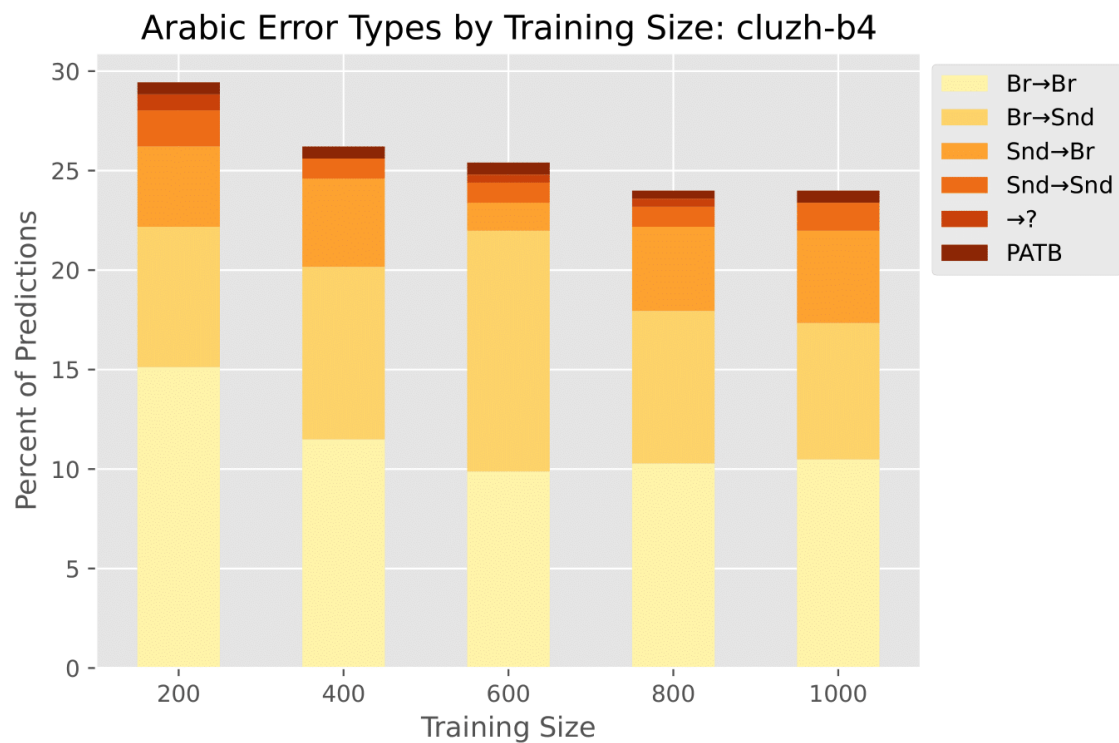
Arabic Error Types by Training Size: cluzh-b4



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

Kodner et al. (2023, CogSci)

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: Model Errors

- 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 NNMs are clearly **not learning morphology in the same way as humans**

Kodner et al. (2023, CogSci)

Conclusions

High accuracy doesn't necessarily entail **cognitive plausibility**

- Do models **generalize** like humans?
 - **To novel features, no** – they struggle more than humans
 - **To novel lemmas**, they do fairly well but:
- Do models make **human-like errors**?
 - **No**: too much overirregularization, no developmental regression

When we look at NNMs through these questions, we see that they **aren't cognitively plausible models of acquisition!**

BUT: more thorough evaluation **helps us understand why!**

Thank you!!



Caleb Belth
U. of Utah



Deniz Beser



Salam Khalifa
Stony Brook



Jordan Kodner
Stony Brook



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.**

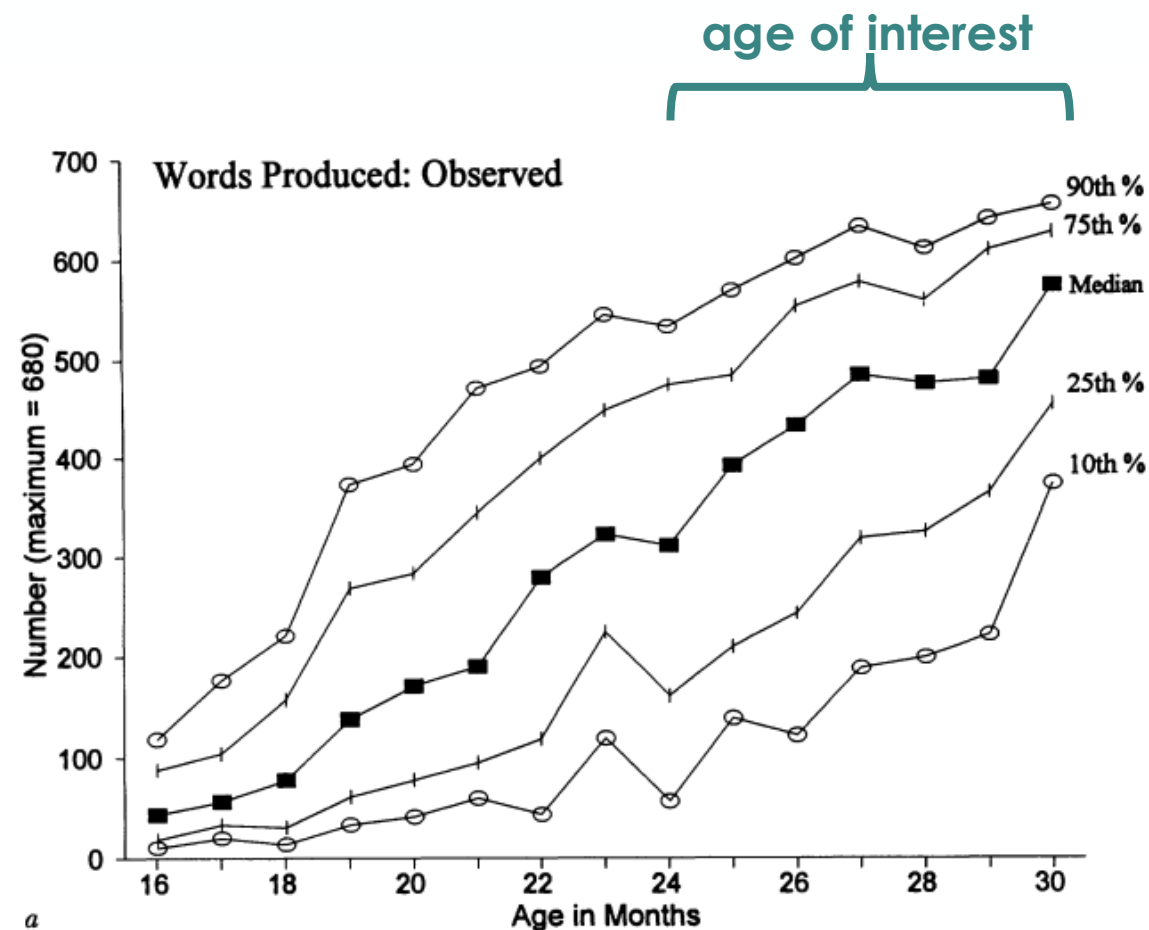


Extra Slides...

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)



(from Fenson et al 1994)

Morphological Inflection: Solved?



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

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

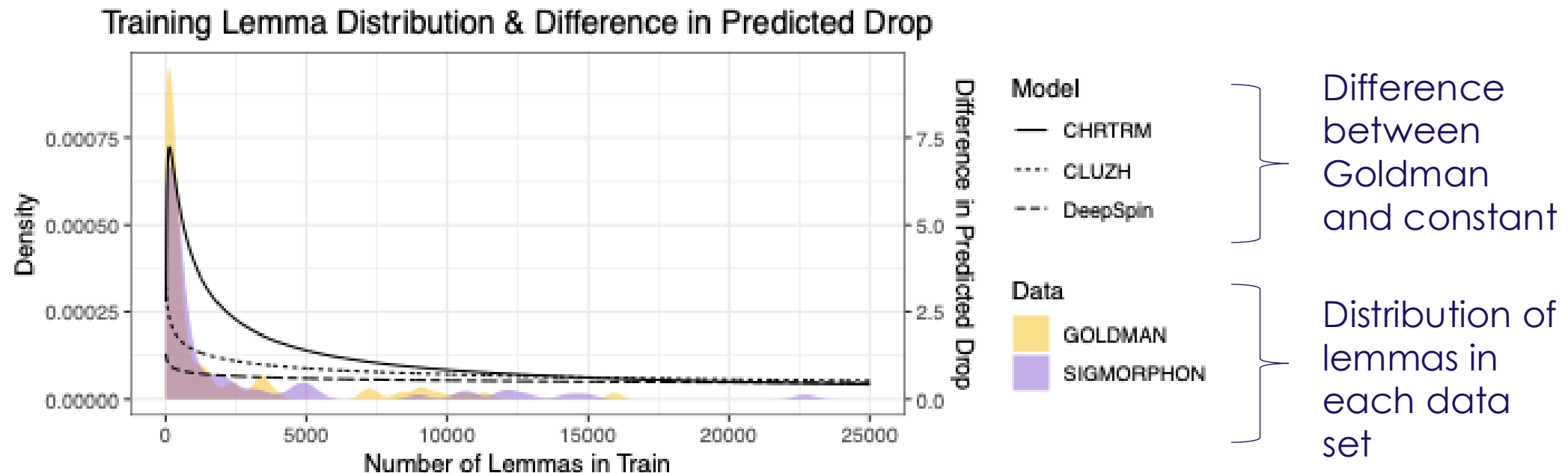
Irish	91.50(uzh-2)	77.10(uzh-1)	37.70(uzh-1)
Scottish-gaelic	—	94.00(iitbhu-iiiith-1)	74.00(iitbhu-iiiith-2)

**Morphological Inflection
isn't solved!**

Explaining the Discrepancy

Goldman: predicted accuracy on L lemmas minus predicted accuracy on $0.7L$, both on Goldman

Constant: predicted accuracy on L lemmas on Sigmorphon minus on L lemmas on Goldman



Payne & Kodner (2025, ACL)