Child Errors as a Window into Morphological Acquisition

Sarah Brogden Payne
Stony Brook University
sarah.payne@stonybrook.edu

Ling 444: First Language Development
October 18, 2024

Background: Morphological Acquisition

Children **learn inflectional morphology**

- From highly ***sparse, skewed*** input
- On <1000 lemmas
- Despite ***exceptions***
- With complex systems of ***allomorphy***

Challenging problem!

Background: Morphological Acquisition

Children make **systematic errors** cross-linguistically

- ***Overregularization:*** e.g. "feel-feeled"
- ***Omissions of Marking:*** e.g. "Papa have it"

Almost all errors

Why **these** errors and not others?

What do the errors tell us about:

- Acquisition?
- The resulting grammar?

Models of morphological acquisition should address these questions

Outline

Why computational modeling?

What makes a good model?

Previous work: The Past Tense Debate

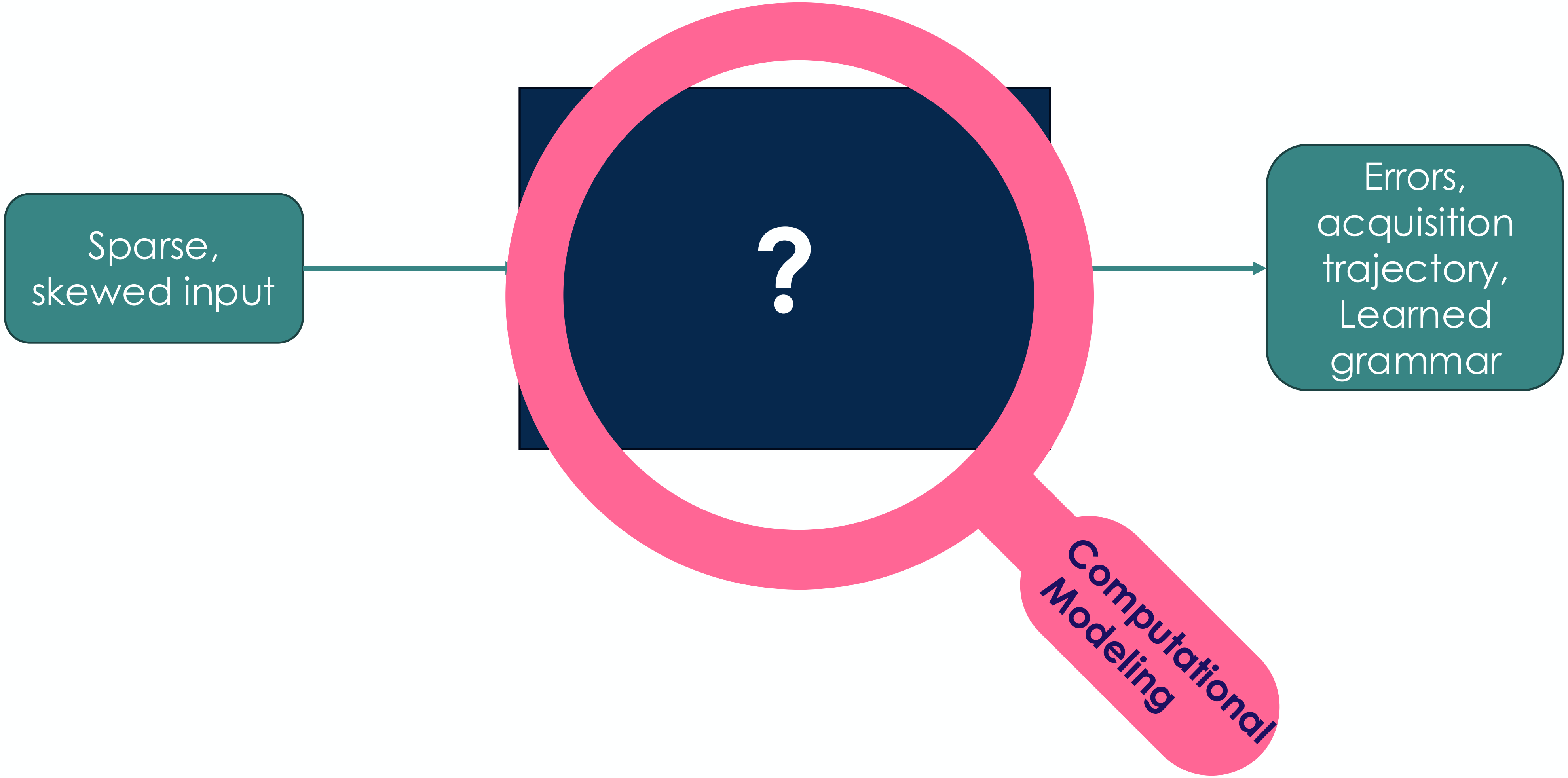
Proposal

- **ATP:** mapping features to form
- **SCL:** learning inflectional classes

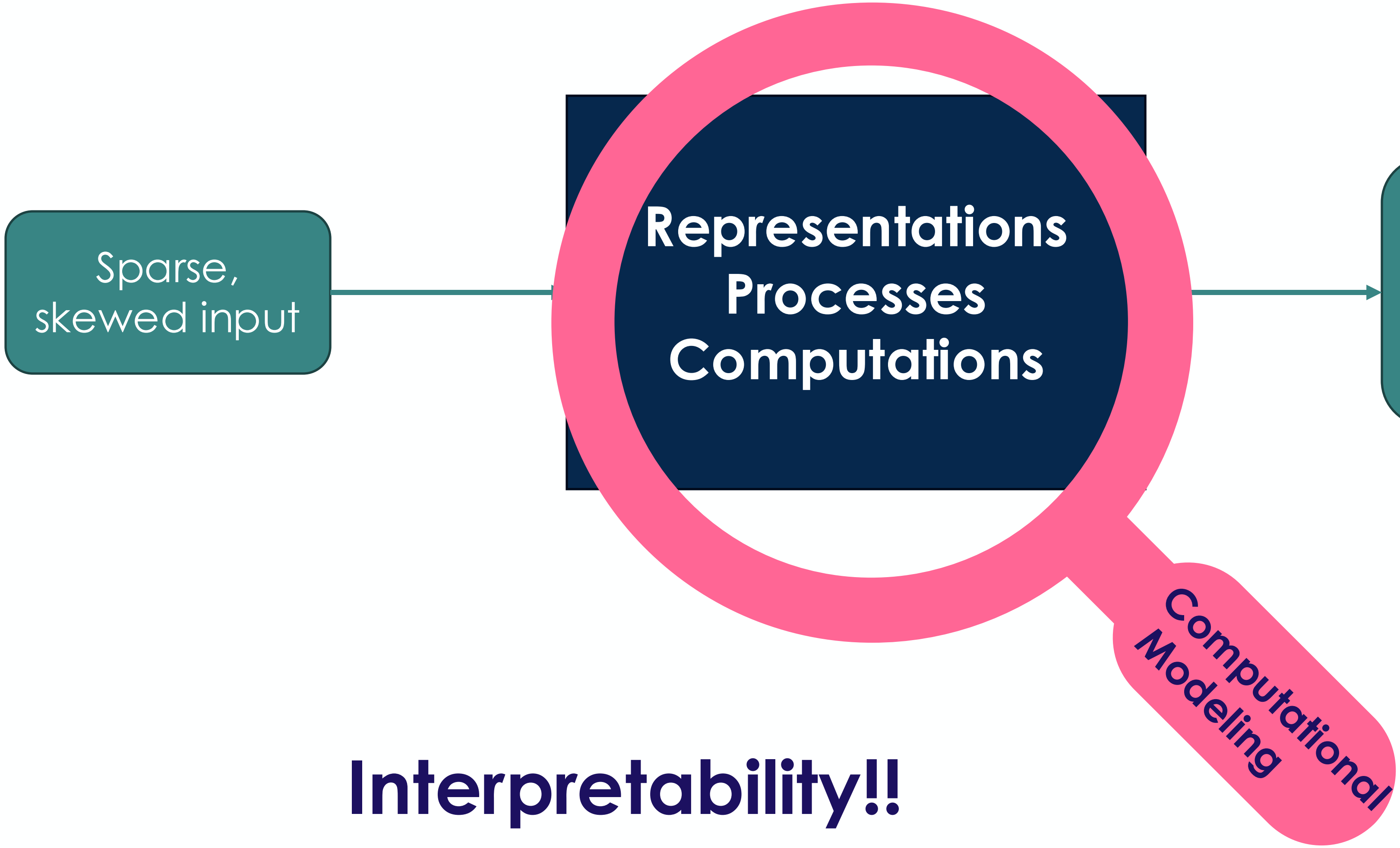
Discussion & future work

Why Computational Modeling?

A Mechanistic Account



A Mechanistic Account



Interpretability!!

A Mechanistic Account

What Happens

- Sparse, skewed input
- Errors
- Acquisition trajectories

Description, not explanation

Why it Happens

- What are the **mental computations** that allow the child to learn from this input?
- How do these computations yield the errors we see?
- How do they lead to the trajectories we observe?

What Makes a Good Mechanistic Account?

What Makes a Good Model?

Input:

- ***Small*** vocabulary
- ***Sparse*** paradigms

Errors:

- ***Omissions***, not substitutions
- ***Over-regularizations***, not over-irregularizations
- Developmental ***regression***

Interpretability: why does it do what it does?

Input Sparsity: Early Vocabulary

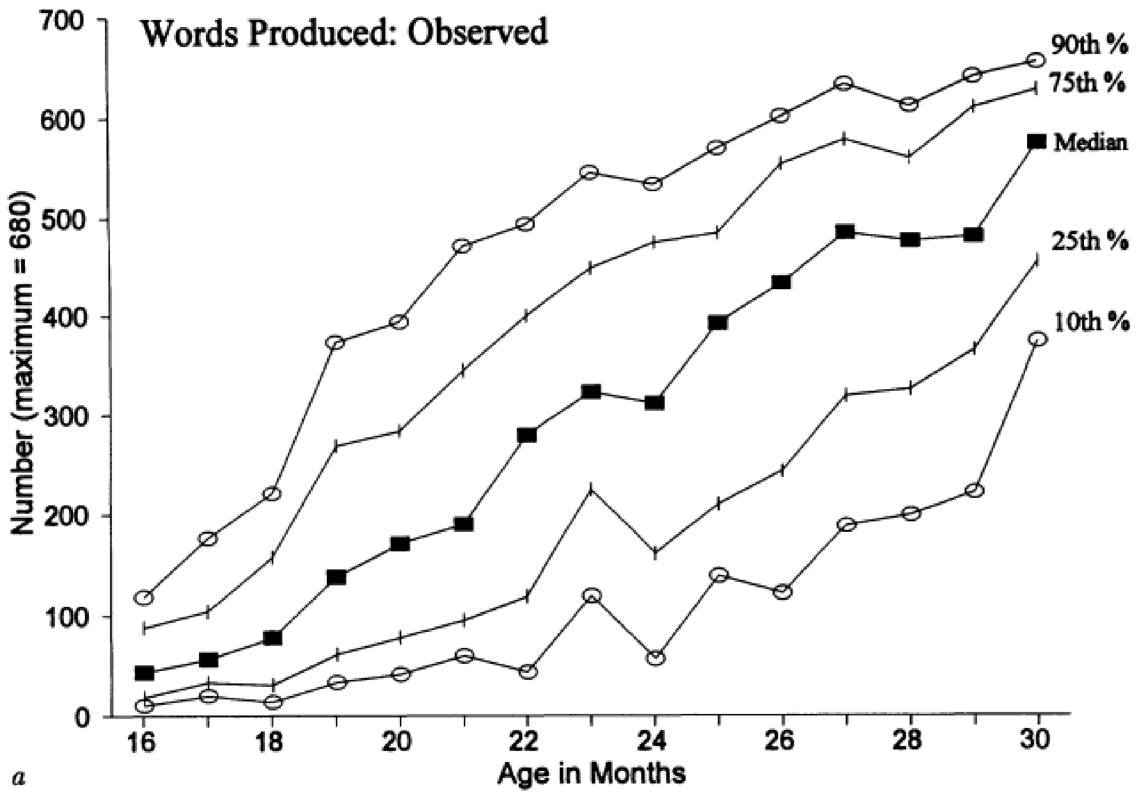
At 2;0: 200-500 words cross-linguistically

At 3;0: <1000 words cross-linguistically

Early vocabulary makeup:

- ~50% **nouns**
- ~25% **verbs**

age of interest



(from Fenson et al 1994)

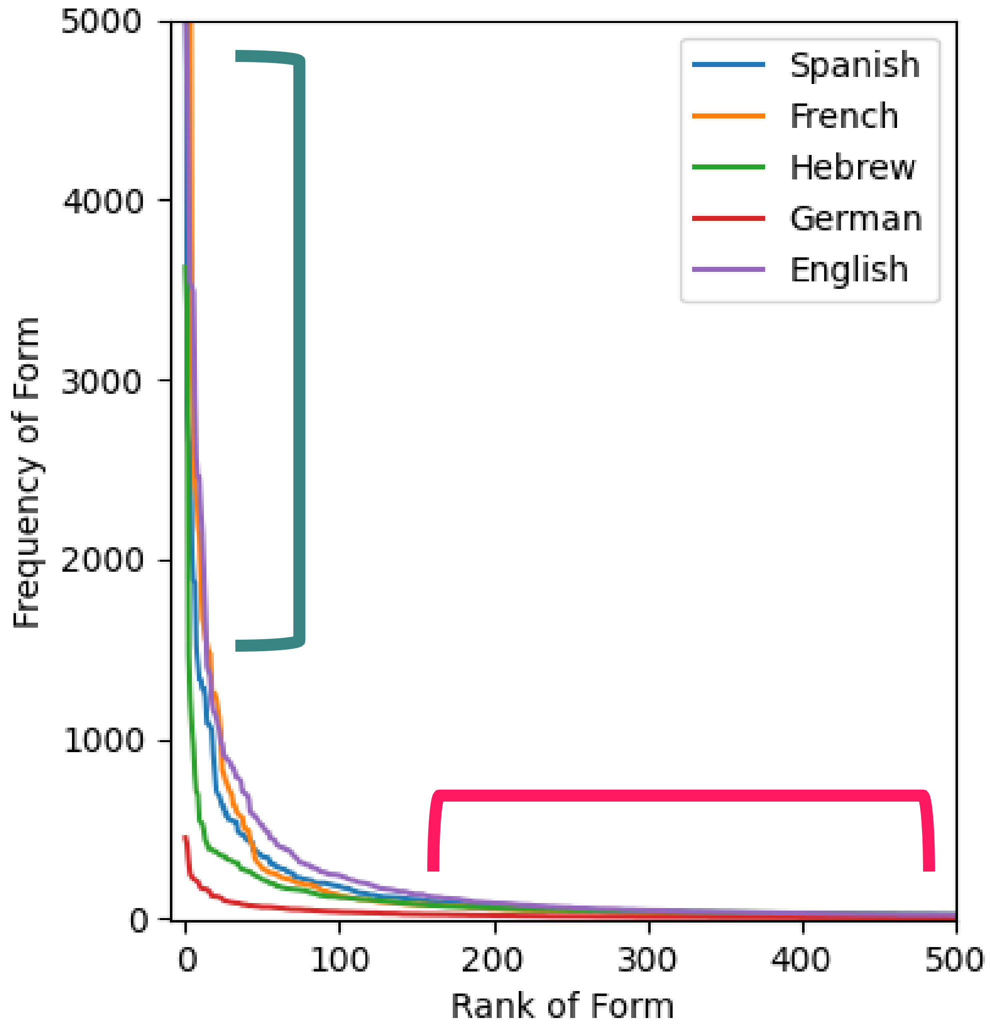
The Long Tail: Zipf's Law

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)**



The Long Tail: Paradigm Saturation

How many of its possible inflected forms does a word **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)

The Long Tail: Paradigm Saturation

How many of its possible inflected forms does a word **actually occur** in?

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

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

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

(Chan 2008, Lignos & Yang 2016)

The Long Tail: Paradigm Saturation

How many of its possible inflected forms does a word **actually occur** in?

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

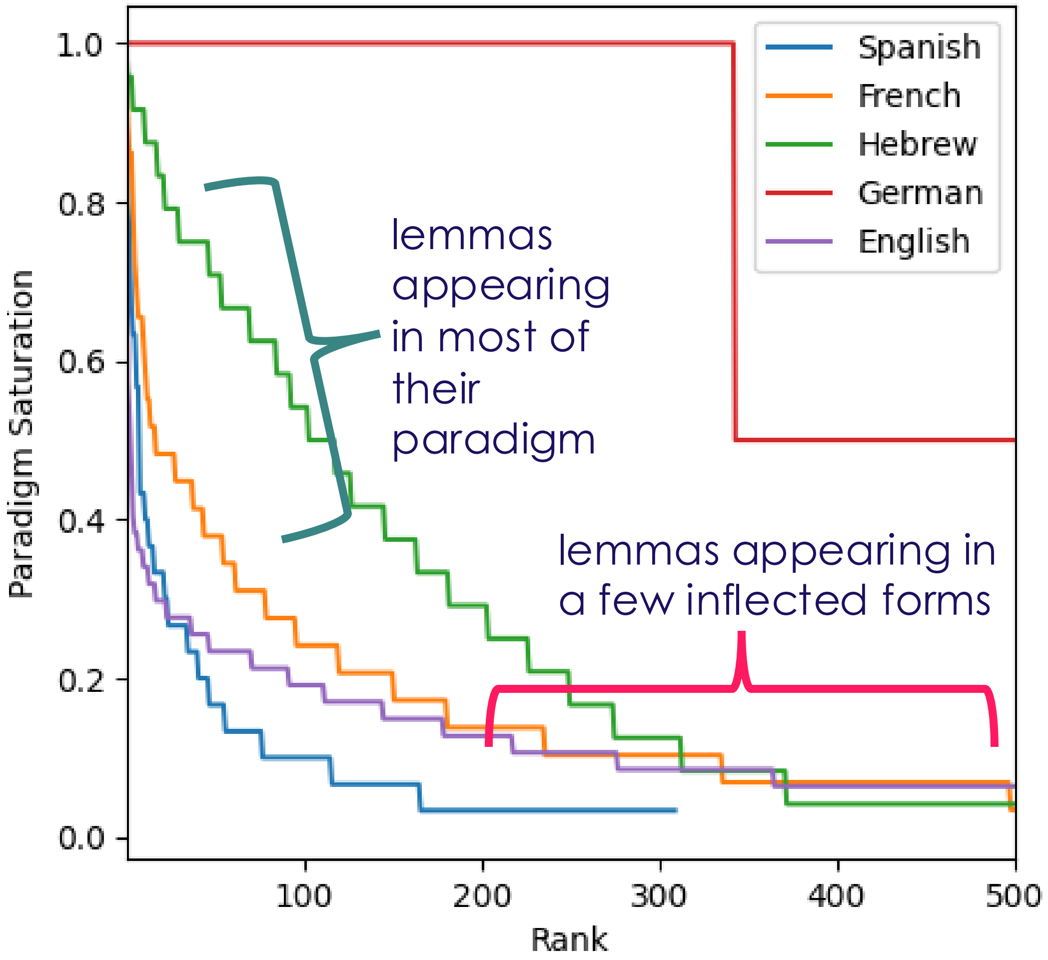
	Present	Preterite	Imperfect	Conditional	Future
1SG	***amo***	*trabajé*	***amaba***	*trabajía*	***amaré***
2SG	*tomas*	***amaste***	*mirabas*	*mirarías*	*esperás*
3SG	***ama***	*esperó*	***amaba***	*espería*	*tomará*
1PL	***amamos***	*miramos*	*mirabamos*	*tomaríamos*	*miraremos*
2PL	*tratáis*				
3PL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

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

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

(Chan 2008, Lignos & Yang 2016)

The Long Tail: Paradigm Saturation



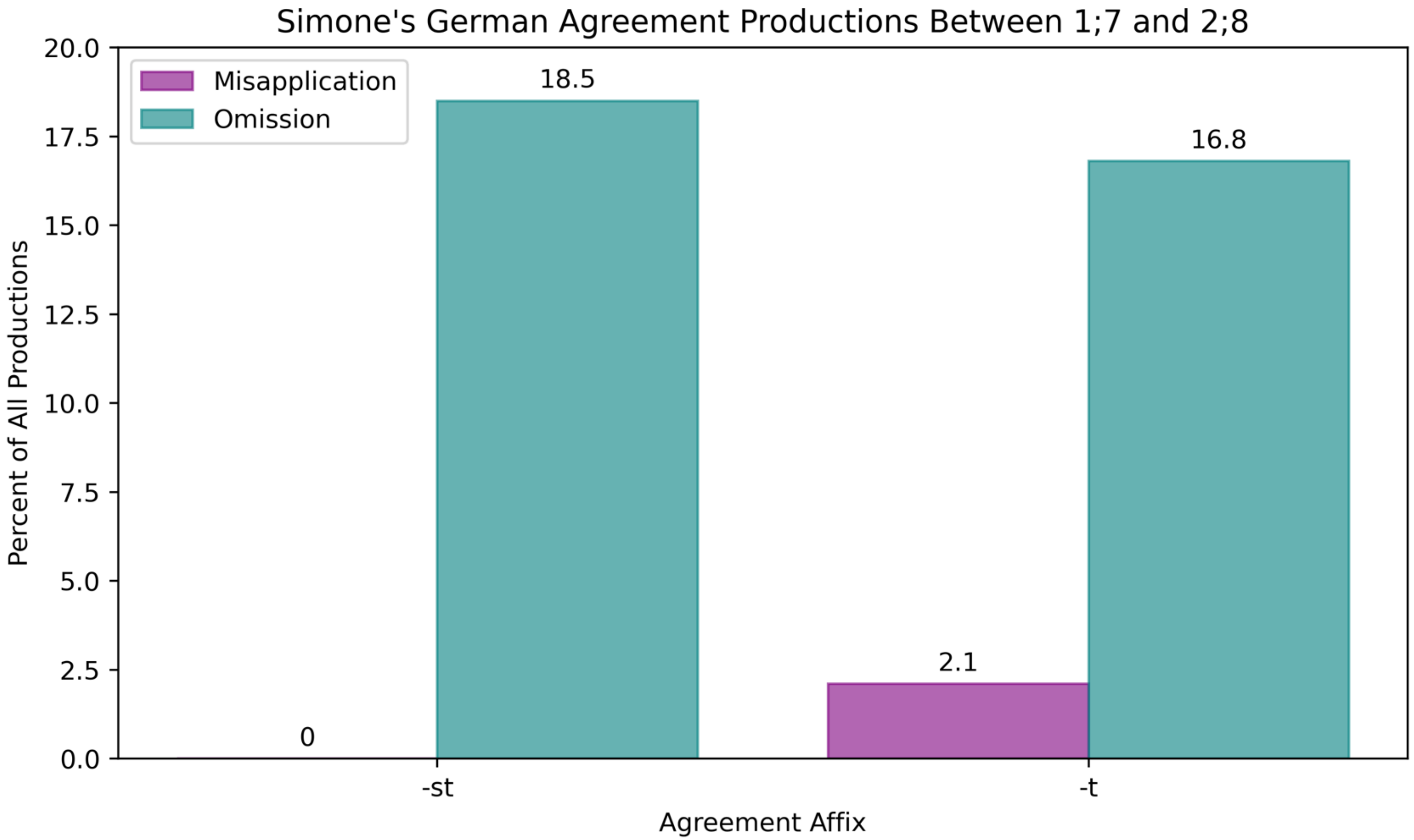
Errors: Omissions vs. Substitutions

Omissions

- e.g., “Papa have it”

Substitutions

- e.g. “I has it”



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

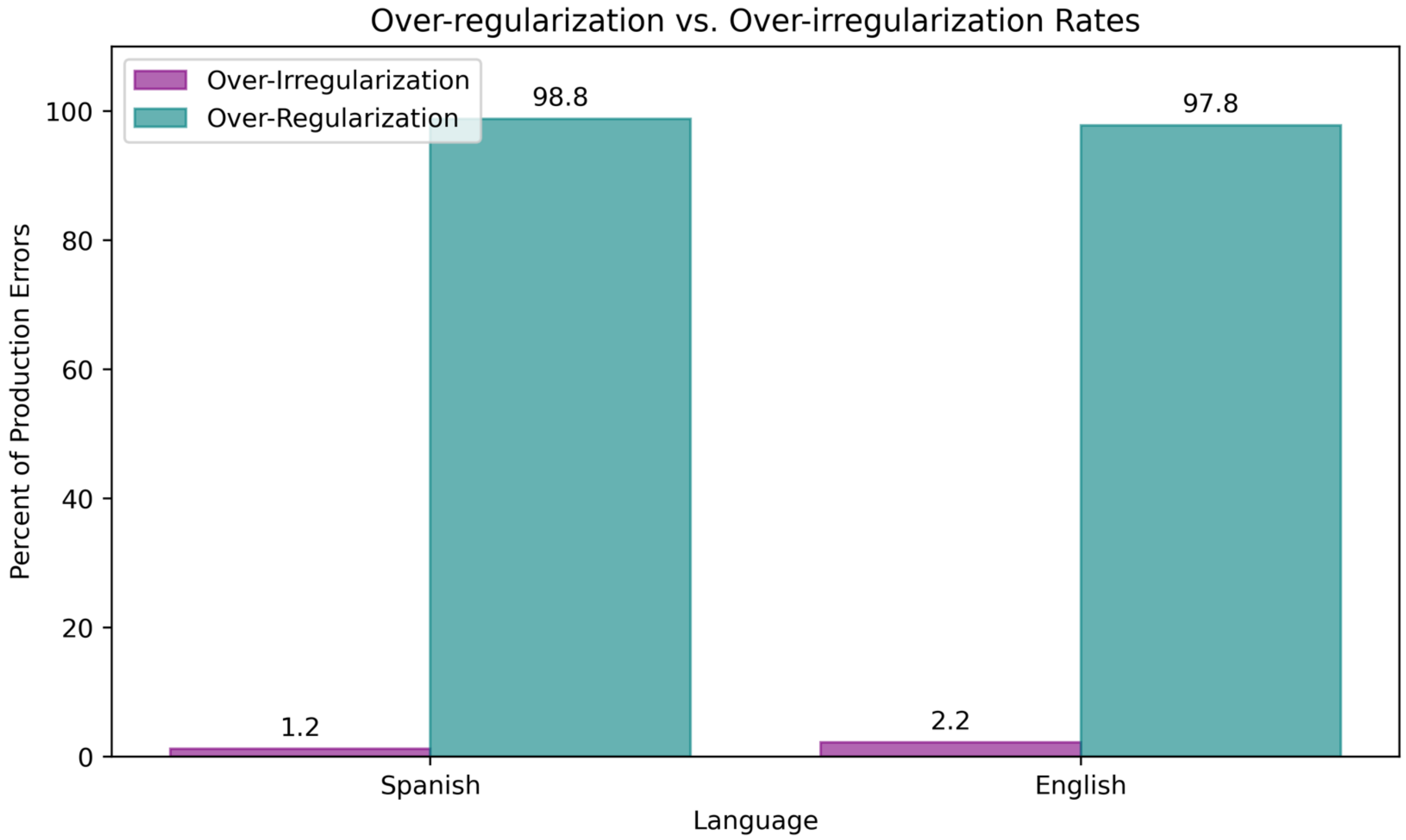
Errors: Over-regularization

Over-regularization

- e.g. ***feel-feeled***

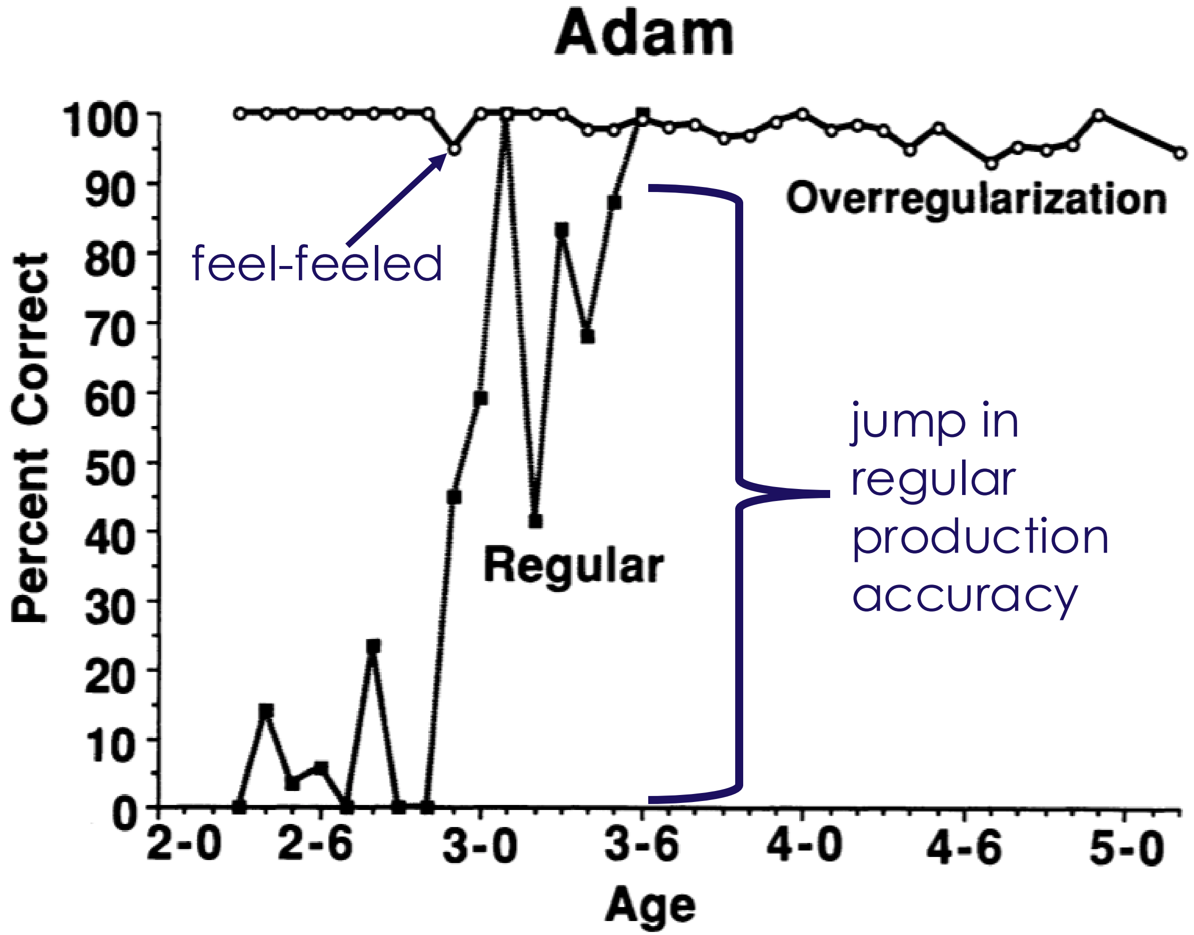
Over-irregularization

- e.g. ***bite-bote***



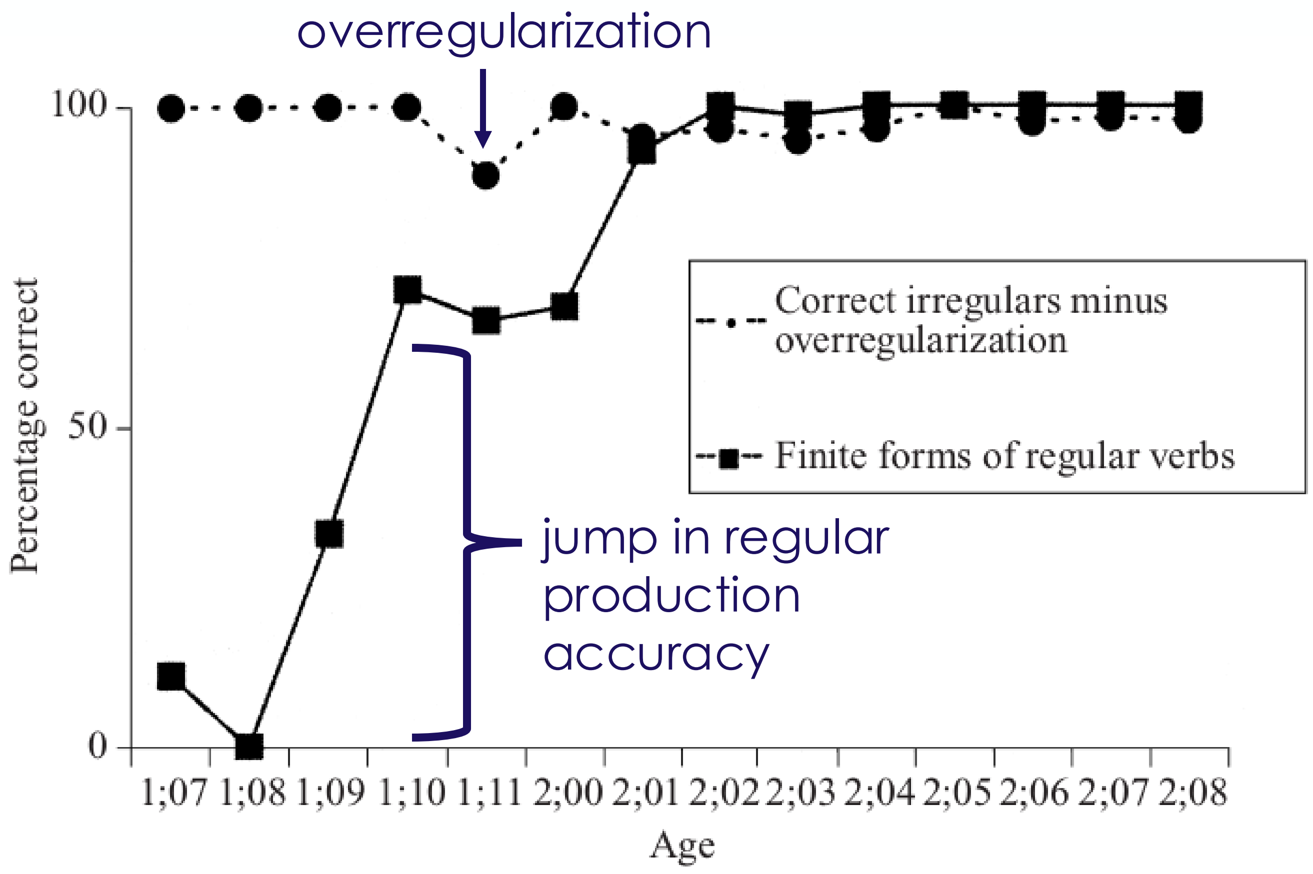
(Maslen et al 2004, Xu & Pinker 1995, Clahsen et al 2002)

Background: Developmental Regression



(from Marcus et al 1992)

Background: Developmental Regression



(from Clahsen, Aveledo, and Roca 2002)

Summary: What Makes a Good Model?

Learn from:

- ***Small*** vocabulary
- ***Sparse*** paradigms

Errors:

- ***Omissions***, not substitutions
- ***Over-regularizations***, not over-irregularizations
- Developmental ***regression***

Interpretability: we want to know why it does what it does

The Past Tense Debate(s): Are Neural Networks Good Models?

Are Neural Networks Good Models?

So true bestie!

Yes, our neural network models morphological acquisition!

James Mclelland

David Rumelhart

- ✅ Trained on **plausible data**
- ✅ Exhibits **developmental regression**
- ✅ **Over-regularizes**

Not Really...

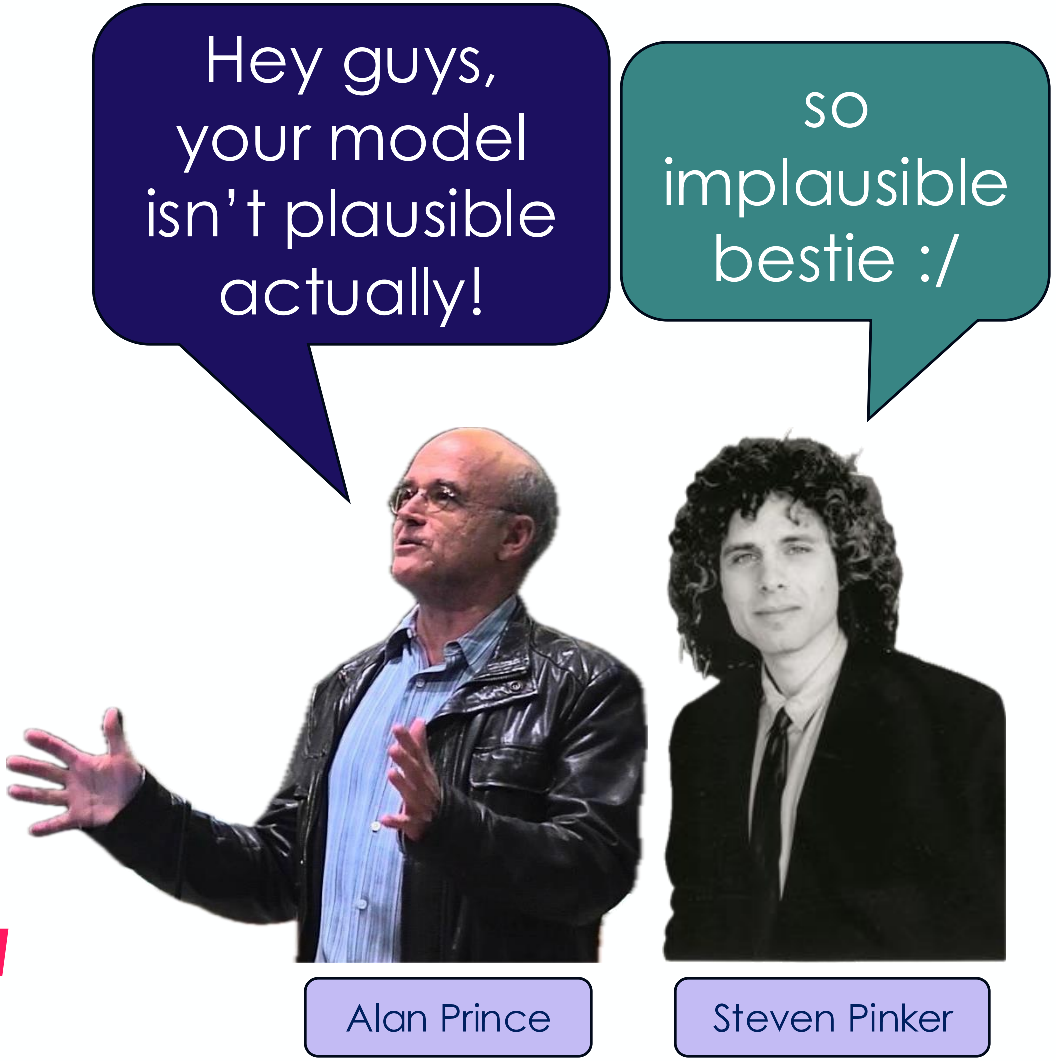
🚨 Not actually trained on **plausible data**

- First trained on **80% irregulars**
- Then trained on **80% regulars**

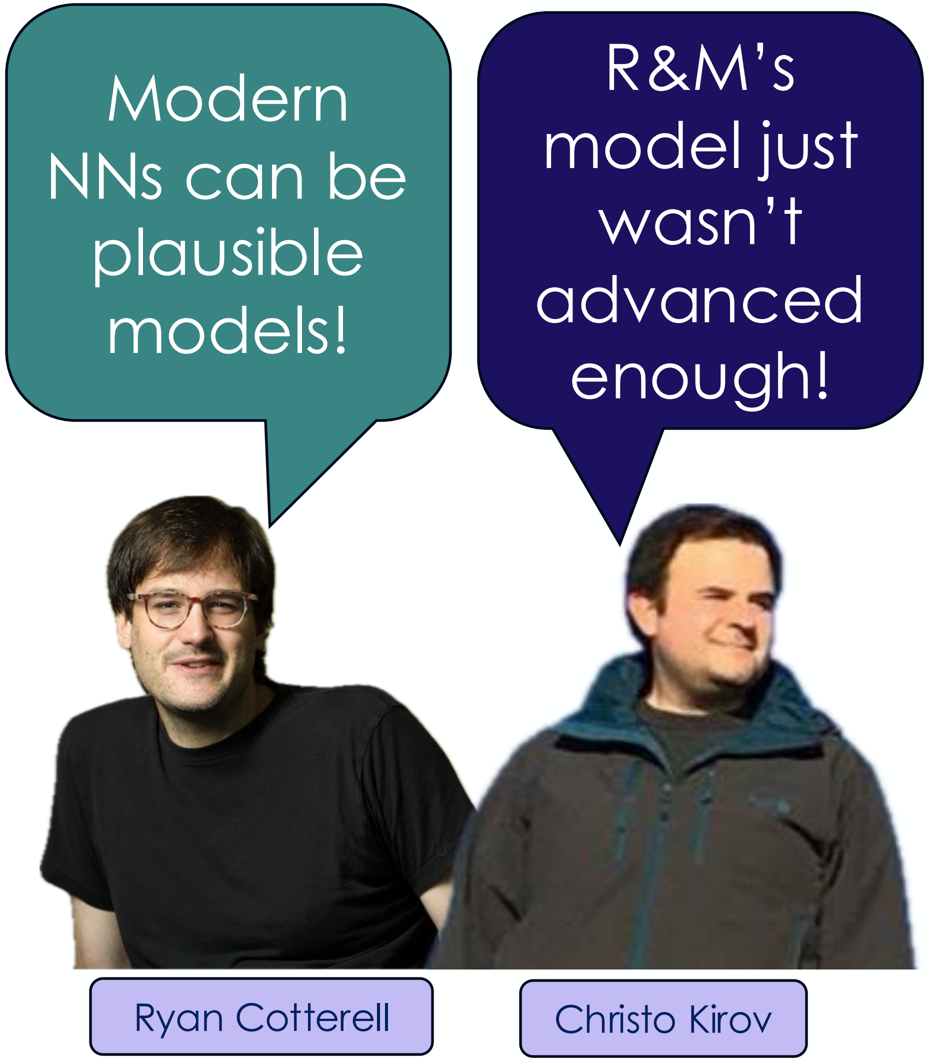
🚨**Developmental regression** results from implausible training data

🚨**Over-irregularizes**

- ***sip-sept, type-typeded, mail-membled***



Are we there yet?



✅ Today's NNs **overcome practical limitations!**

- Near 100% test accuracy ✅
- Learn several inflectional classes ✅

✅ Trained on **developmentally representative data**

✅ Main errors = **over-regularizations**

No, we aren't...

🚨 Still **over-irregularizes** way more than humans

- Predictions don't match well with **wug test judgments**

🚨 No **developmental regression**

🚨 **Implausible training data**

- Trained on **> 3500 verbs in their full paradigm**
- Children know < 350 verbs at 3;0
- Would need to see **> 15k unique words** to see 3,500 in full paradigm

Hey, so your model also isn't plausible!

Engineering ≠ science!

This is getting ridiculous

Maria Corkery

Yevgen Matusevych

Sharon Goldwater

Summary: The Past Tense Debate(s)

What have we gotten from ~30 years of NN research?

- **Better accuracy**
- **More developed architecture**

Good for engineering!

What haven't we gotten?

- Still ***overproduce irregulars***
- Still ***no developmental regression***
- Still ***data-hungry***

Persistence of issues ⇒ **fundamental difference between neural models & human language faculty**

What would they tell us about acquisition, anyway?

- **They aren't interpretable!**

Proposal:
Recursive, Rule-based Learning

Mechanistic Account of Errors

Past tense debate(s):

Create a model in a certain existing framework

- In this case, **connectionism**

Check whether it can learn from **plausible data**

Check whether it **matches errors**

No a-priori reason to expect either

What if:

The design of our model was motivated by:

- The need to learn from **small, sparse data**
- The **types of errors** we expect it to make

Expect it to follow developmental patterns!

Error-Motivated Modeling

Children **over-regularize** & don't **over-irregularize**

Account for this with **rule-based mappings:**

- Learn rule like **Past ⇒ -ed**
- Apply rule when no exception is known
  - **Over-regularization** when exception not yet learned
  - **Developmental regression** when rule first learned

Abduction of Tolerable Productivity (ATP): recursively learn productive rules & their exceptions

Error-Motivated Modelling

Children **omit** inflectional affixes, but don't **substitute** them

Account for this with **initially-underspecified inflectional categories**

- Must learn e.g. that English contrasts **+3SG with -3SG**
- Underspecified category can't be productively mapped to form, so **omit inflection**

Sufficient Contrast Learner (SCL): recursively learn inflectional categories

Preliminaries: The TSP

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:

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

Sufficient positive evidence

Generalize a rule applying to *N* items and seen applying to *M* iff:

$$\underbrace{N - M}_{\text{worst-case e}} \leq \theta_N = \frac{N}{\ln N}$$

(Yang 2016)

Preliminaries: The TSP

Experimental evidence for the ***Tolerance-Sufficiency Principle***

$$\theta_{16} = 5.7$$

Emond & Shi (2020)

Preliminaries: Training Data

Children learn **frequent forms earlier** (Goodman et al 2008)

- Use ***most frequent forms from CHILDES***

Children use of **distributional cues** to learn meaning

- Intersect CHILDES with ***UniMorph features as a proxy for these cues***

Input: **(lemma, inflected, features)**

Language	Lemma	Inflected	Features
English	*walk*	*walked*	{V, PAST, 3, SG}
Spanish	*amar*	*amaban*	{V, 3, PL, PAST, IMPFV}
German	*Sache*	*Sachen*	{N, FEM, PL}

Mapping Features to Form: Abduction of Tolerable Productivity

(Belth et al 2021)

Abduction of Tolerable Productivity

Apply TSP **recursively**

- Given ***N*** items, do **enough** of them take **-x affix**?
  - If yes, **productive rule learnt!**
  - If not, **subdivide** into disjoint subsets & **recurse**

Abduction of Tolerable Productivity

Apply TSP **recursively**

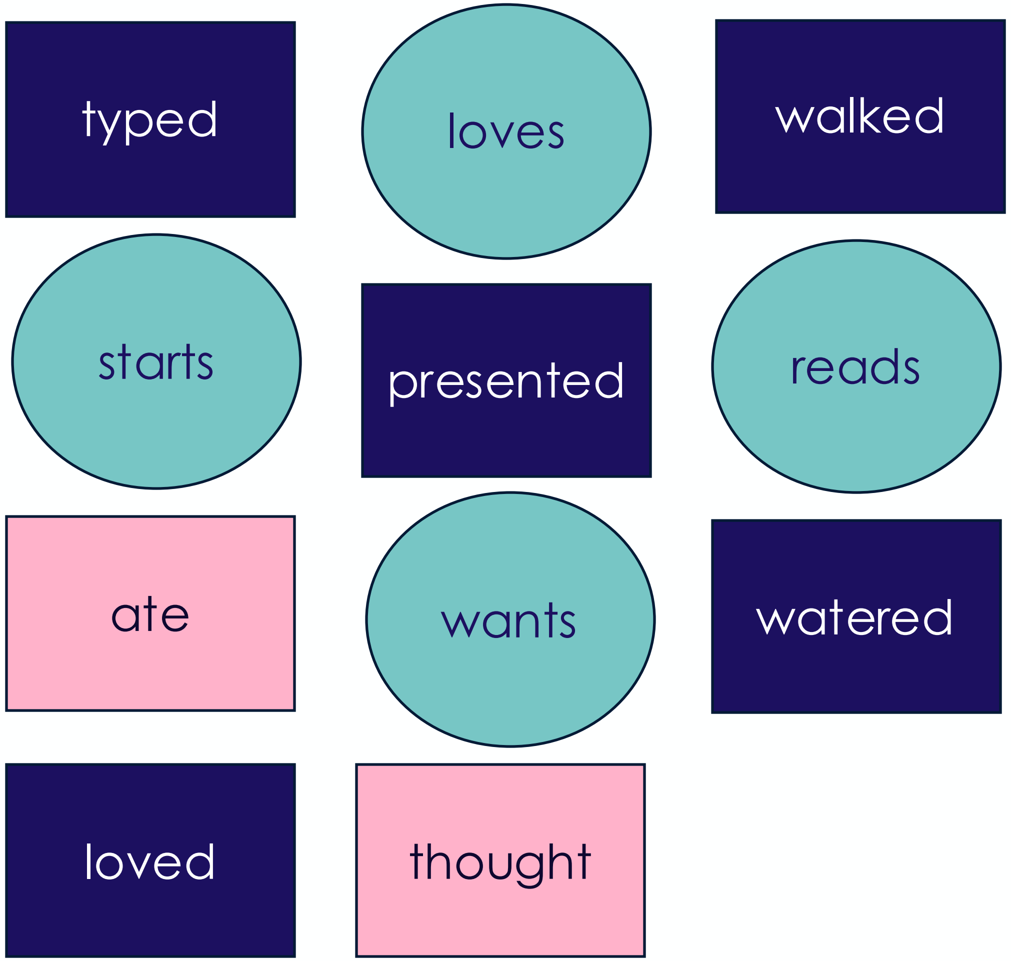
- Given ***N*** items, do **enough** of them take **-x affix**?
  - If yes, **productive rule learnt!**
  - If not, **subdivide** into disjoint subsets & **recurse**

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

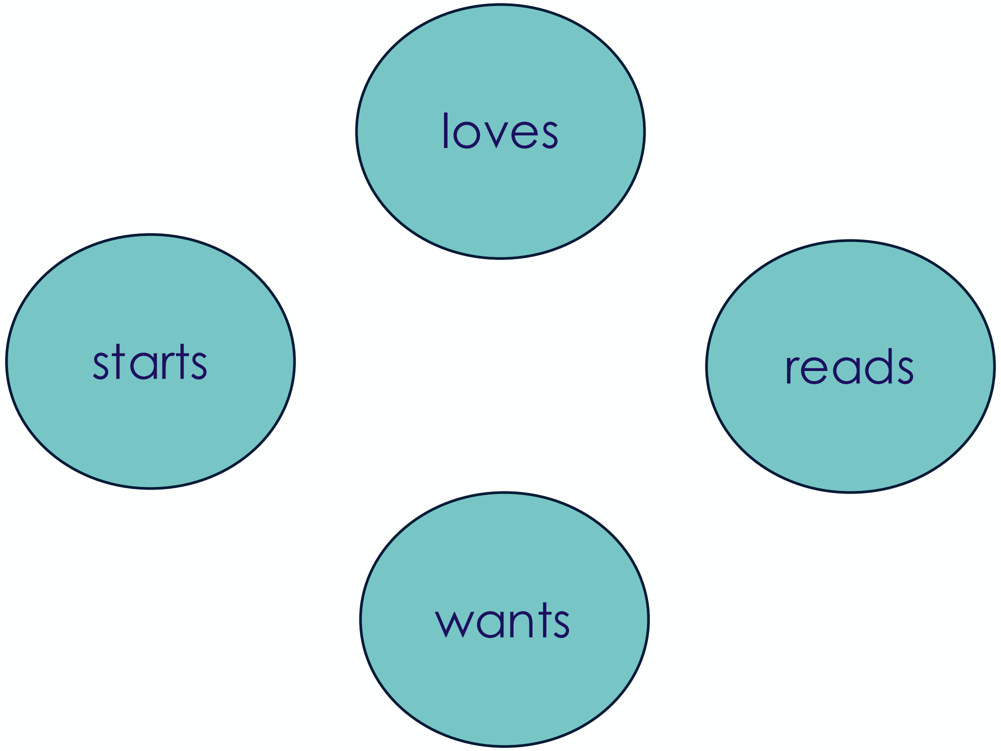
watered

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***

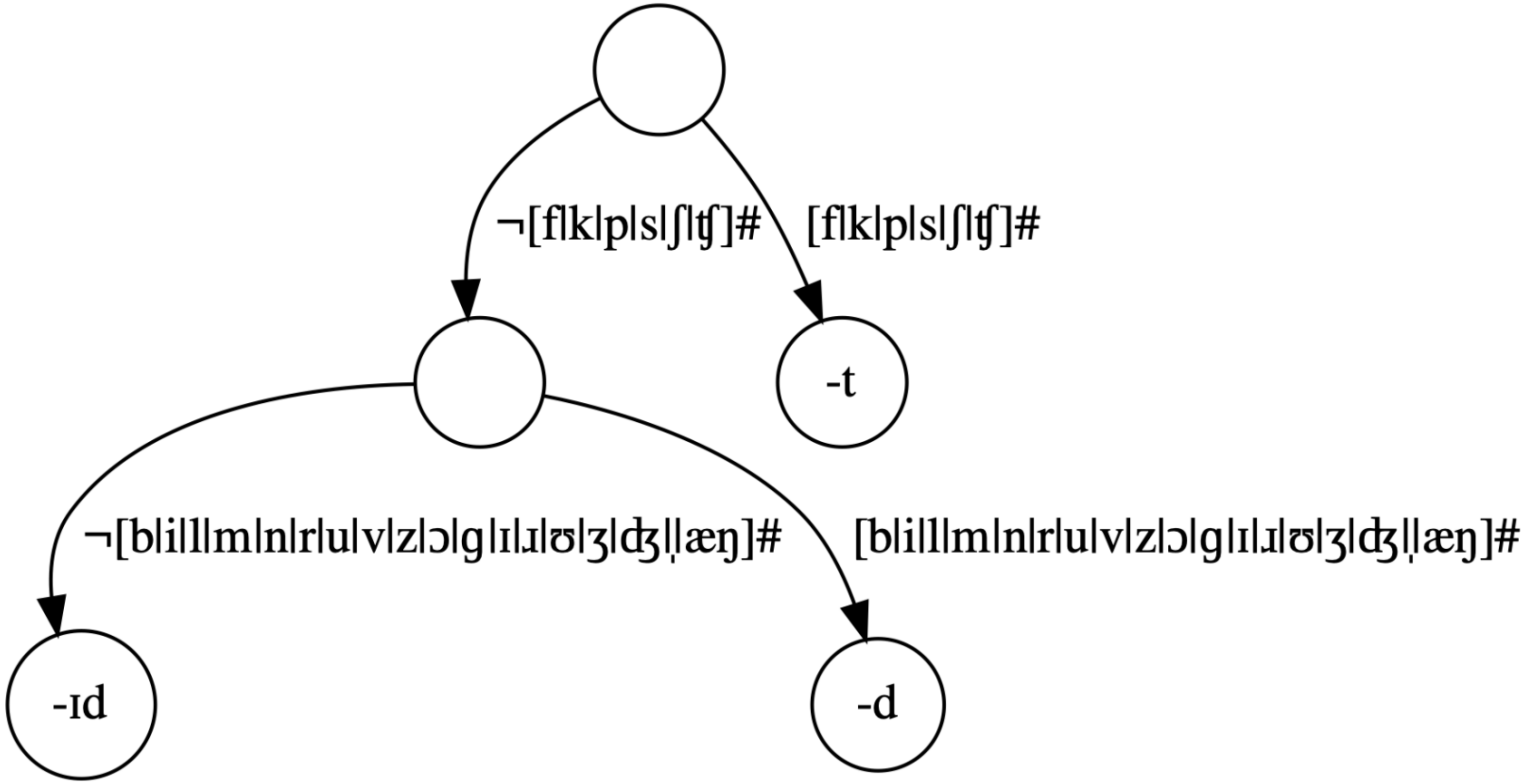
ATP Model: Toy Example



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

ATP Model: Sample learning trace

English past tense: morphophonological conditioning



ATP Model: Inflection and Generation

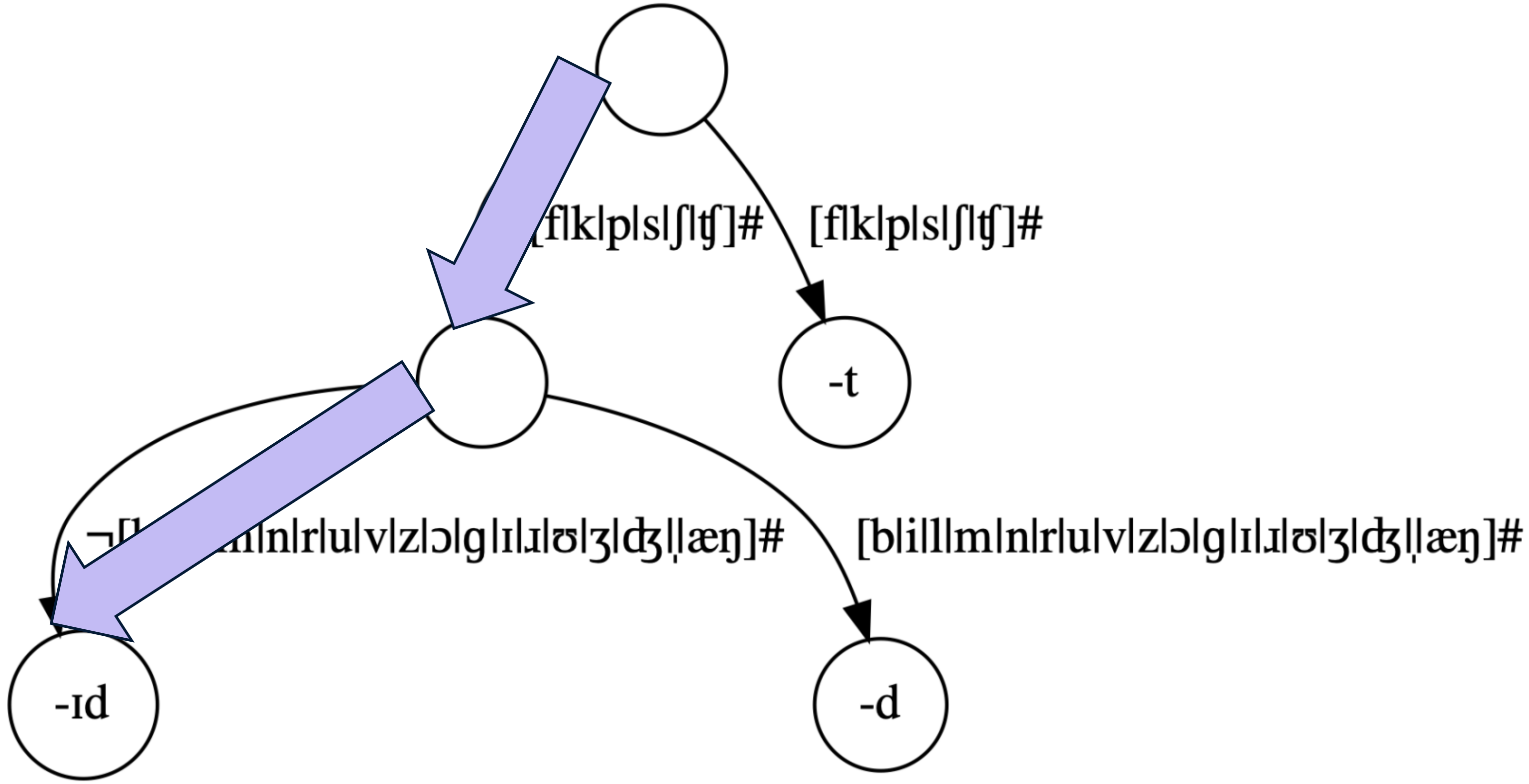
During test, given **novel forms & features** to inflect

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

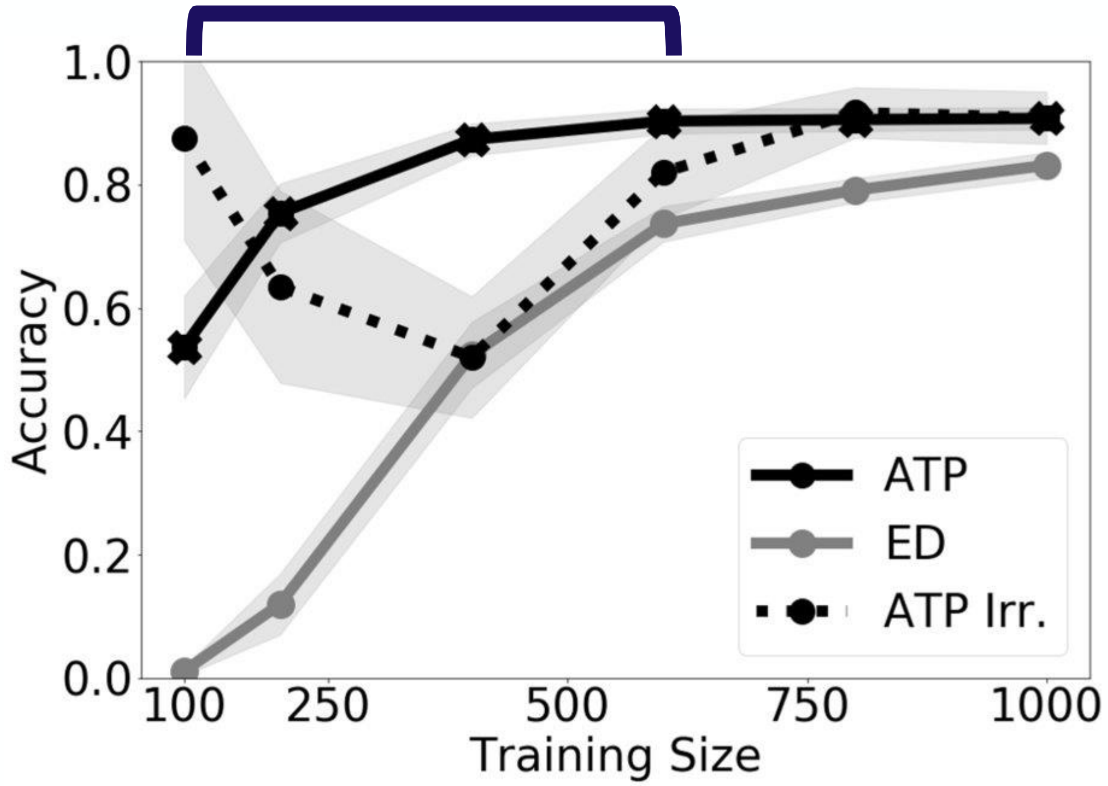
ATP Model: Sample learning trace

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

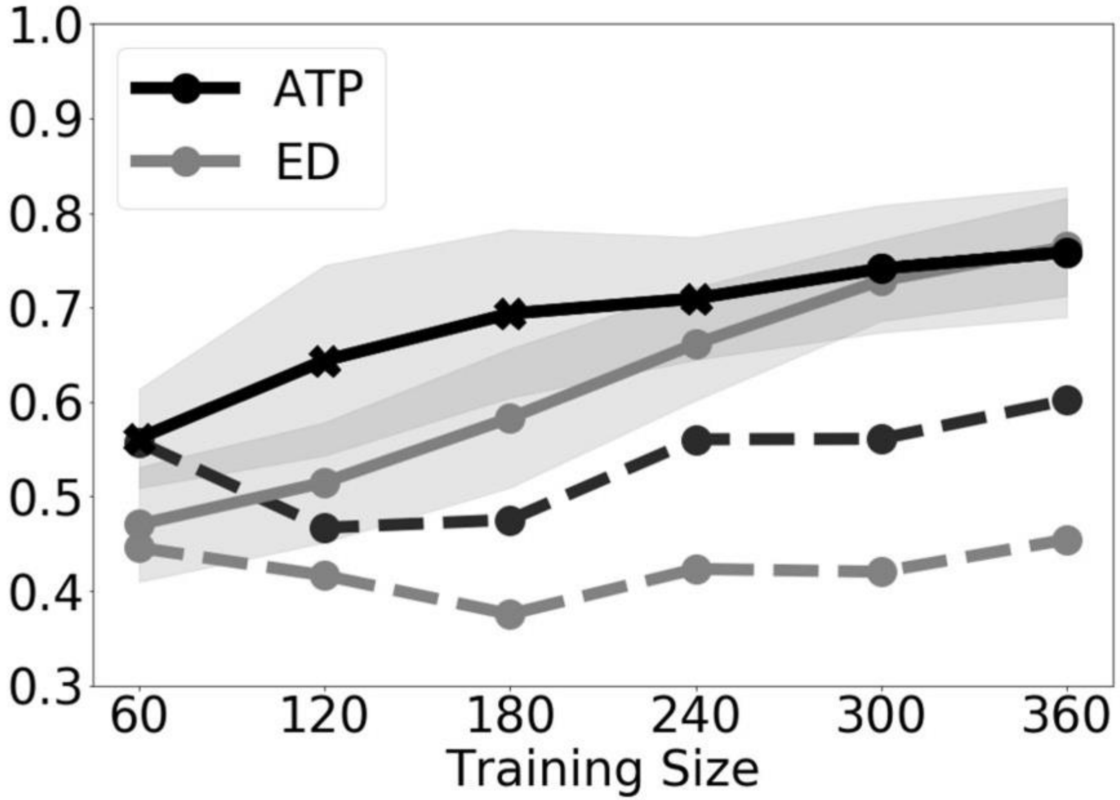
ATP: German Results

Relies less on gender than K&C

- **Solid lines** = gender info given at test
- **Dashed lines** = gender info not given at test

Trained on **plausible vocabulary**

- **400** inflected forms



(b) German

ATP: Summary

Children **overregularize**

- ***So does ATP!***

Children show **developmental regression** when learning some paradigms

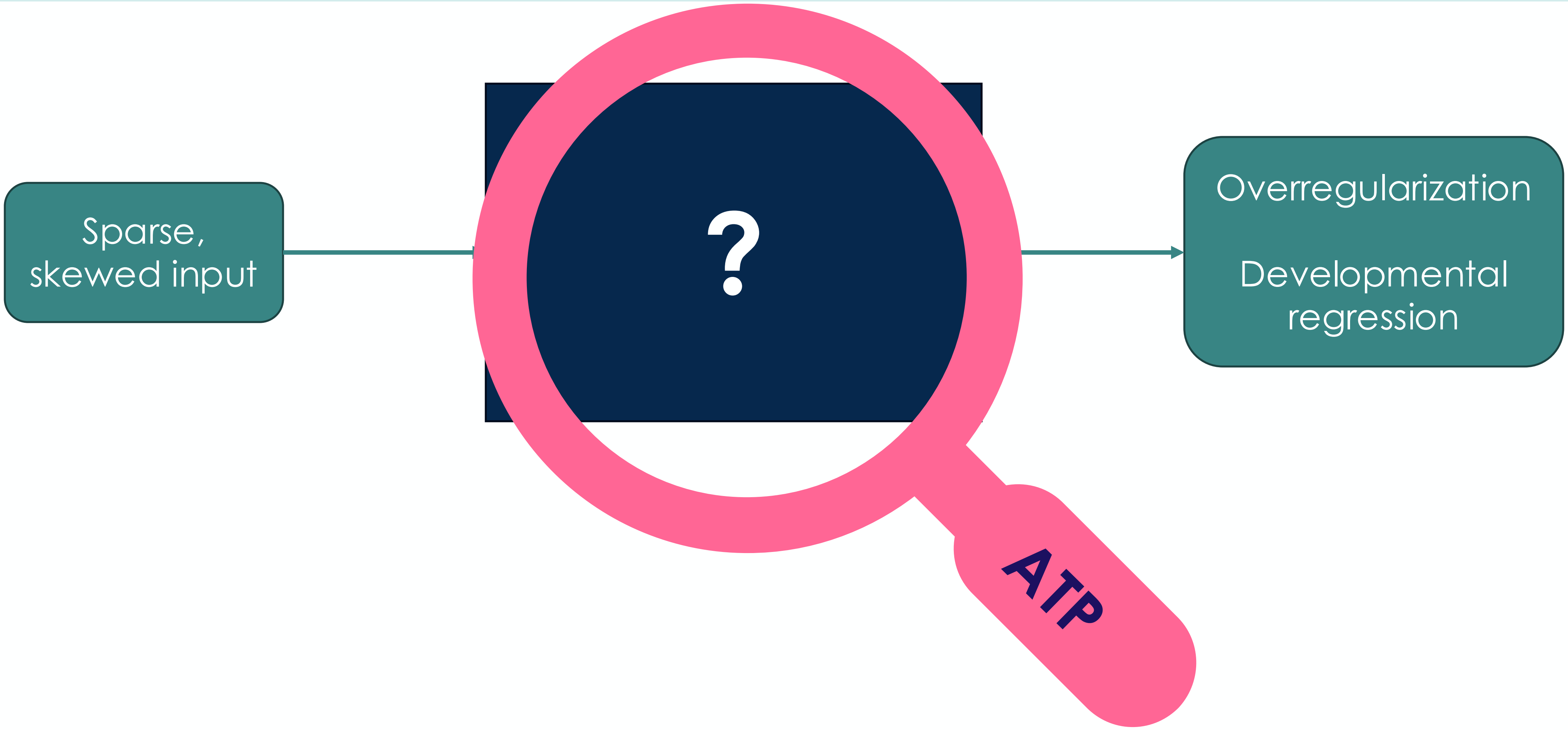
- ***So does ATP!***

Children learn from **extremely sparse, skewed input**

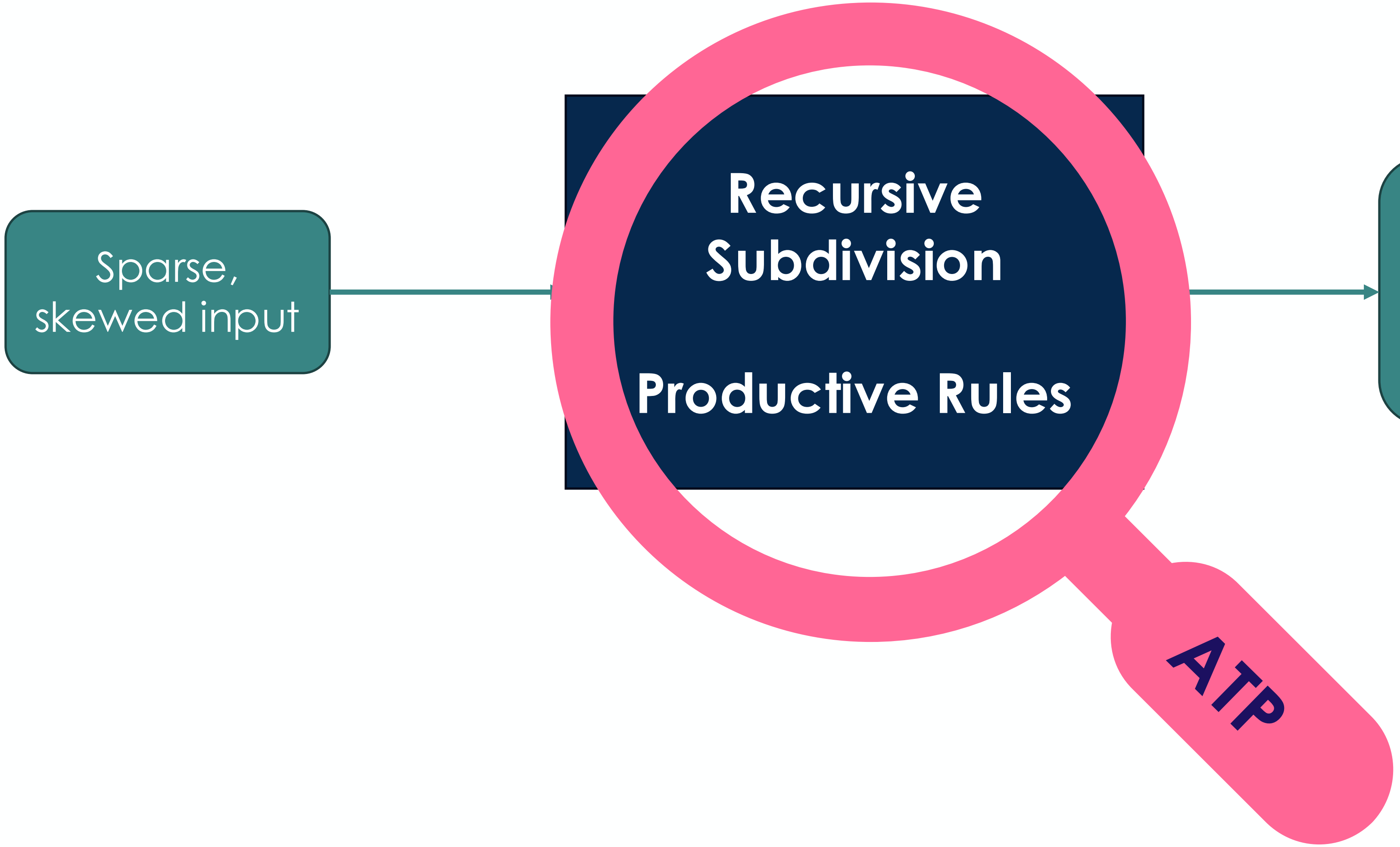
- ***So does ATP!***

ATP gives a mechanistic account of *why* these errors occur and what the resulting grammar looks like

ATP: Summary


Sparse, skewed input
?
ATP
Overregularization
Developmental regression

ATP: Summary


Sparse, skewed input
Recursive Subdivision
Productive Rules
ATP
Overregularization
Developmental regression

Learning which Features are Marked:
Sufficient Contrast Learner

(Payne 2022, 2023)

SCL Model: Preliminaries

Principle of Contrast: **distinct forms ⇒ distinct meanings**

- e.g. ***walk*** and ***walked*** must mean something different

Collisions: one lemma in multiple inflected forms

- e.g. ***walk-walked*** tells us that **±PAST** is marked

TSP: when are there enough collisions to learn marking?

- e.g. don't learn from ***am-are*** that **1 vs. 2** marked in English
- Learn from ***walk-walked, sing-sang***, etc. that **±PAST** marked

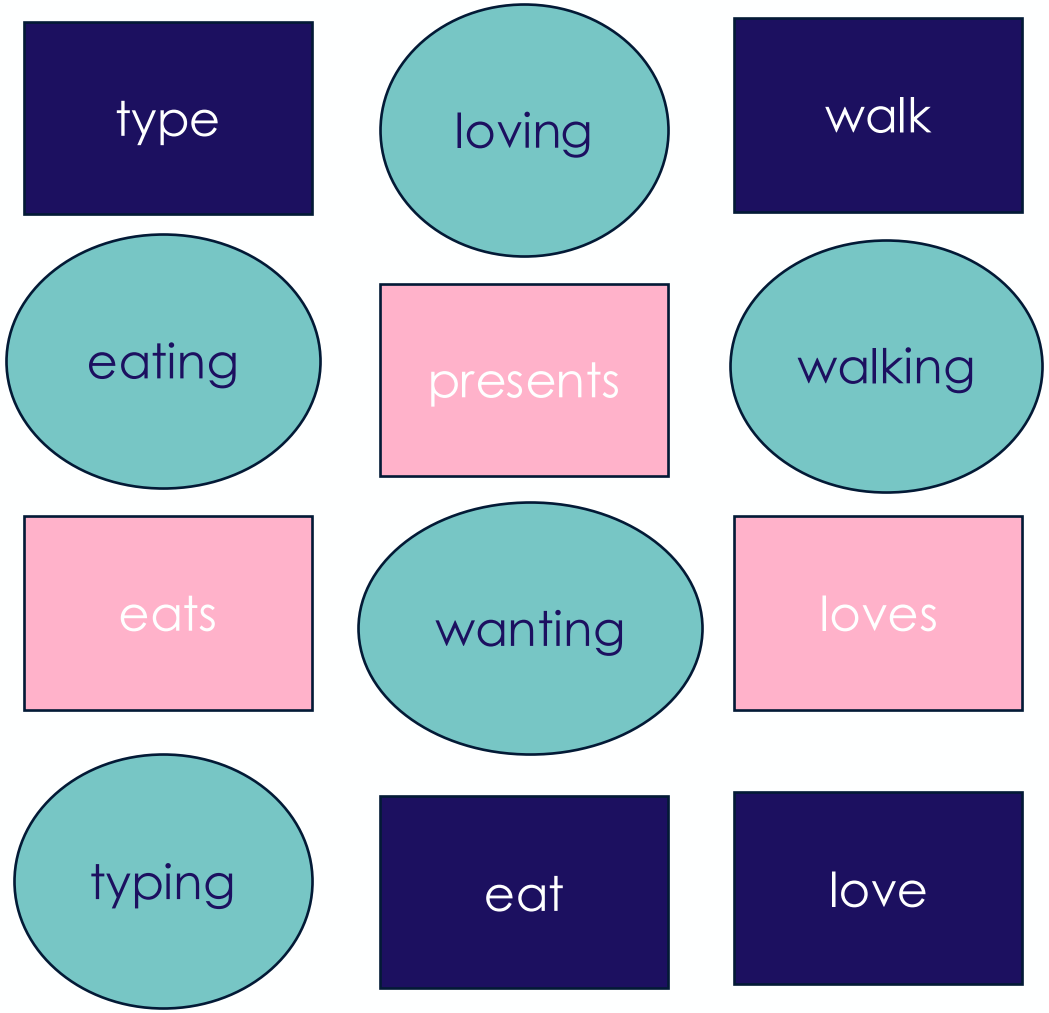
SCL Model: Collisions

Apply TSP **recursively** again!

- Input taken in **incrementally**
- When ***j*^th^ input** encountered, is there a collision?
- If so, do enough forms appearing in **inflection A** also appear in **inflection B** in a different form?
  - If yes, **productive contrast learnt!** Subdivide and recurse
  - If no, **continue** to take in input

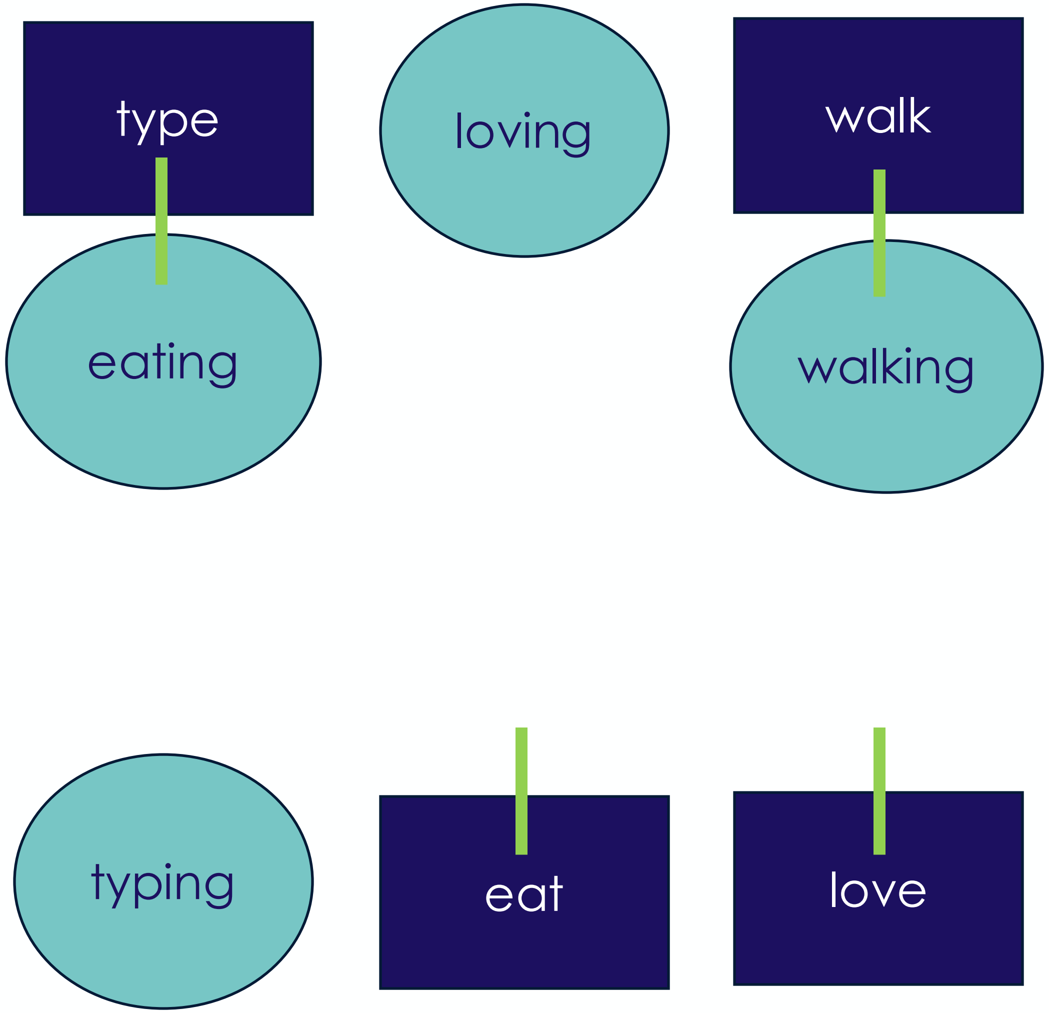
Apply to **English verbs, German noun plurals, Spanish verbs**, and **Hebrew verbs**

SCL Model: Toy Example



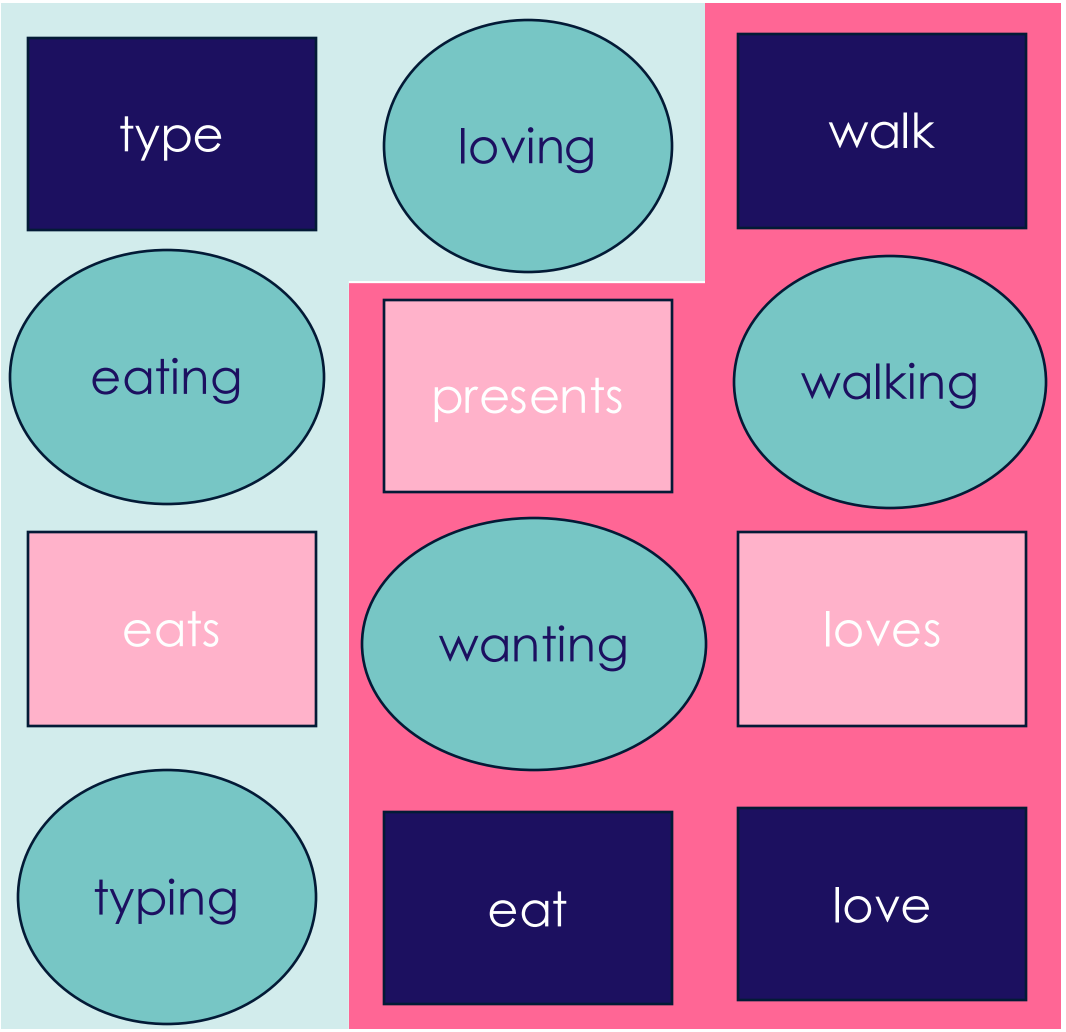
- Collision: ***walk-walking***
- **±PARTICIPLE** marked?

SCL Model: Toy Example



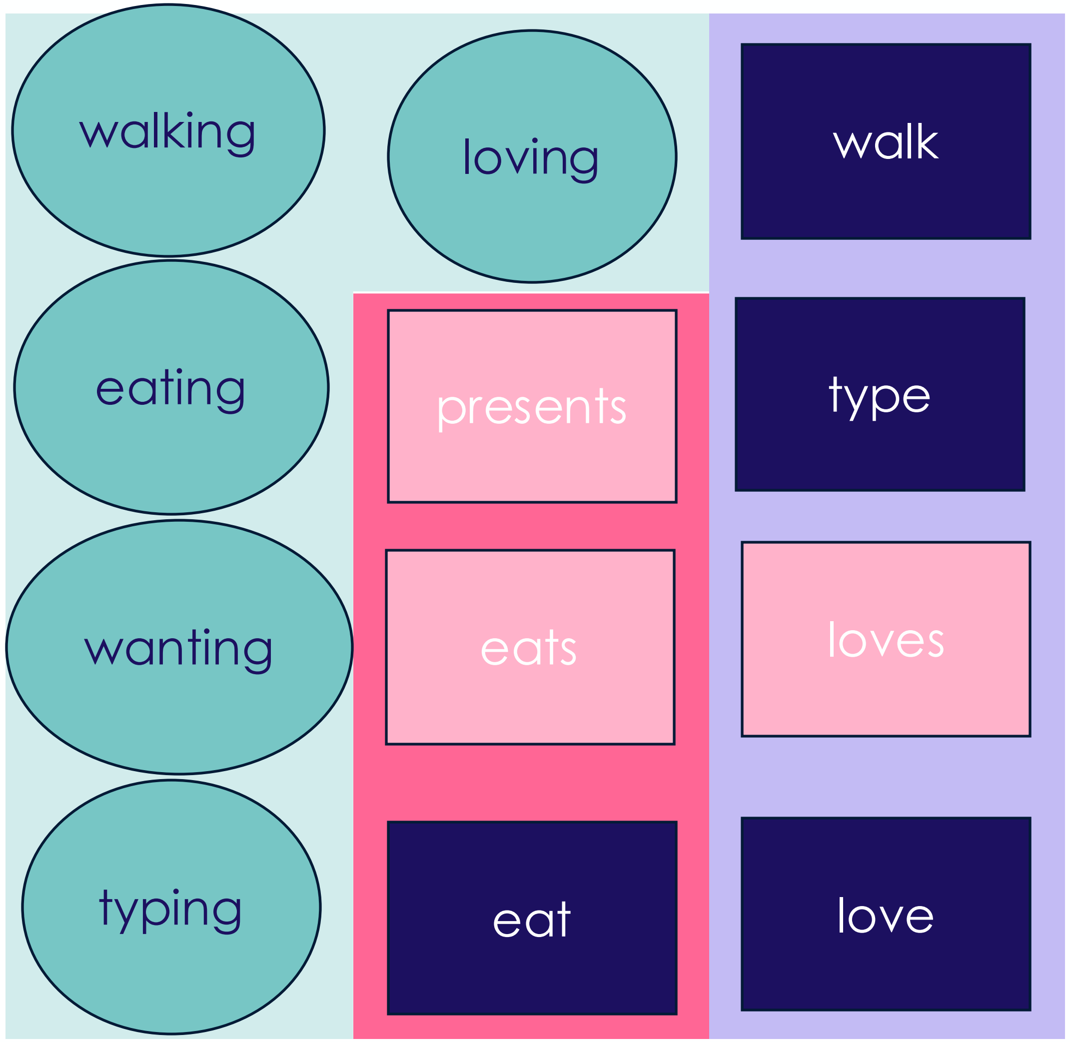
- Collision: ***walk-walking***
- **±Participle** marked?
  - 5 participles, 4 collisions (not ***wanting***)
  - $N - M = 1 < \theta_5 = 3$ ✅
- Contrast 1 productive! **±Participle marked**

SCL Model: Toy Example



- Collision: ***walk*-*walking***
- ±Participle marked?
  - 5 participles, 4 collisions (not ***wanting***)
  - $N - M = 1 < \theta_5 = 3$
- Contrast 1 productive! **±Participle marked**
- **Subdivide** into +Participle and –Participle forms

SCL Model: Toy Example



- Collision: ***walk-walking***
- **±PARTICIPLE** marked?
  - 5 participles, 4 collisions (not ***wanting***)
  - $N - M = 1 < \theta_5 = 3$ ✅
- Contrast 1 productive! **±PARTICIPLE marked**
- **Subdivide** into **+PARTICIPLE** and **–PARTICIPLE** forms
- Recursively learn that **± 3SG** marked

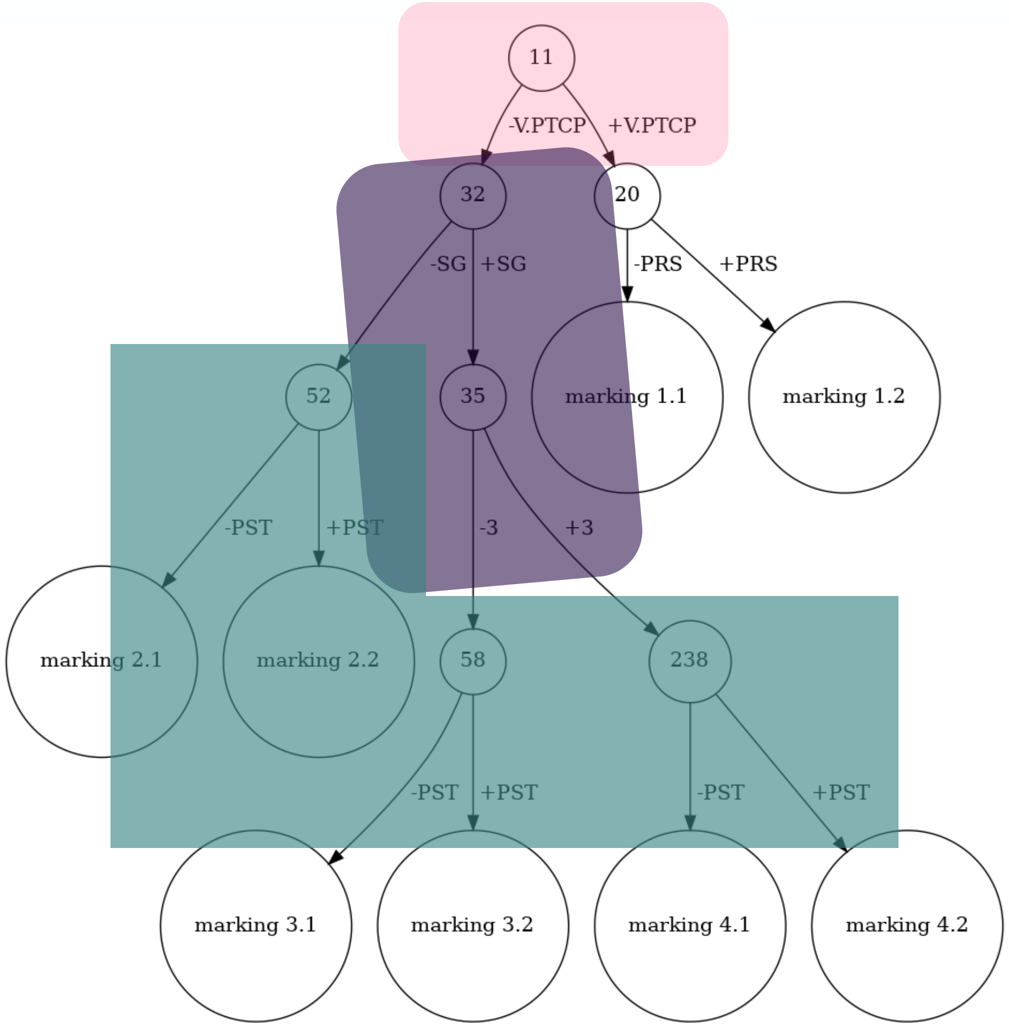
SCL: English Results

Plausible **order of acquisition**

1. **Participle** (*–ing*)
2. **3Sg** (*-s*)
3. **Past** (*-ed*)

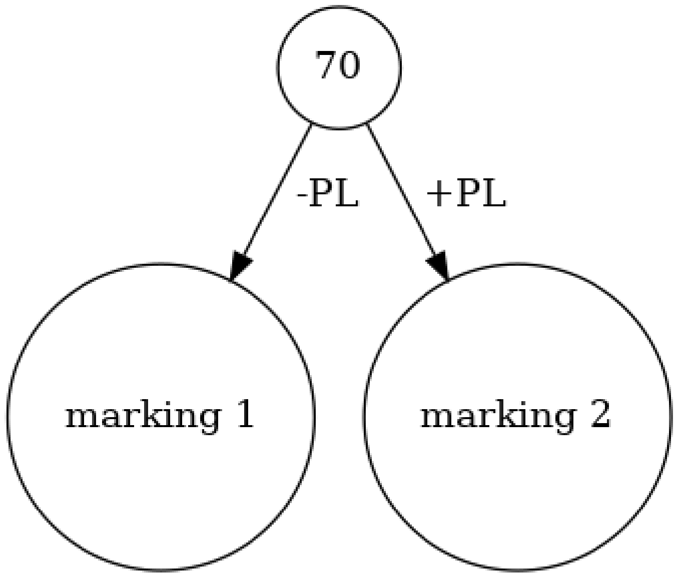
Plausible vocabulary size:

- **112** lemmas
- **238** inflected forms



SCL: German results

- Plausible vocabulary size:
  - **66** lemmas
  - **70** inflected forms
- Well under vocab size at which plural affix overapplication begins



SCL: Spanish Results

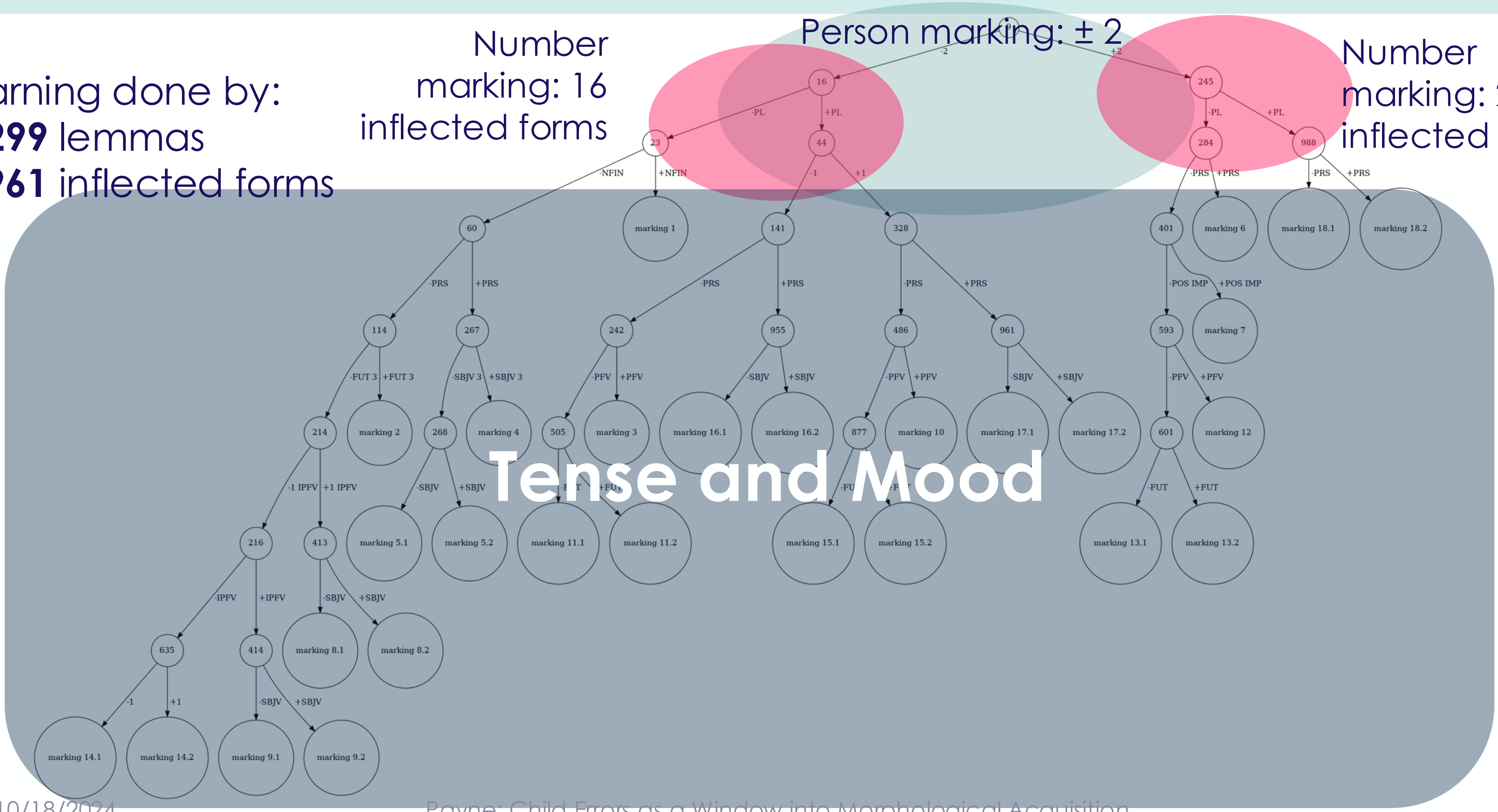
- Spanish order of acquisition:
  - **Finiteness & person marking**: 1;7
  - **Number marking**: 1;7-2;0
    - Second plural emerges later than other agreements in many learners
  - **Tense**: 2;0-2;2
  - **Mood**: 1;7-2;2

(Montrul 2004)

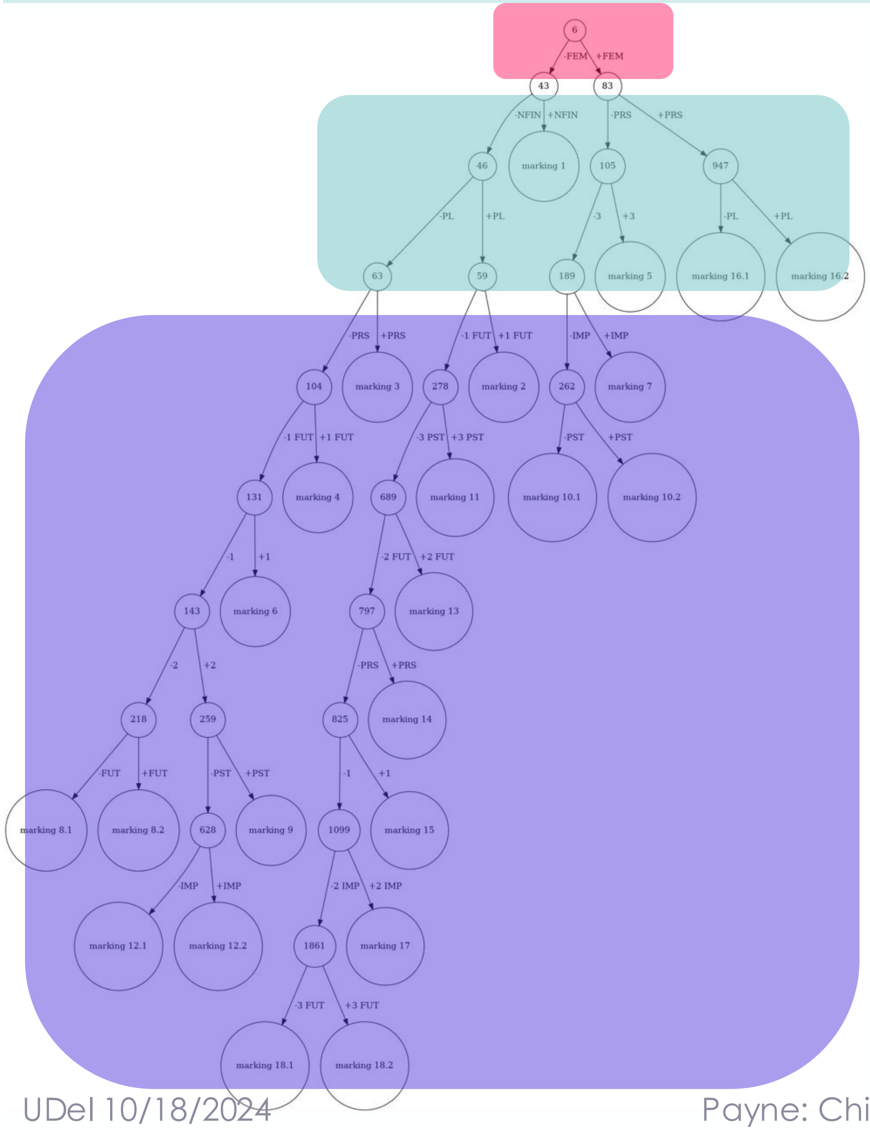
SCL: Spanish Results

Learning done by:

- **299** lemmas
- **961** inflected forms



SCL: Hebrew Results



Hebrew order of acquisition:

- **Person, number, gender** before tense
- **Person vs. number** varies
- **Gender** appears before or at the same time as **number**

Our model:

- Order of acquisition:
  - **Gender**, **person & number**, **tense**
- Vocab size:
  - **323** lemmas
  - **1861** inflected forms

SCL Implications: Root Infinitives

Root Infinitive (RI) Stage: stage of omission errors

Cross-linguistically, **“richer” morphology ⇒ shorter RI stage**

Richer morphology also means **more subdivision**

- TSP tolerates more exceptions for **smaller *N***
- More subdivision ⇒ **smaller *N***
- Smaller *N* ⇒ **quicker learning** of inflectional categories

SCL gives a **mechanistic account** of cross-linguistic differences

(Philips 1995, Legate & Yang 2007)

SCL: Summary

Children **omit affixes**

- ***SCL gives an account for why!***

Children show clear **order of acquisition effects**

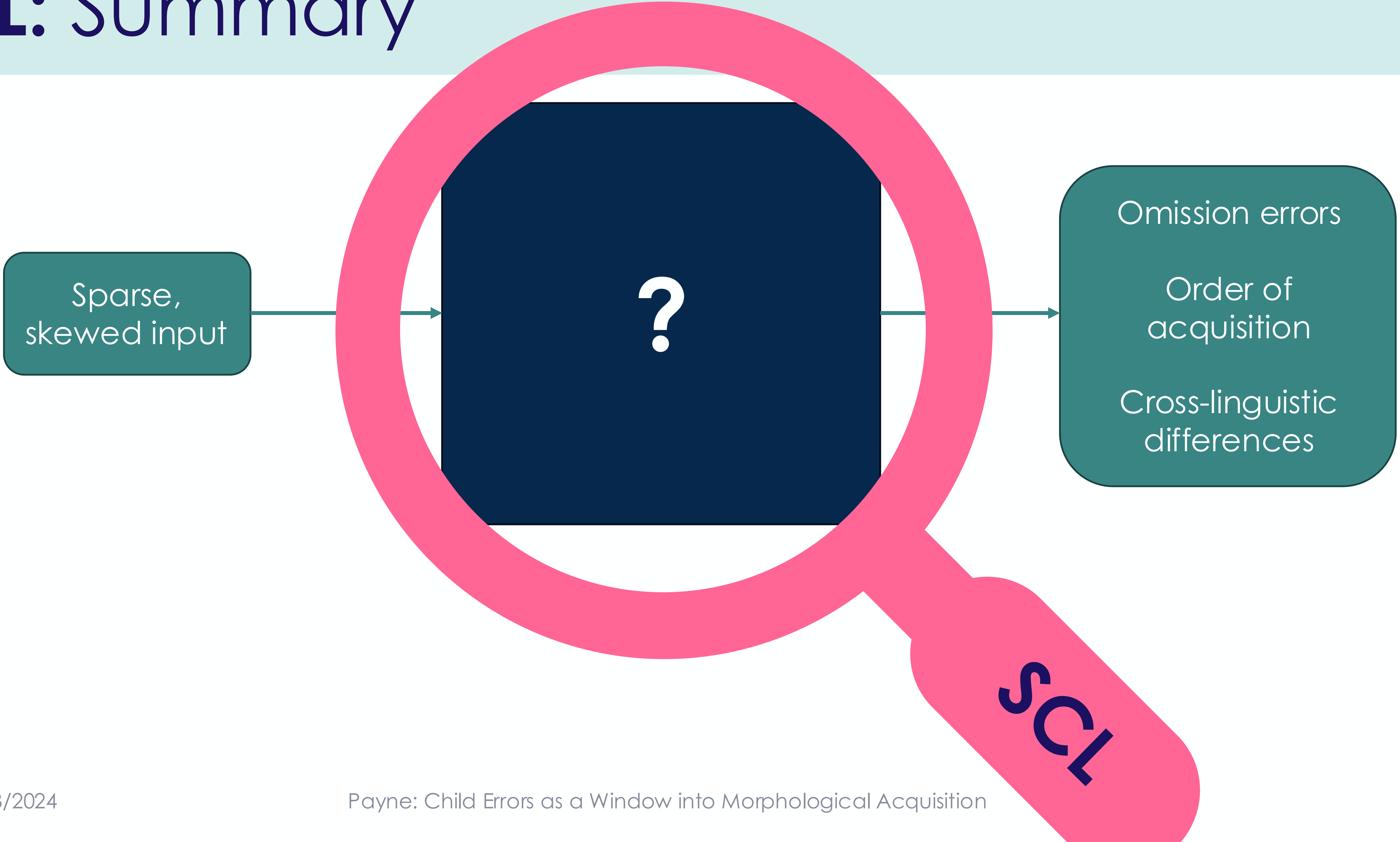
- ***So does SCL!***

Children learn from **extremely sparse, skewed input**

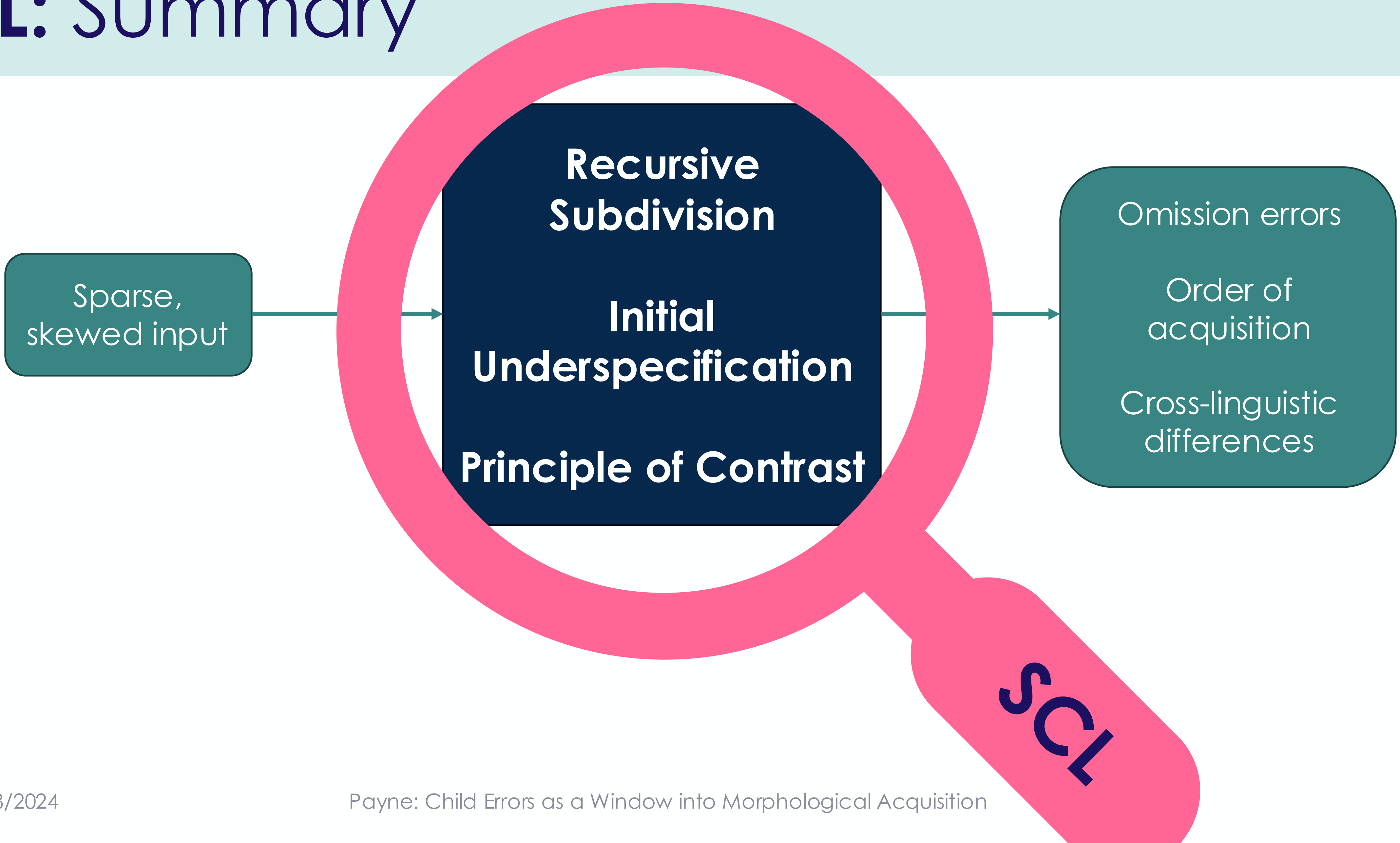
- ***So does SCL!***

SCL gives mechanistic account of order of acquisition, omission errors and cross-linguistic differences in acquisition

SCL: Summary


Sparse, skewed input
?
SCL
Omission errors
Order of acquisition
Cross-linguistic differences

SCL: Summary


Sparse, skewed input
Recursive Subdivision
Initial Underspecification
Principle of Contrast
Omission errors
Order of acquisition
Cross-linguistic differences
SCL

Discussion & Future Directions

What Makes a Good Model?

Input:

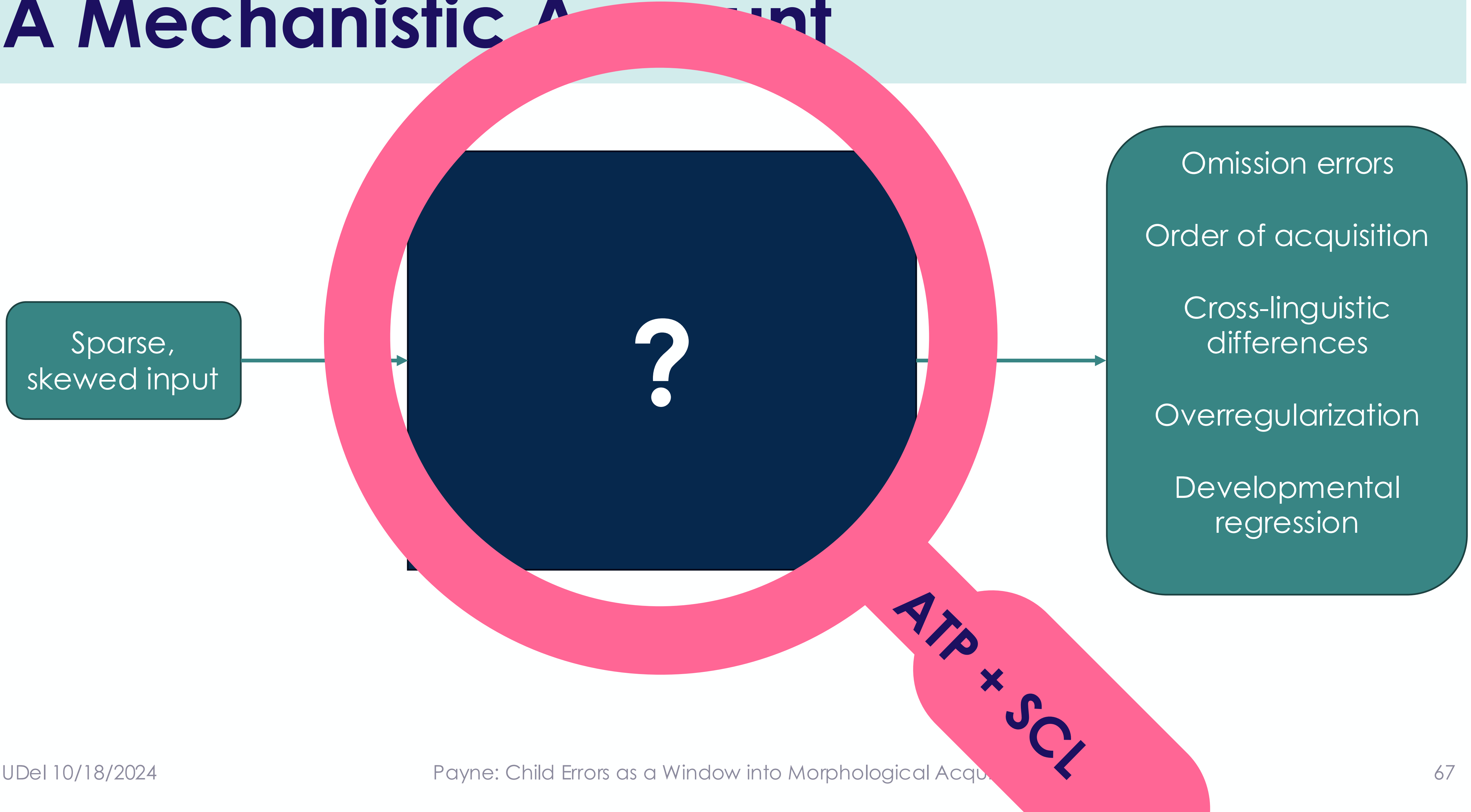
- ***Small*** vocabulary
- ***Sparse*** paradigms

Errors:

- ***Omissions***, not substitutions
- ***Over-regularizations***, not over-irregularizations
- Developmental ***regression***

Interpretability: why does it do what it does?

A Mechanistic A… unt



A Mechanistic A unt

Sparse, skewed input

Recursive Subdivision

Productive Rules

Initial Underspecification

Principle of Contrast

ATP + SCL

Omission errors

Order of acquisition

Cross-linguistic differences

Overregularization

Developmental regression

Conclusion: Getting the Right Stuff Wrong

	NNs	ATP	SCL
Learn from plausible data	❌	✅	✅
Account for over-regularization and developmental regression	❌	✅	---
Account for omission and the RI stage	❌	---	✅
Interpretability	❌	✅	✅

Future Work

SCL learns the features required by **ATP**

- **Combination** of these two learning strategies
- First learn the inflectional classes, then map them to form

Expand ATP:

- Handle **templatic & agglutinative** morphology

Expand SCL:

- Explore model **subdivision predictions**
- Learn features from **distributional information**

Thank you!!

Caleb Belth
University of Utah

Deniz Beser
ISI

Jodan Kodner
Stony Brook

Charles Yang
UPenn

Special thanks to **Charles Yang**, **Jordan Kodner**, **Caleb Belth**, **Jeff Heinz**, **Julie Anne Legate**, **Mark Aronoff**, and **Bob Berwick.**

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

Background: Earliest Stages

Early Segmentation ≠ Early Understanding of Use

Early Segmentation

- English learners:
  - **0;6:** segment ***–s***
  - **0;8:** segment ***–ing***
  - **Both:** not ***pseudomorphemes***
- French learners (0;11):
  - Segment **-e** (infinitive & pastp)
  - Don't segment ***pseudomorpheme*** *-u*

≠

Early Understanding of Use

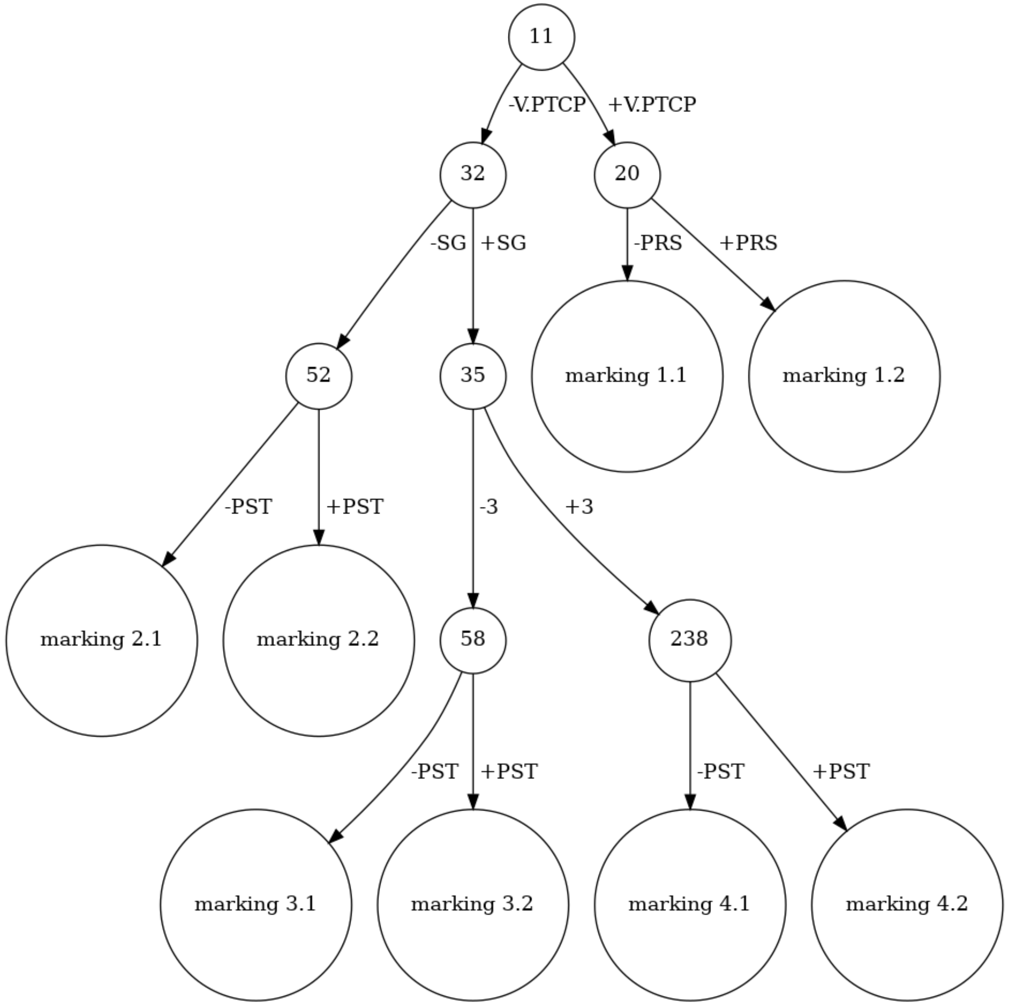
- English-learners:
  - **1;7:** can't use **–s** as cue to subject number
  - **5;0:** use **–s** as a cue to subject number in comprehension

(Kim & Sundara 2021, Marquis & Shi 2012, Soderstrom et al 2002, Johnson et al 2005)

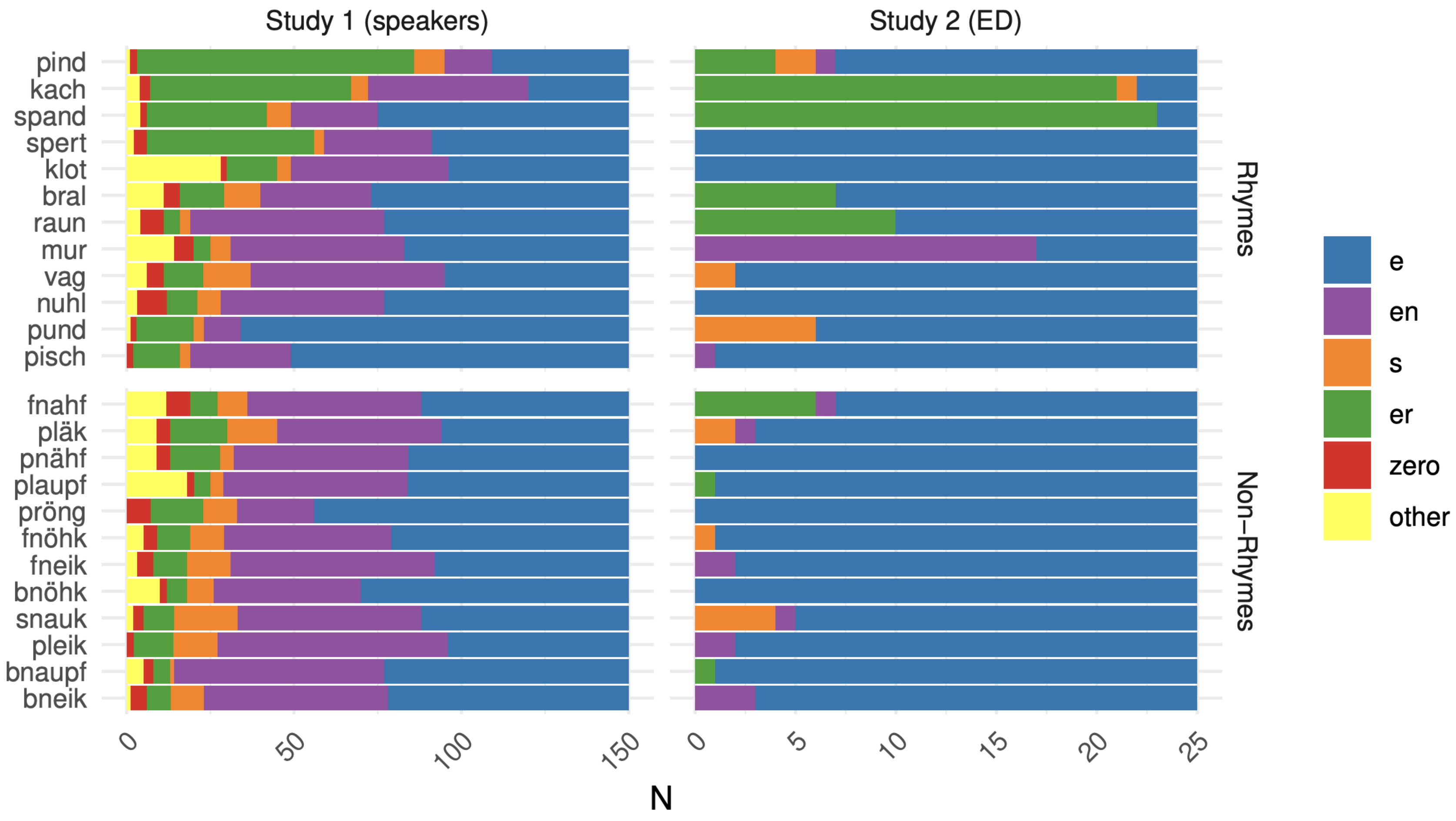
SCL: English Results

Learning past tense separately for each agreement?

- **Yang, Ellman, and Legate (2015):** past tense acquired later for learners of AAE
- **Difference in input** = agreement, not tense
- TSP tolerates ***relatively fewer exceptions*** for larger *N*

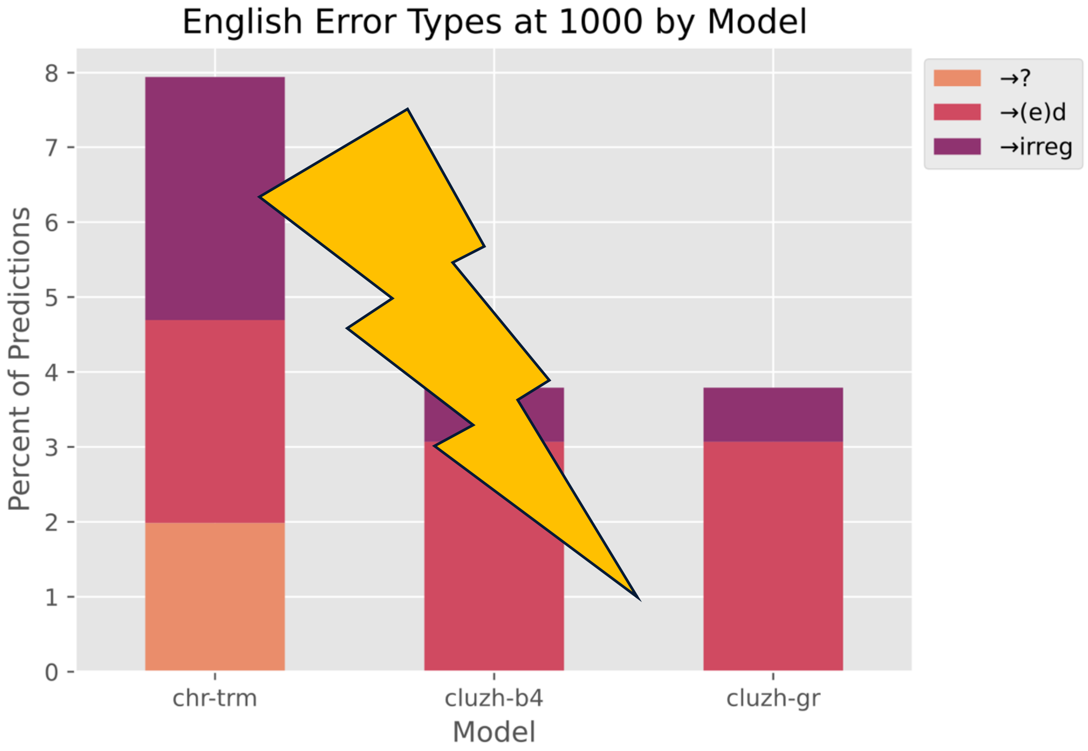


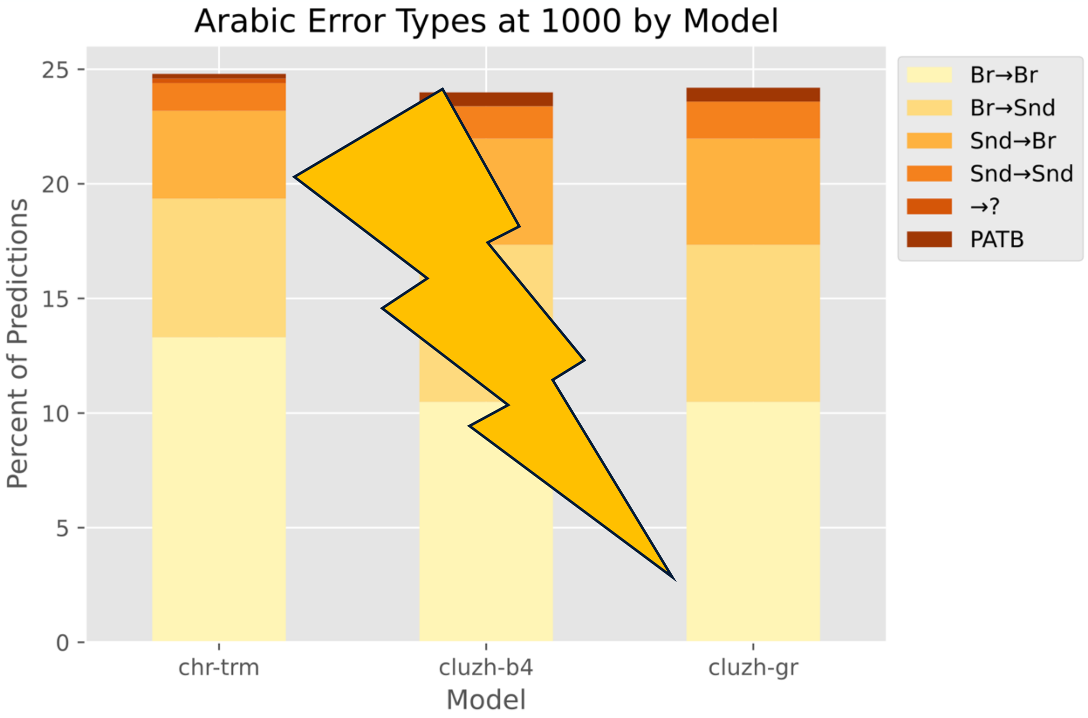
German Noun Plurals: We really aren't there



Other Neural Models: Still not there!

- Ongoing work with **Salam Khalifa, Jordan Kodner & Zoey Liu:** test ***more models*** on ***more paradigms,*** find ***same problems***

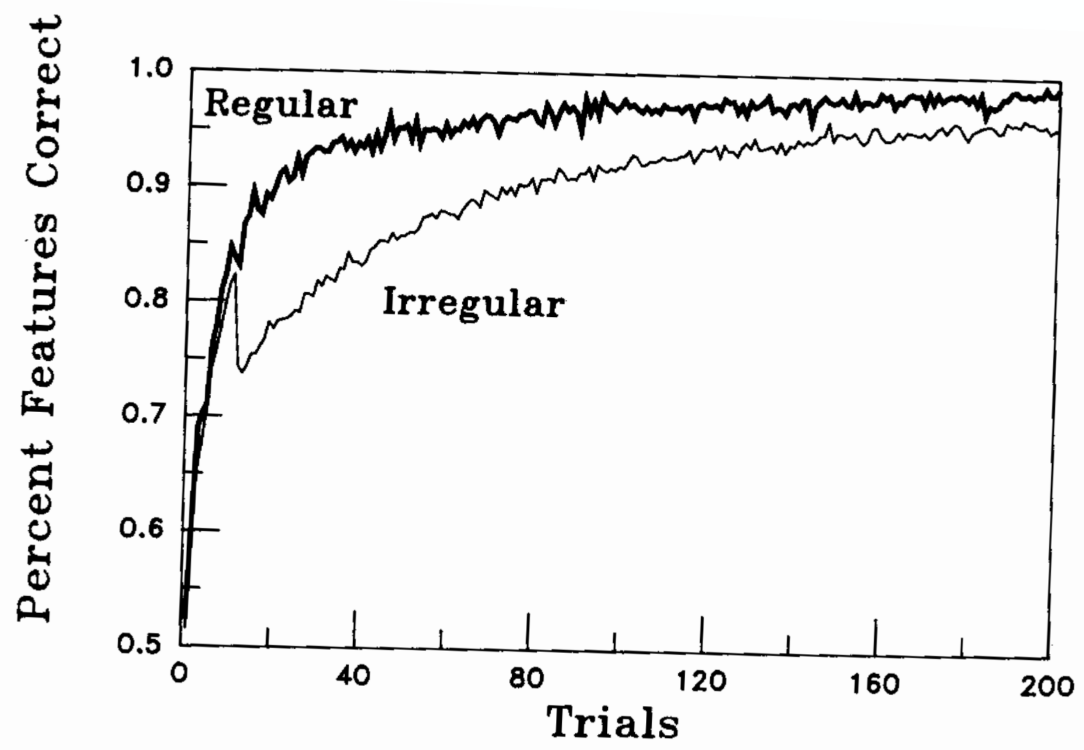




Background: The Past Tense Debate

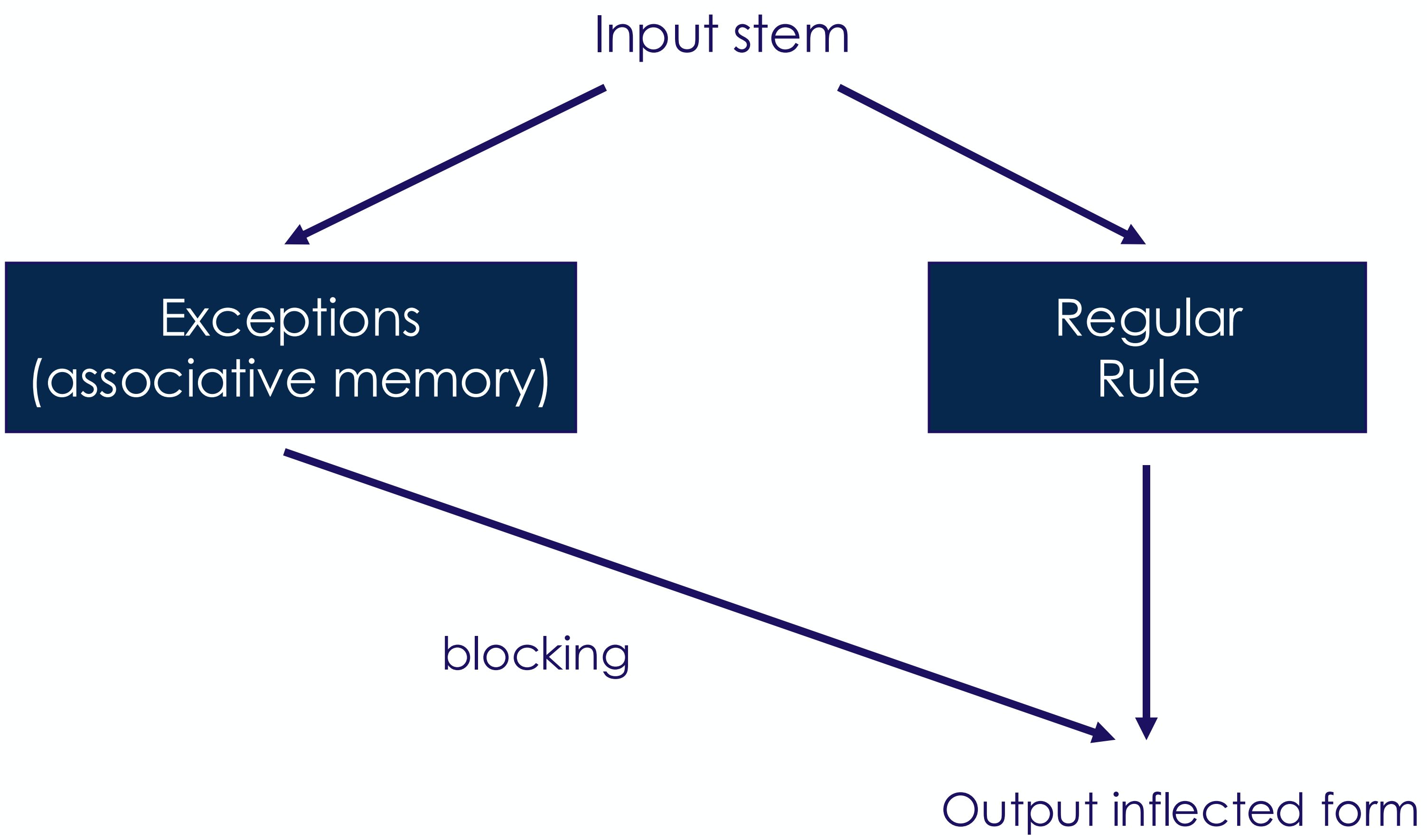
- **Rumelhart & McClelland:** ***single-route, connectionist*** model can:
  - Exhibit ***developmental regression***
  - Exhibit ***overregularization***

  ∴ **Rule-like behavior!**



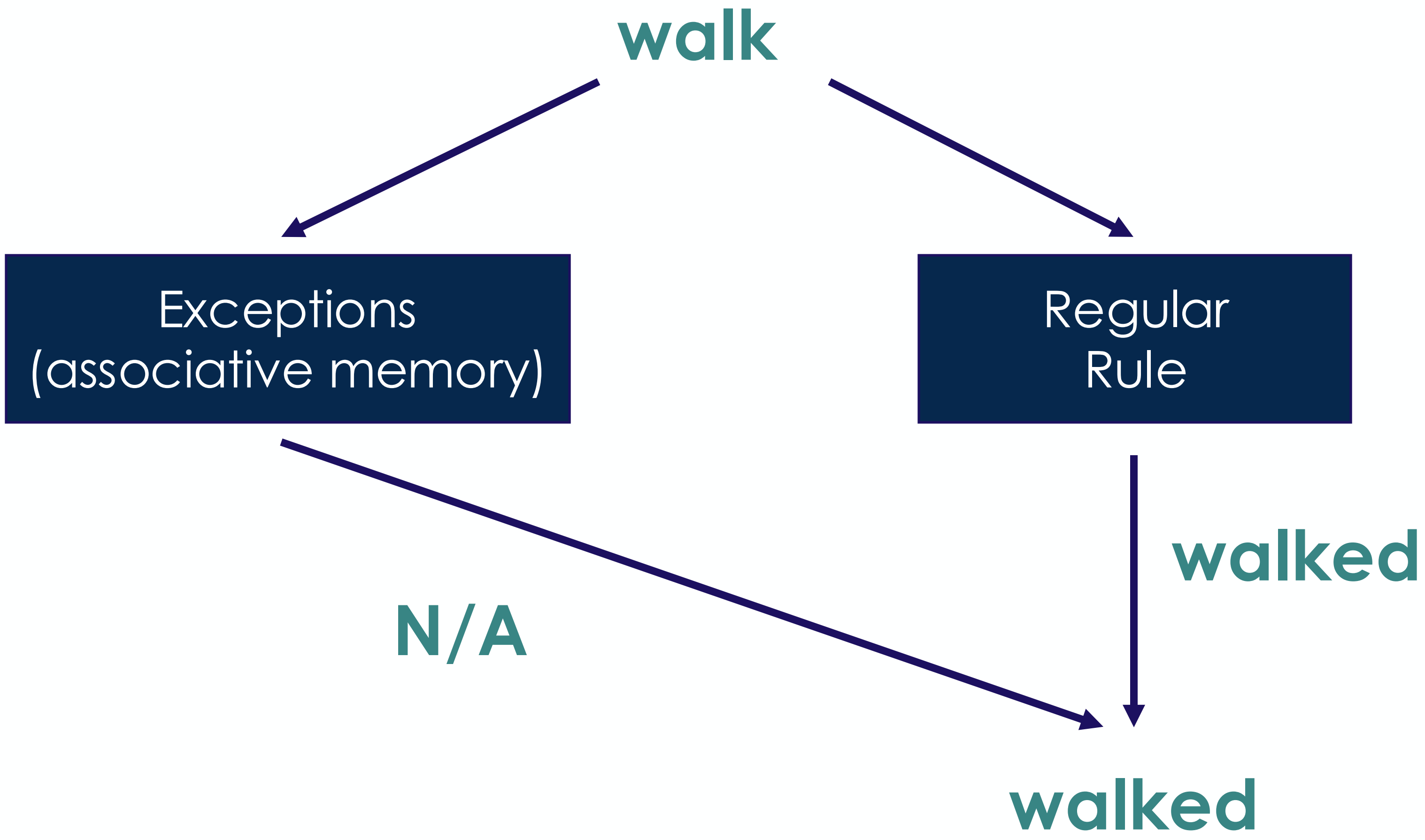
Background: The Past Tense Debate

Pinker & Prince's *dual-route* model:



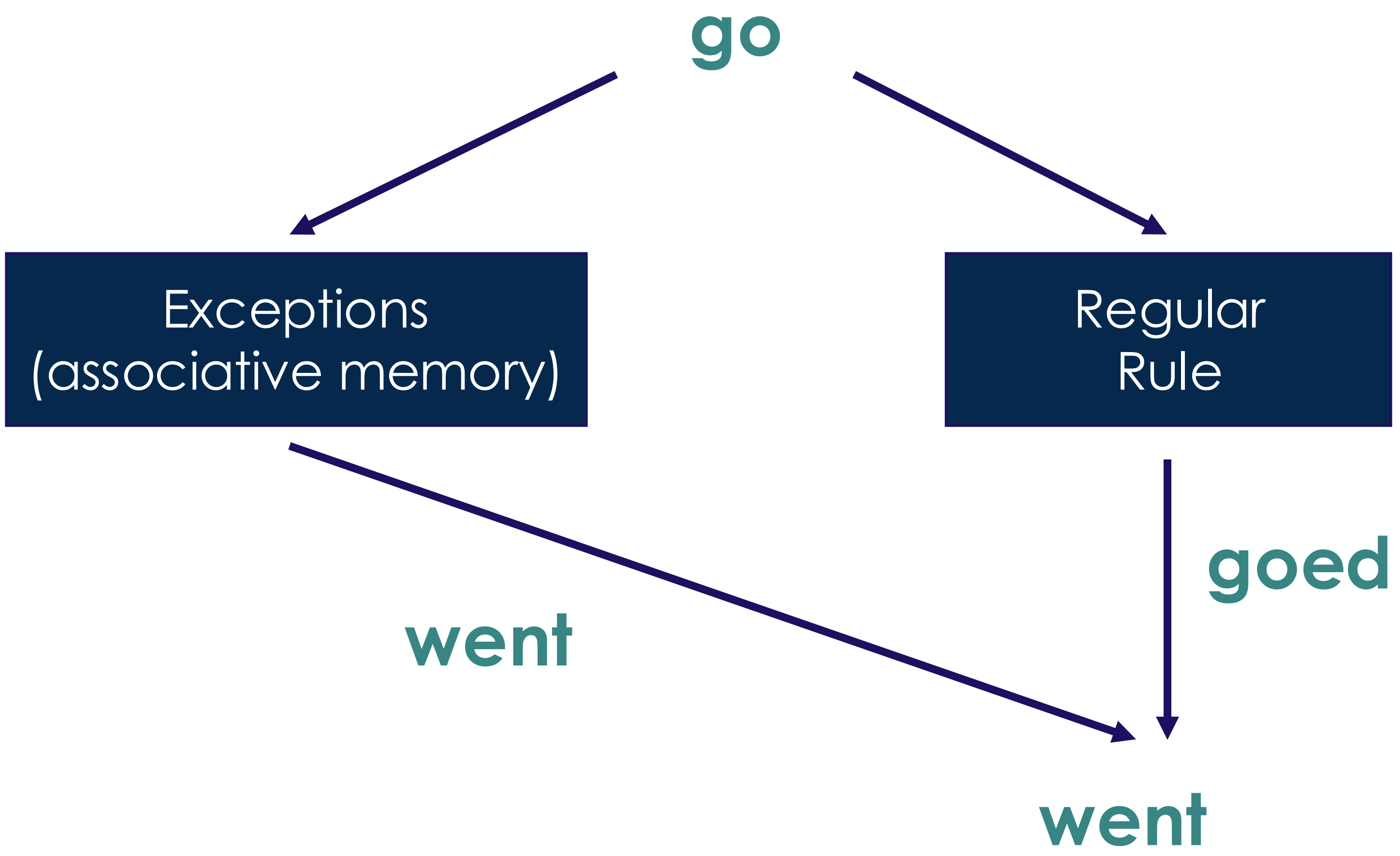
Background: The Past Tense Debate

Pinker & Prince's *dual-route* model: regular inflection



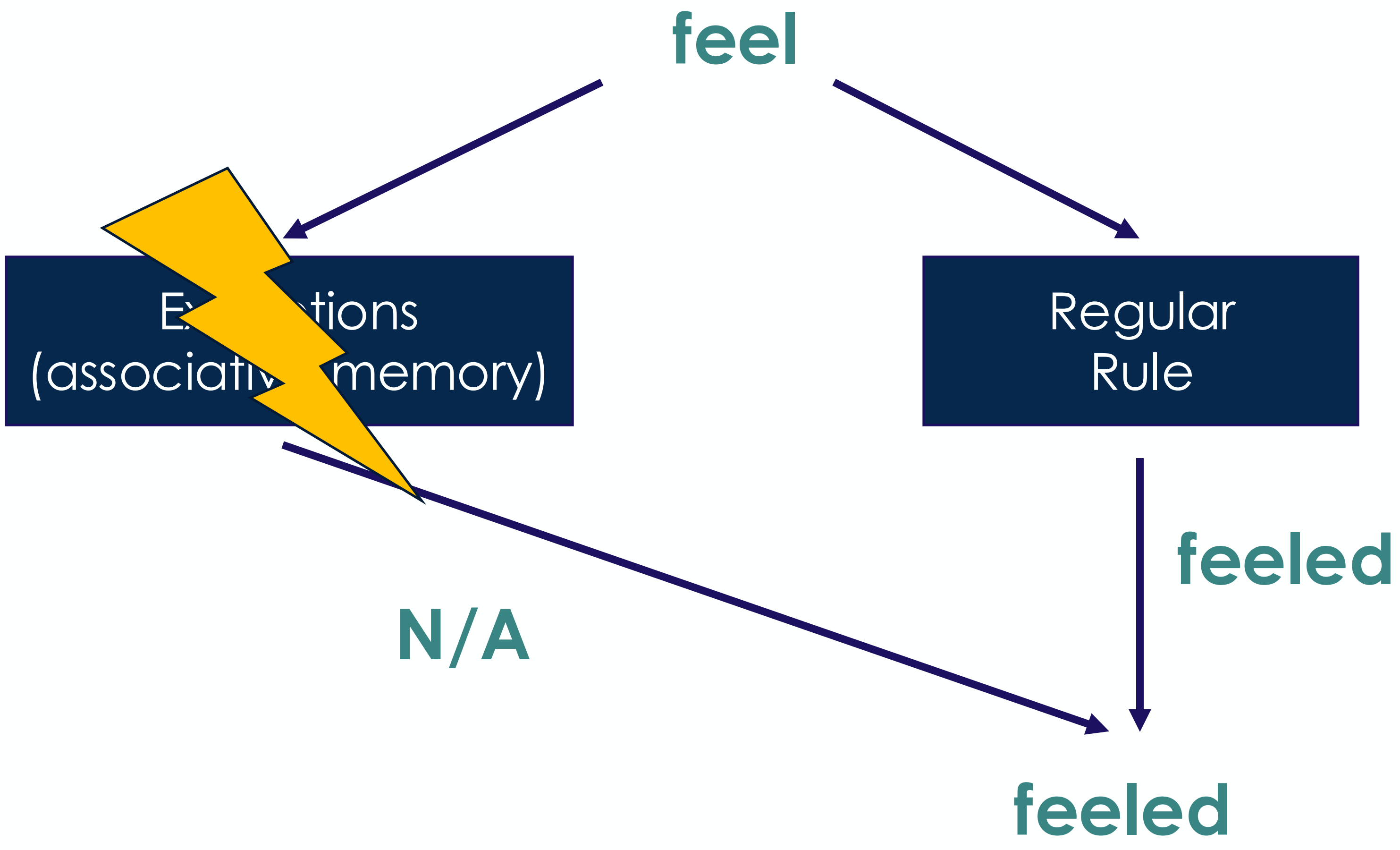
Background: The Past Tense Debate

Pinker & Prince's ***dual-route*** model: irregular inflection



Background: The Past Tense Debate

Pinker & Prince's ***dual-route*** model: overregularization



Background: The Past Tense Debate

Pinker & Prince's ***dual-route*** model:

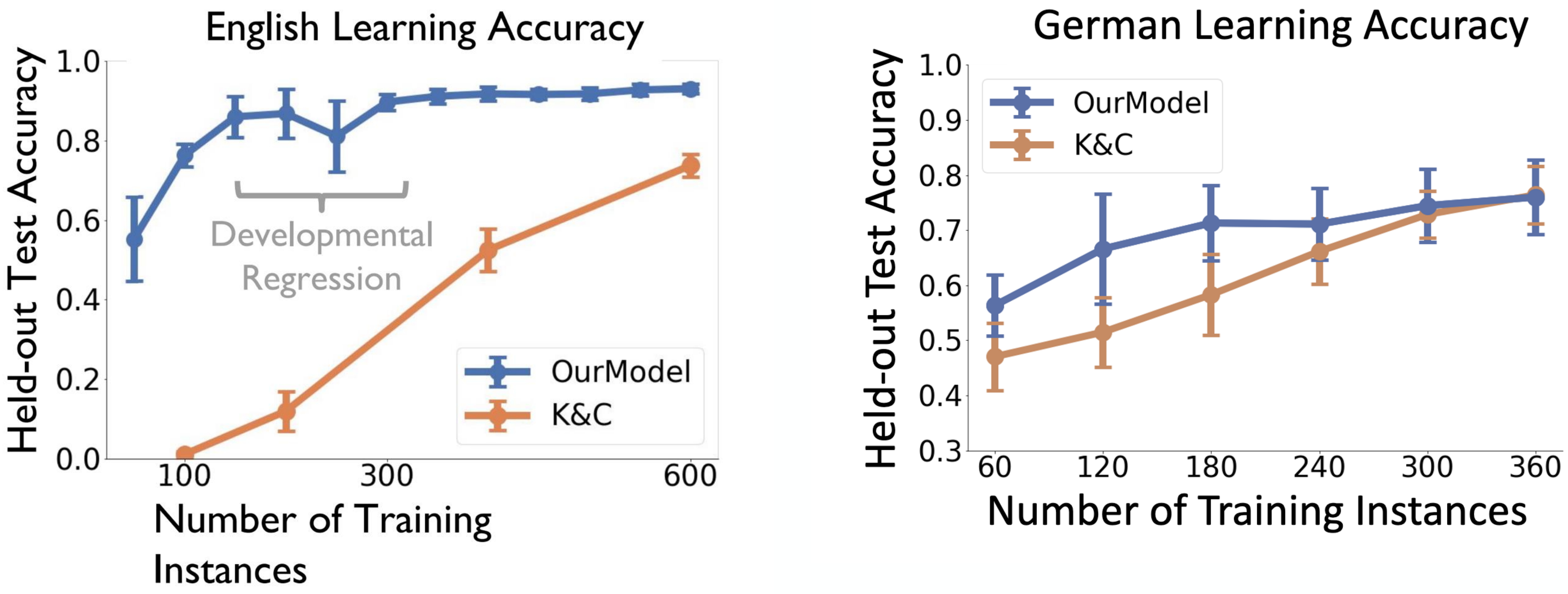
- Interpretable model of the grammar!
- But how are rules actually learned?

ATP: German Results

- Correlates better with nonce word predictions than K&C:

	Neuter			Unknown		
	%R	%NR	ρ	%R	%NR	ρ
-(e)n	0.17	0.04	-0.26	0.19	0.23	0.43
-e	0.27	0.35	-0.14	0.45	0.62	0.01
-∅	0.11	0.0	0.55	0.07	0.00	0.55
-er	0.44	0.17	0.53	0.29	0.0	0.46
-s	0.01	0.44	0.3	0.01	0.15	0.64
other	0.00	0.00		0.00	0.00	

ATP: Results


English Learning Accuracy
Held-out Test Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
100
300
600
Number of Training Instances
Developmental Regression
OurModel
K&C
German Learning Accuracy
Held-out Test Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
60
120
180
240
300
360
Number of Training Instances
OurModel
K&C