Root Infinitives and the Acquisition of Morphological Marking

Sarah Brogden Payne
sarah.payne@stonybrook.edu

PLC 47
March 19, 2023

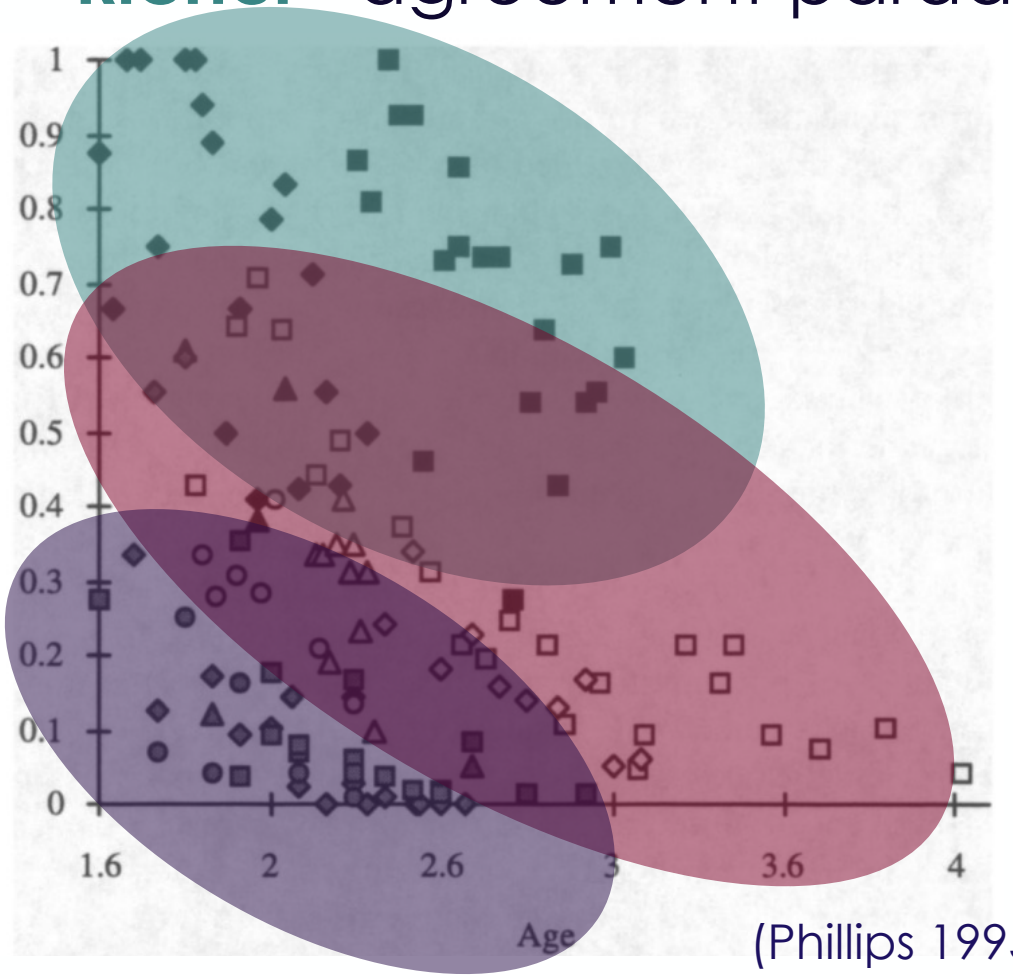
Background: Root Infinitives (RIs)

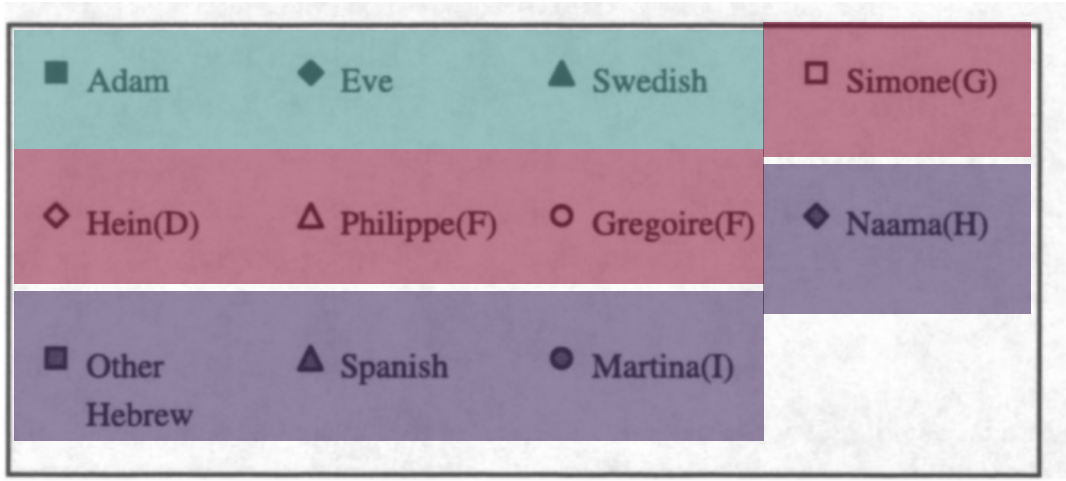
- Use of a non-finite verb in the matrix position:
  - **English:** Papa **have-INF** it
  - **Swahili:** mbaza …aza ku-ni-chund-a
    Mbaza **INF-OA$_{1.SG}$-pinch-IND**
  - **French:** Dormir petit bébé
    sleep-INF little baby
  - **German:** mein Kakao hinstelln
    my cocoa **put-INF**
  - **Hebrew:** Lashevet al ha-shulxan
    sit-INF on the-table

(examples from Legate & Yang 2007, Deen 2005)

Background: Cross-Linguistic Differences

- **"Richer"** agreement paradigms ⇒ **shorter & less frequent** RI





(Phillips 1995, Legate & Yang 2007)

Background: a Morphological Problem?

- Are RIs just the **failure to apply a morphological process** at PF?
  - Apply nothing in the absence of productive rule?
- No: **form-position correlations!**

Finite Clauses	Non-Finite Clauses

Background: a Morphological Problem?

- Are RIs just the **failure to apply a morphological process** at PF?
  - Apply nothing in the absence of productive rule?
- No: **form-position correlations!**

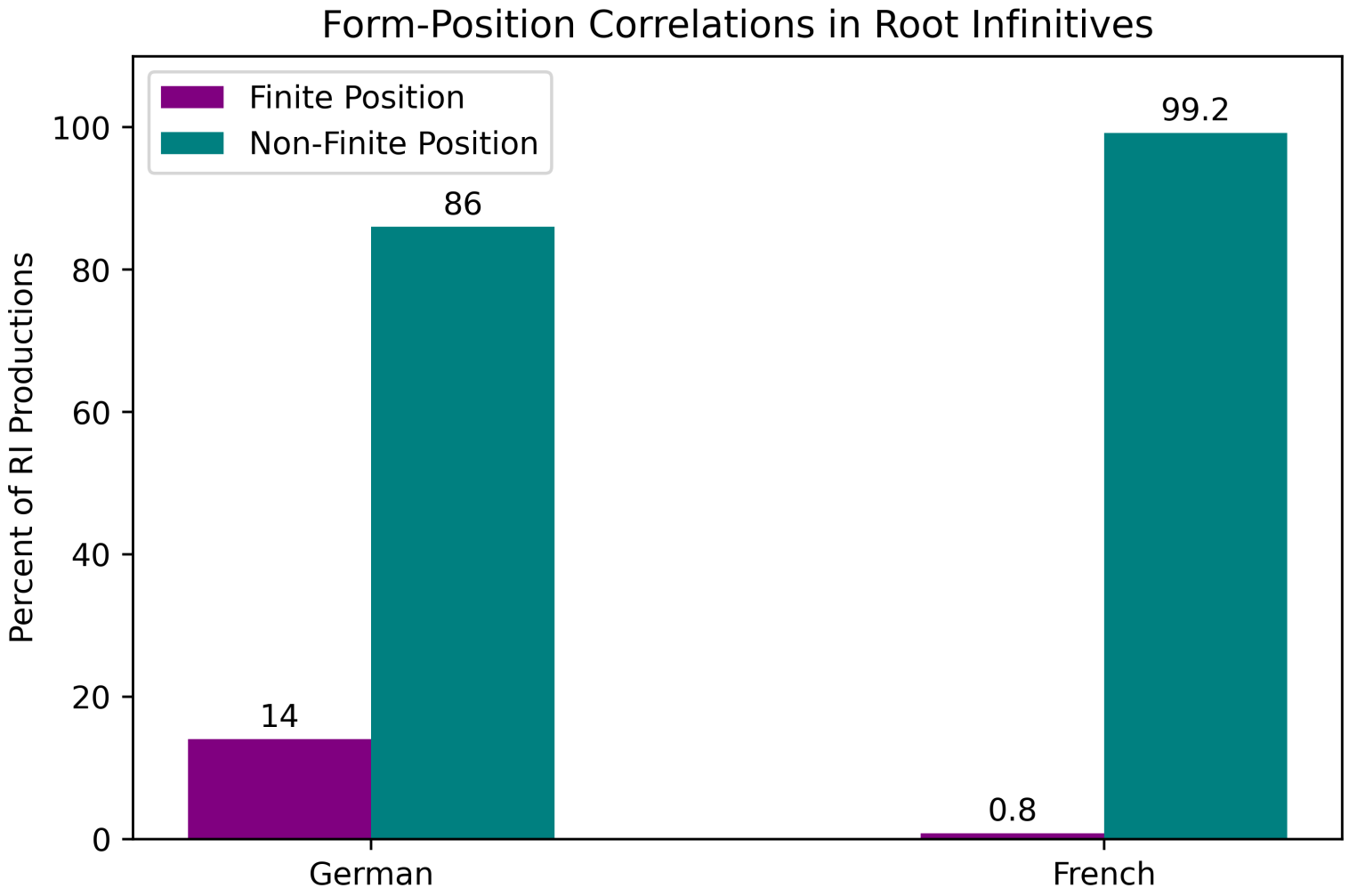
	Finite Clauses	Non-Finite Clauses
German:	V → T → C: **verb high**	V → T: **verb-final**

Background: a Morphological Problem?

- Are RIs just the **failure to apply a morphological process** at PF?
  - Apply nothing in the absence of productive rule?
- No: **form-position correlations!**

	Finite Clauses	Non-Finite Clauses
German:	V → T → C: **verb high**	V → T: **verb-final**
French:	V → T: **before negation**	V-in-situ: **after negation**

Background: a Morphological Problem?



(Poeppel & Wexler 1993, Pierce 1989 & 1992, Phillips 1995)

Proposal

- RIs = **byproduct of the acquisition of inflectional categories**
  - Child must learn ***which inflectional categories are marked***
    - English contrasts **±PAST** but Mandarin doesn't
    - Spanish contrasts **±1** but English doesn't
  - RIs emerge before the child learns that ***their language marks tense***
- More evidence for tense marking in high position ⇒ tense acquired earlier (Legate & Yang 2007)
- Focus: modeling **crosslinguistic differences** in overall **length & frequency** of RI

Proposal

- Model of the **acquisition of inflectional categories**
  - Matches ***developmental findings***
    - **Order of acquisition**
    - **Vocabulary size**
  - Correctly predicts ***cross-linguistic differences*** in RI stage

Preliminaries: Data

- Children learn **frequent forms earlier**
  - Use ***most frequent forms from CHILDES***
- Children use of **distributional cues** to learn meaning
  - Intersect CHILDES with UniMorph ***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}
French	/ʁə.gaʁ.de/	/ʁə.gaʁd/	{V, IMP, PRES, 2, SG}

(Goodman et al 2008, MacWhinney 2000, Kirov et al 2018)

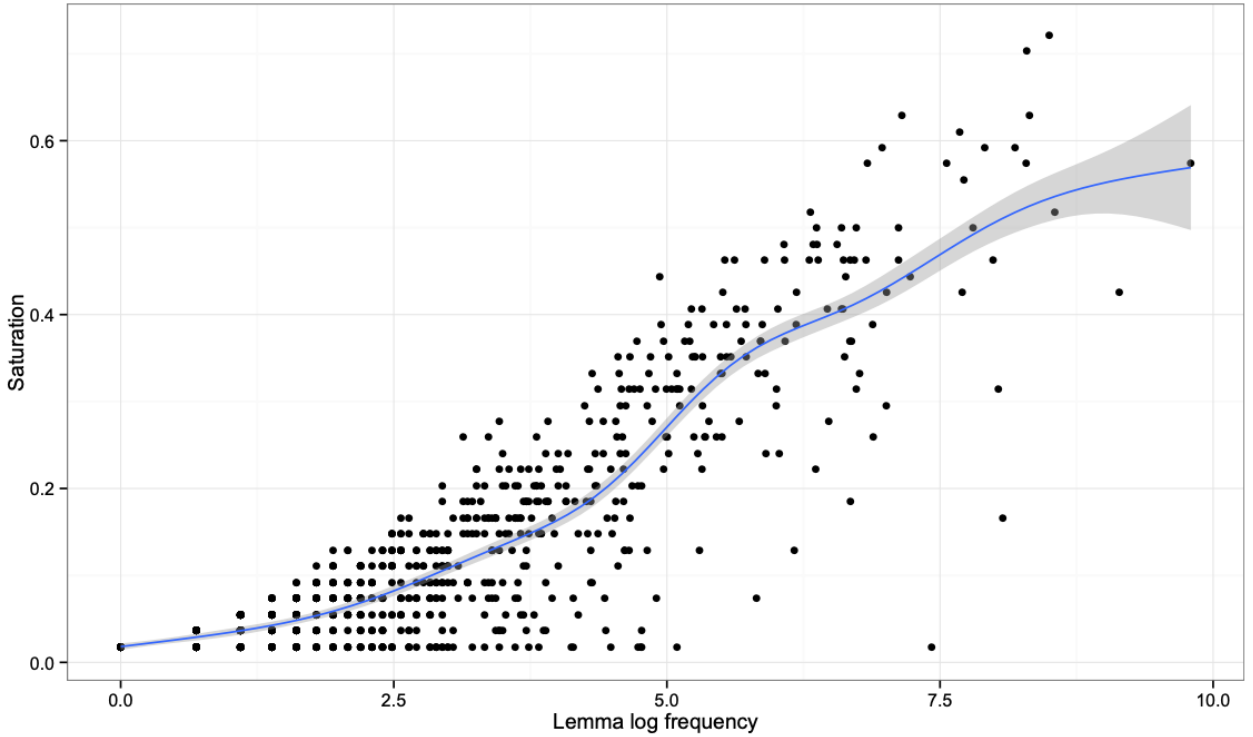
Model: Sufficient Contrast Learner

- **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*** ⇒ **±PAST** is marked
- **Infants sensitive to collisions:** can relate nonce words to their stems as early as 0;6

(Clark & MacWhinney 1987, Kim & Sundara 2021)

Model: The TSP

- Is a **single collision** enough to learn marking?
  - ***I am ~ you are*** ⇒ English marks 1 vs. 2 person?
- Should we require **all lemmas** to have collisions?
  - **Sparsity of the input:** morphological paradigm saturation
  - **Syncretisms:** e.g. ***put-put***



(Chan 2008, Lignos & Yang 2016)

Model: The Tolerance Principle

- When are there "**enough" collisions** to learn that an inflectional category is marked?
  - **Tolerance-Sufficiency Principle:** threshold for generalization based on ***computational efficiency***
  - Given ***N*** items, ***M*** of which we've seen doing ***X***, all do ***X*** iff:

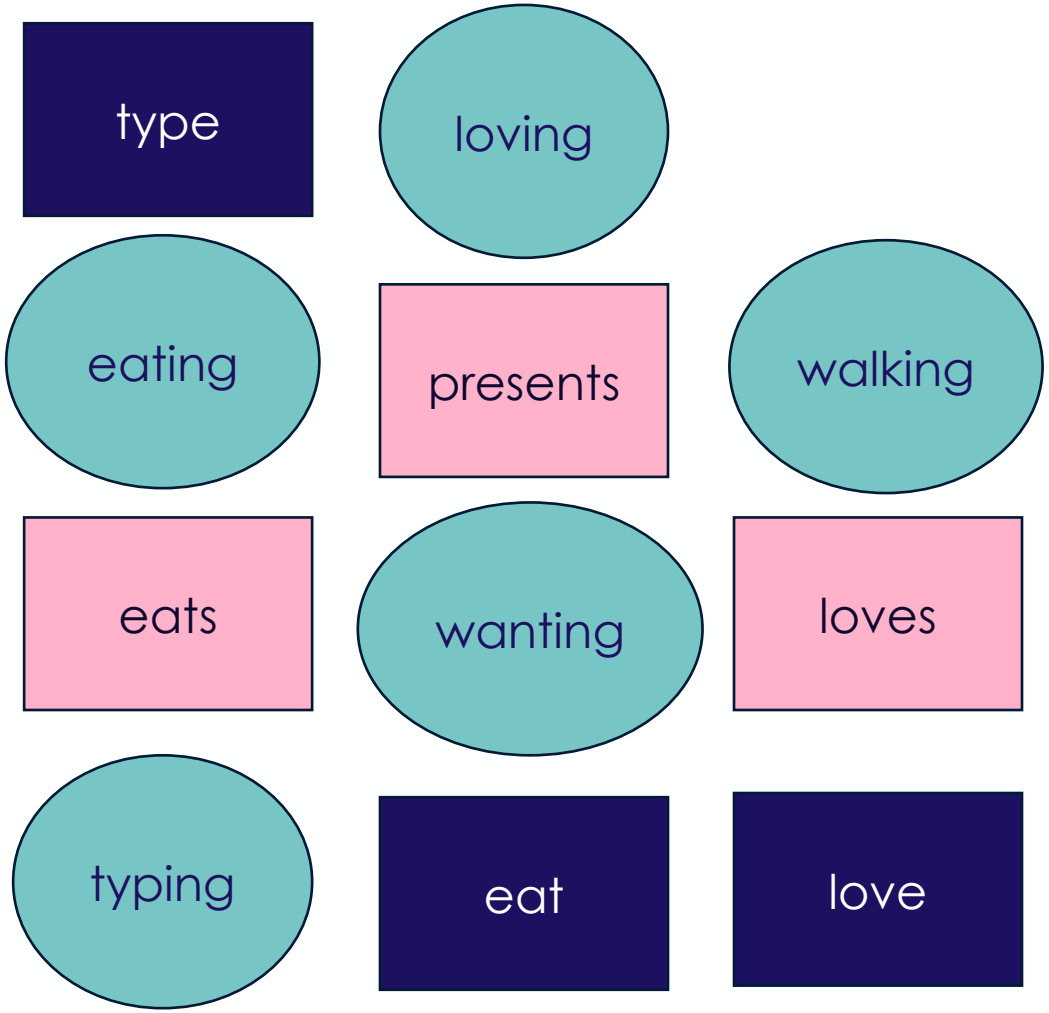
$$N - M \leq \theta_N = \frac{N}{\ln N}$$

(Yang, 2016)

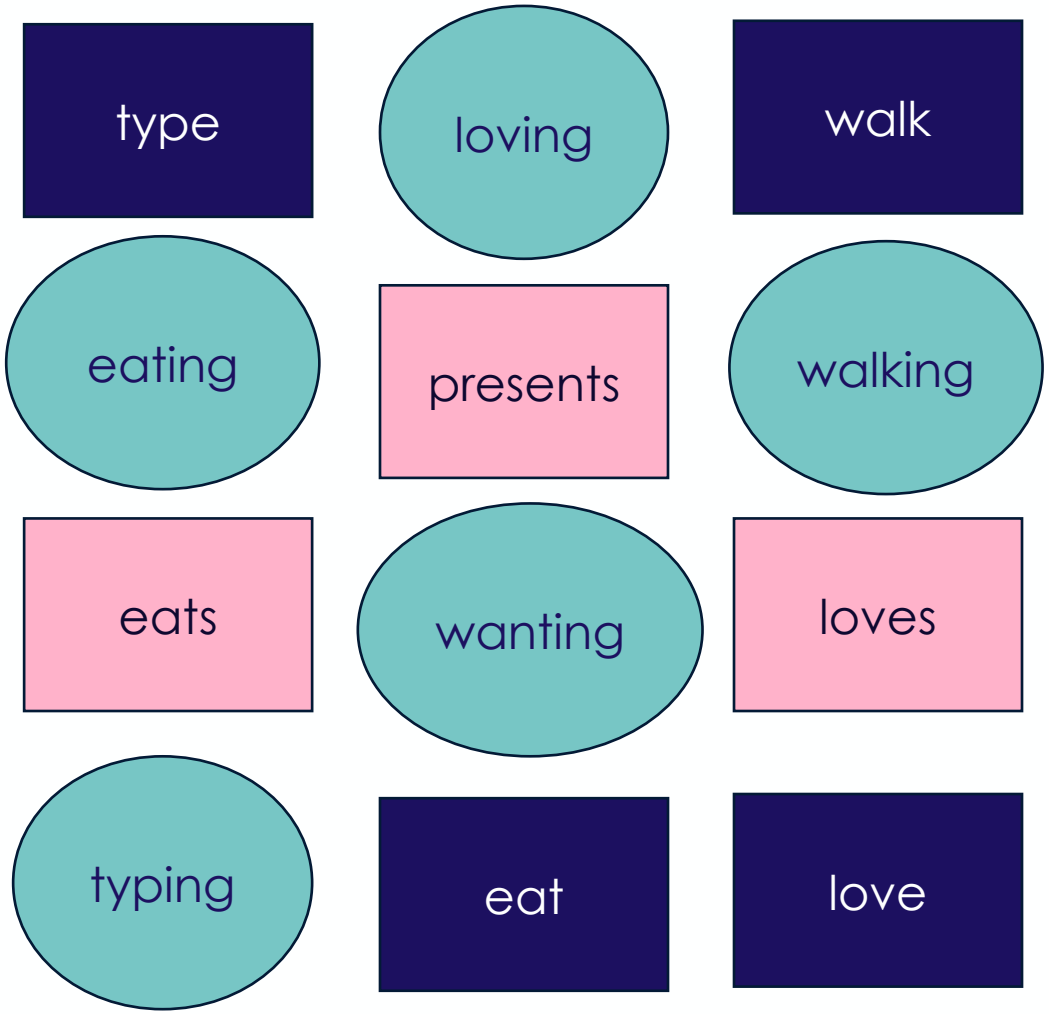
Model: Recursive Subdivision

- Take in input **incrementally**
- If inflection ***A*** (less frequent) has a collision with inflection ***B*** (more frequent):
  - **Do enough $(A - \theta_A)$ verbs that appear in *A* appear in *B* in a different form than *A*?**
- If **enough** words have a collision (by TSP):
  - Subdivide the input based on the **feature difference between *A* and *B***
  - **Recurse** on each resulting set

Model: Toy Example

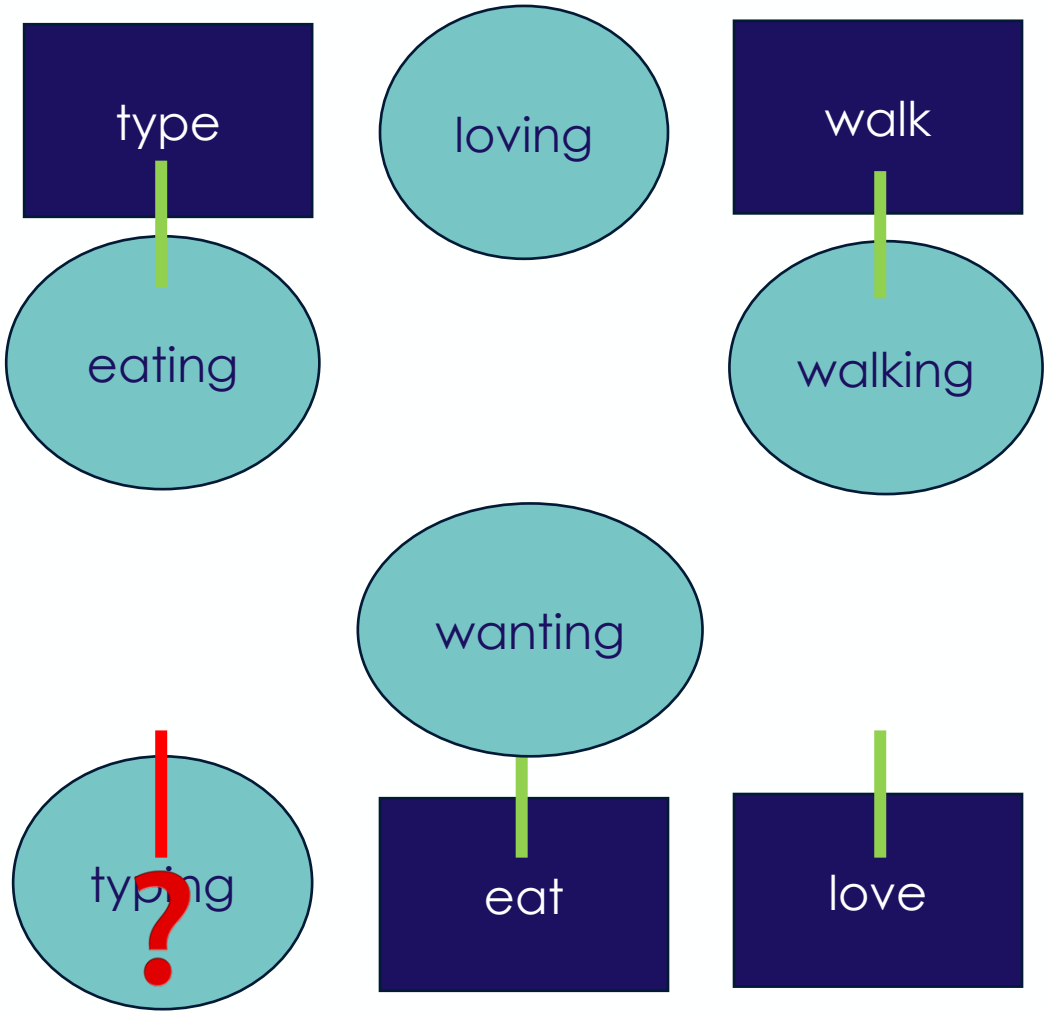


Model: Toy Example



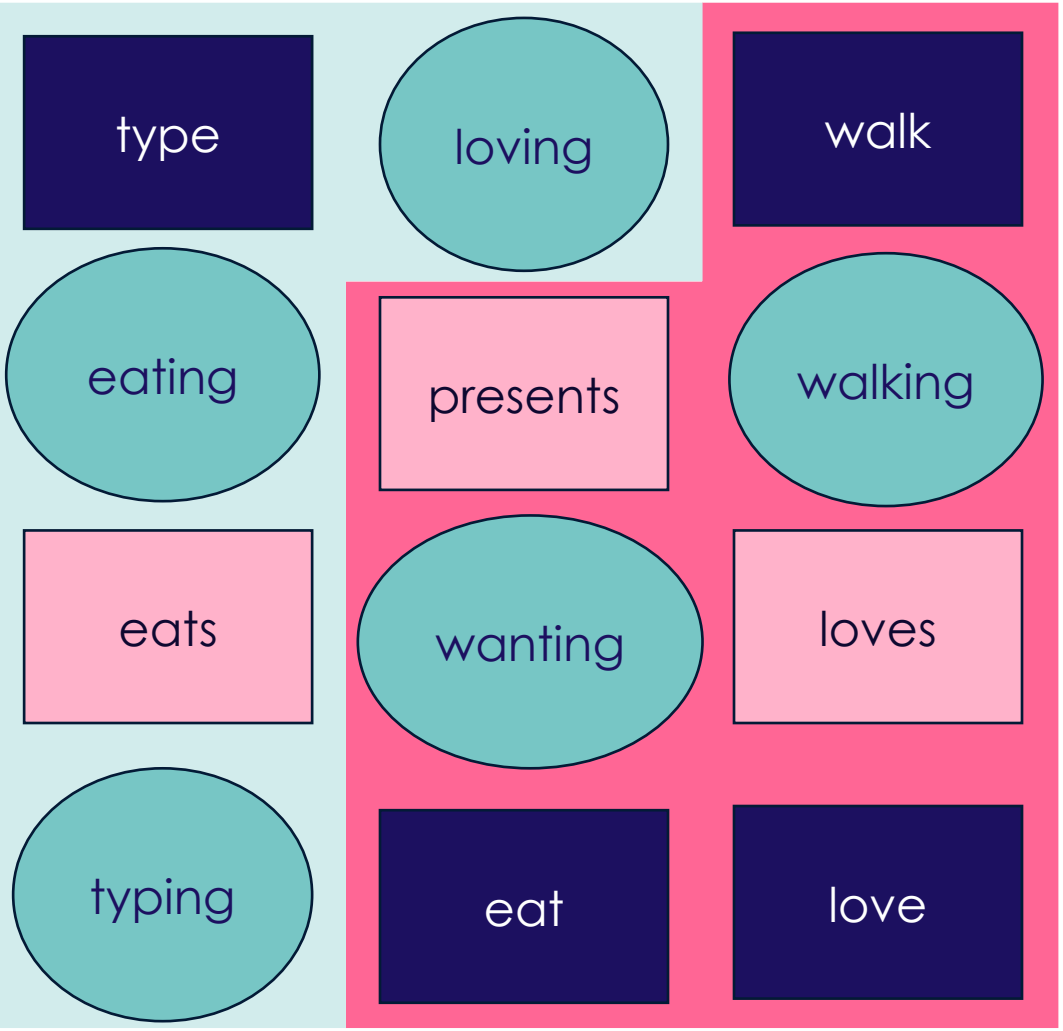
- Collision: ***walk*~*walking***
- **±PARTICIPLE** marked?

Model: Toy Example



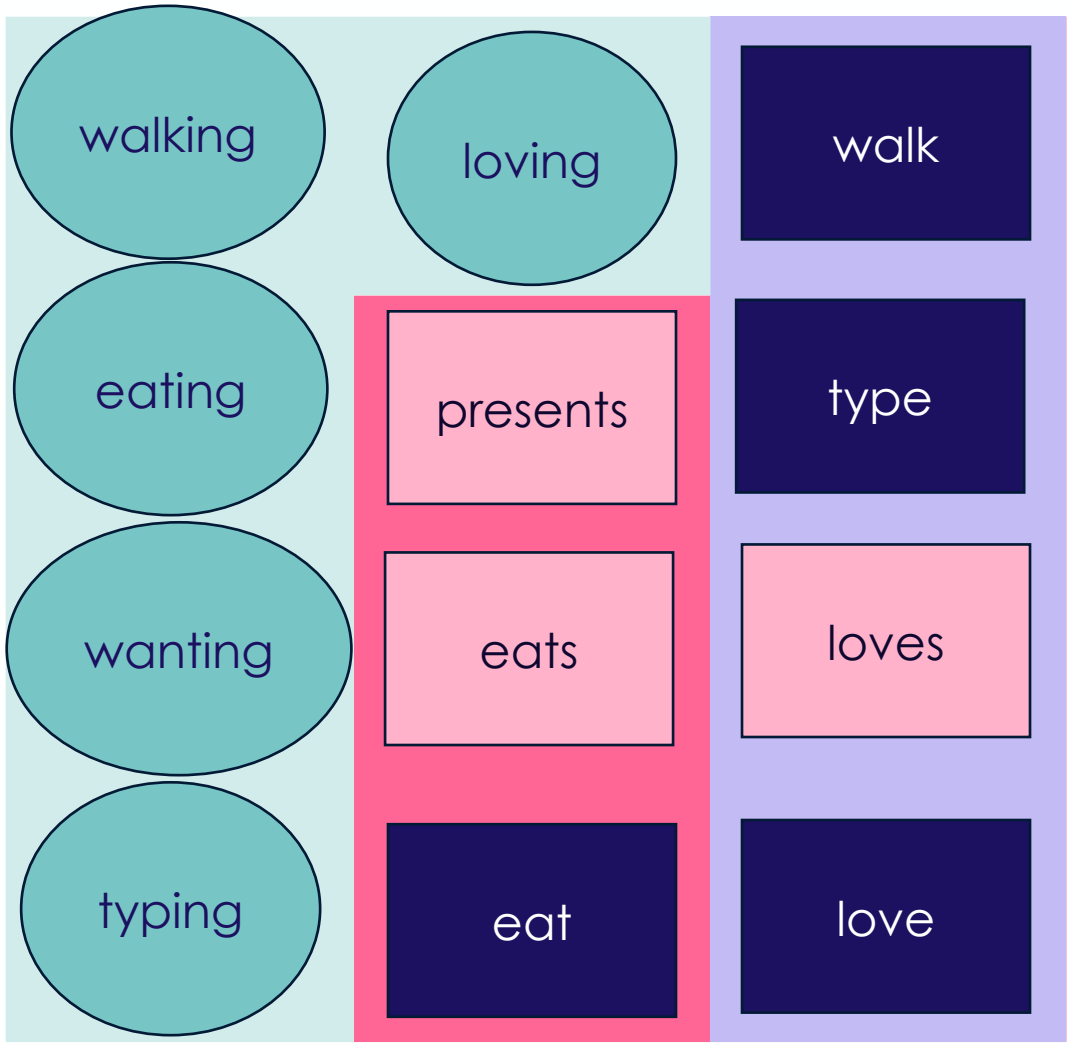
- Collision: ***walk*~*walking***
- **±PARTICIPLE** marked?
  - 5 participles, 4 collisions (not ***wanting***)
  - $N - M = 1 < \theta_5 = 3$ ✅
- Contrast 1 productive! **±PARTICIPLE marked**

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

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 **±3.SG** marked

Experiments

- **English vs. French vs. Spanish** verbs (following Legate & Yang 2007)
  - **English:** ***longest & most frequent*** RI
  - **French:** in the middle
  - **Spanish:** ***shortest & least frequent*** RI
- Does our model **match developmental findings**?
  - Order of acquisition
  - Vocabulary size
- Can it account for **cross-linguistic differences** in RI?

Predictions

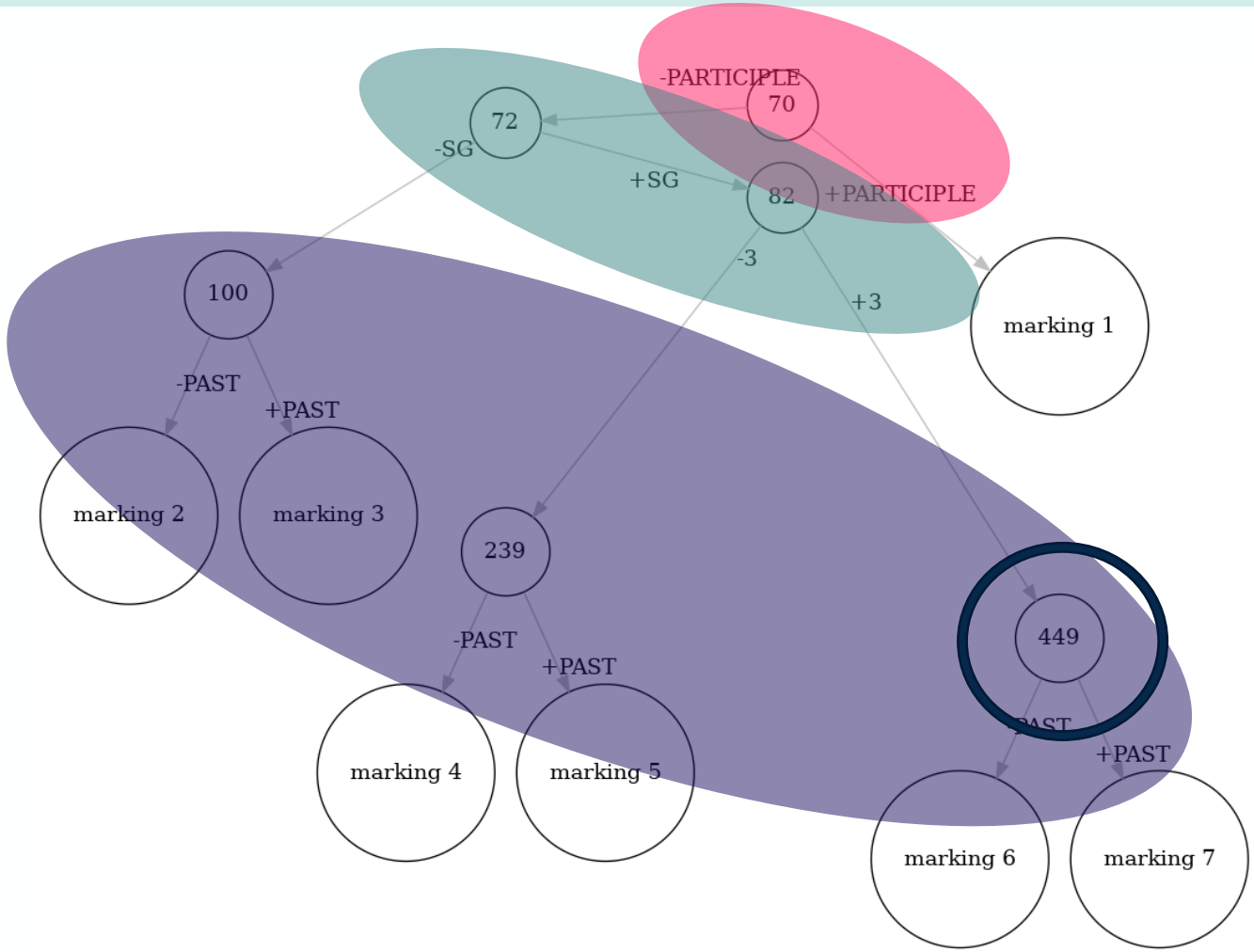
- All 3 languages: **subject agreement before tense**
- Richer agreement paradigm ⇒ **more subdivision**
- More subdivision ⇒ **smaller *Ns***
- Smaller *Ns* ⇒ **learn tense more quickly**
  - **TSP tolerates relatively more exceptions for smaller *N***
  - $\theta_{10} \approx 4 = \mathbf{40\%}$ but $\theta_{100} \approx 21 = \mathbf{21\%}$
- Learn tense more quickly ⇒ **shorter RI**

∴ **Richer agreement paradigm ⇒ shorter RI**

Results: English

- Order of acquisition:
  - **PARTICIPLE**
  - **3.SG**
  - **PAST**

  } Matches with developmental findings
- Vocabulary size:
  - At 3;0 know ≤ **250 verb stems**
  - Done learning at **188 stems**
- Tense emerges:
  - By **449 inflected** forms (**188 stems**)



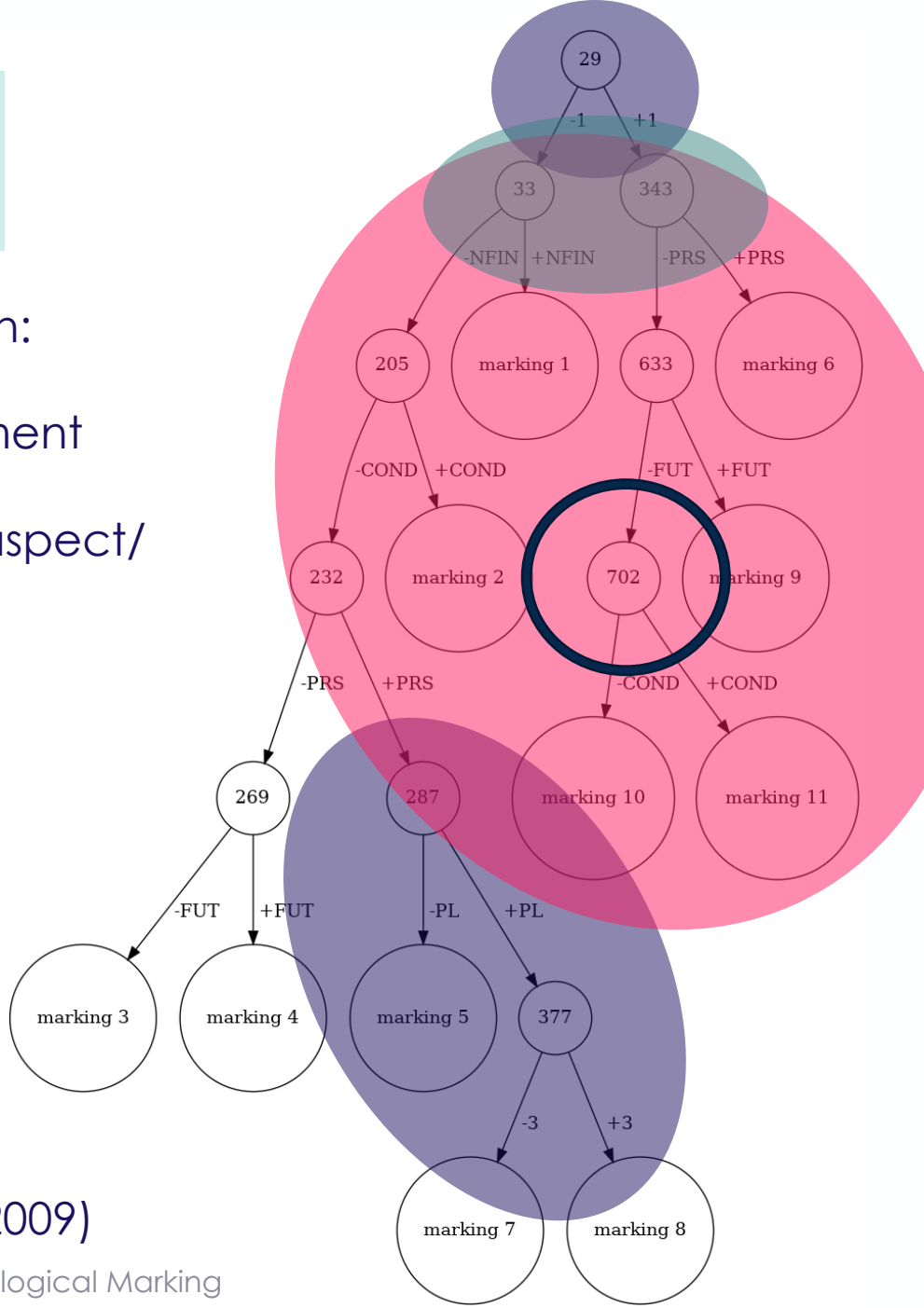
(Fenson et al. 1994, Bornstein et al. 2004, Berko 1958, Brown 1973)

Results: French

- Order of acquisition:
  - Subject agreement **early & late**
  - **Tense/aspect/mood** after ±1

Children: subject agreement before tense/aspect/ mood

- Vocabulary size:
  - At 1;8, children know **≤ 400 words**
  - Done learning at **232 stems**
- Tense emerges:
  - By **343 inflected** forms **(124 stems)**

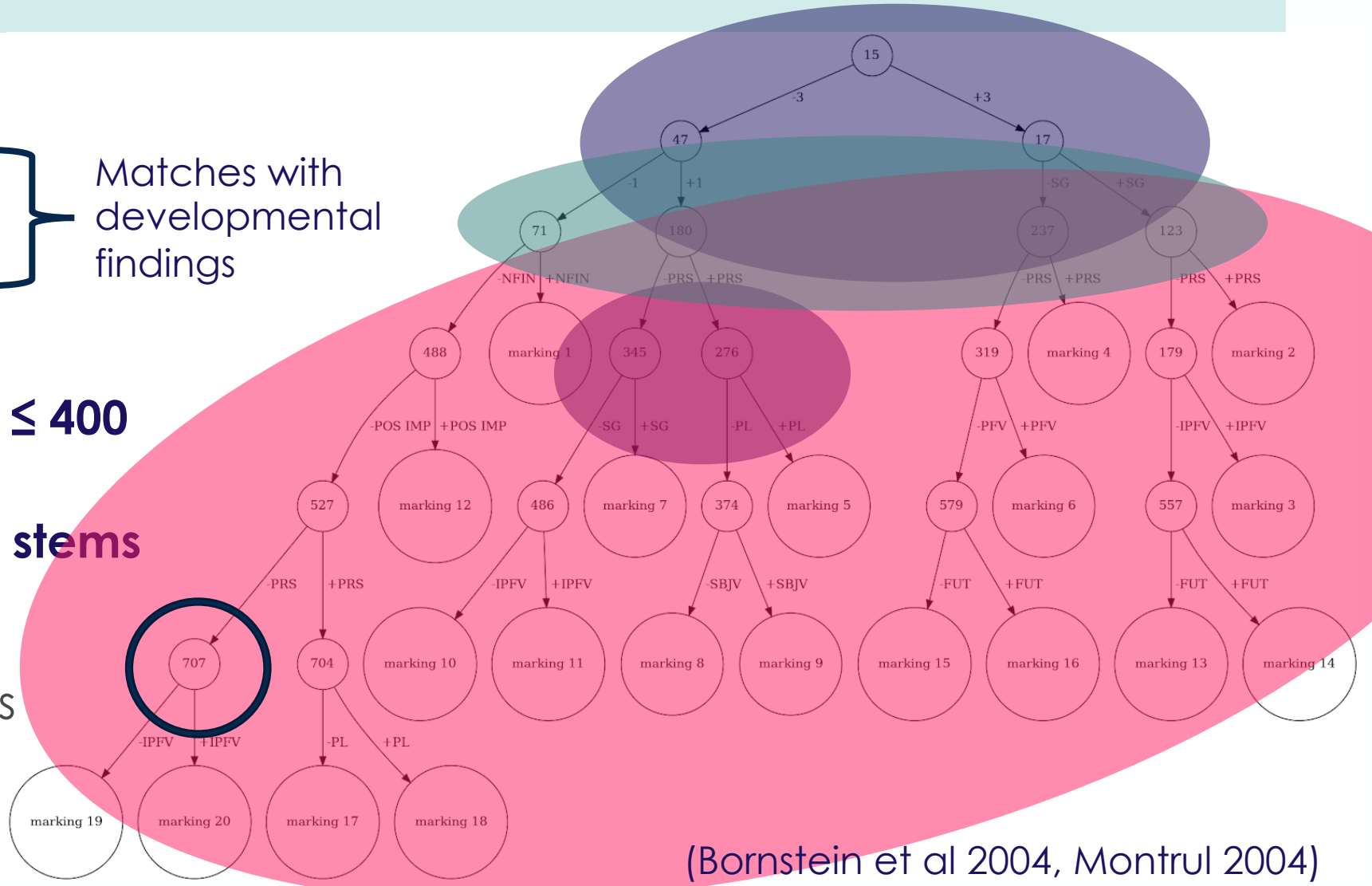


(Bornstein et al 2004, Prevost 2009)

Results: Spanish

- Order of acquisition:
  - **Subject agreement**
  - **Tense/aspect mood**

  } Matches with developmental findings
- Vocabulary size:
  - At 1;8, children know **≤ 400 words**
  - Done learning at **230 stems**
- Tense emerges:
  - By **237 inflected** forms **(103 stems)**



(Bornstein et al 2004, Montrul 2004)

Results: Cross-linguistic Differences

- Length of RI Stage in children:

Spanish < French < English

- Number of stems on which our model learns tense marking:

Spanish (103) < French (124) < English (188)

- Number of inflected forms on which our model learns tense marking:

Spanish (237) < French (343) < English (449)

Discussion

- Our model: **mechanistic account** of RI stage as a byproduct of the **acquisition of inflectional categories**
  - Relies only on **inequality between inflected forms**
- Future work:
  - Apply to **more languages**
  - Combine with **grounded/distributional models** to learn features
  - Investigate high vs. non-high in **French and German**

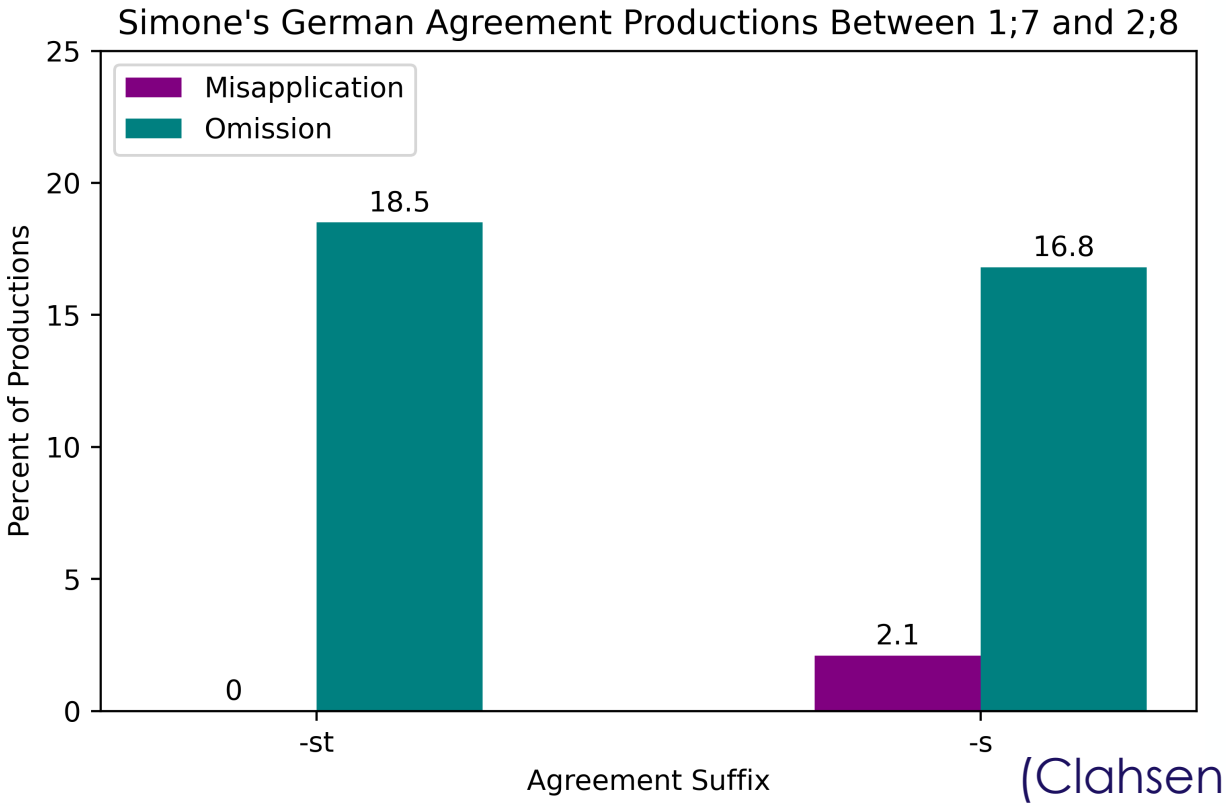
Thank you!!

I am grateful to Charles Yang, Jordan Kodner, Jeff Heinz, Julie Anne Legate, Mark Aronoff, Bob Berwick and his lab, Kyle Gorman and the CUNY Computational Linguistics group, Salam Khalifa, and attendees of the Stony Brook University Brown Bag for comments and feedback.

I am grateful for funding by the Instituted for Advanced Computational Science Graduate Research Fellowship and the National Science Foundation Graduate Research Fellowship.

Background: a Syntactic Problem?

- Are RIs just **a failure of AGREE?**
  - Failure of φ-agreement ⇒ ***substitution errors*** (e.g. I has it)



(Clahsen & Penke 1992, Philips 1995)