

ON LEARNING UNDERLYING FORMS

Adam Jardine



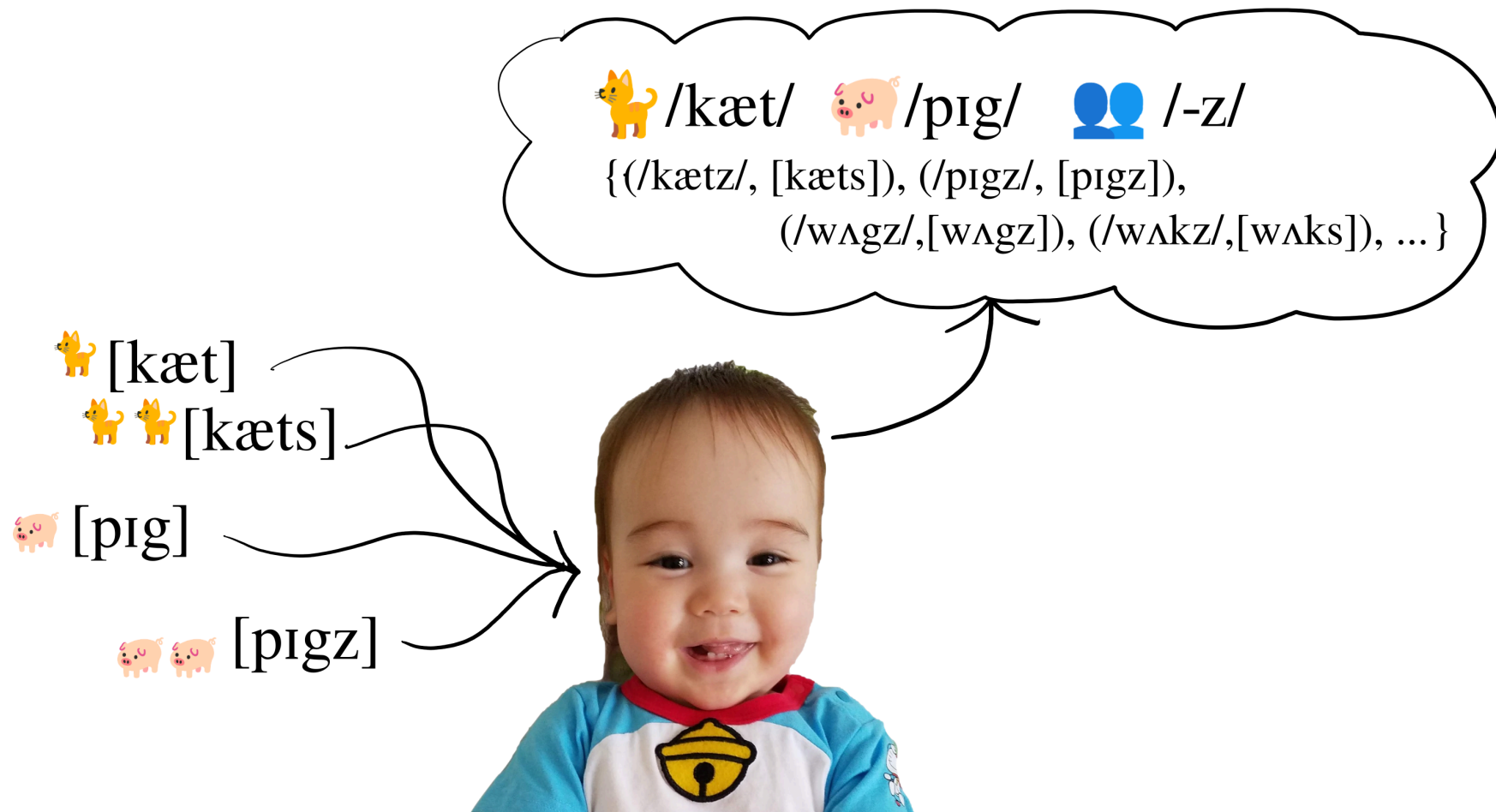
CUNY Colloquium
October 22, 2026

Johor Malay (Onn, 1976)

SR	gloss
[pəŋarah]	NOM-‘direct’
[pəmbawə]	NOM-‘carry’
[pəndaki]	NOM-‘climb’
[pəŋgali]	NOM-‘dig’

	λ	
NOM	$\mapsto$	/pəŋ/
CARRY	$\mapsto$	/bawa/
CLIMB	$\mapsto$	/daki/
$\vdots$	$\vdots$	$\vdots$
NOMDIRECT	$\mapsto$	/pəŋarah/
NOMCARRY	$\mapsto$	/pəŋbawa/
NOMCLIMB	$\mapsto$	/pəŋdaki/
$\vdots$	$\vdots$	$\vdots$

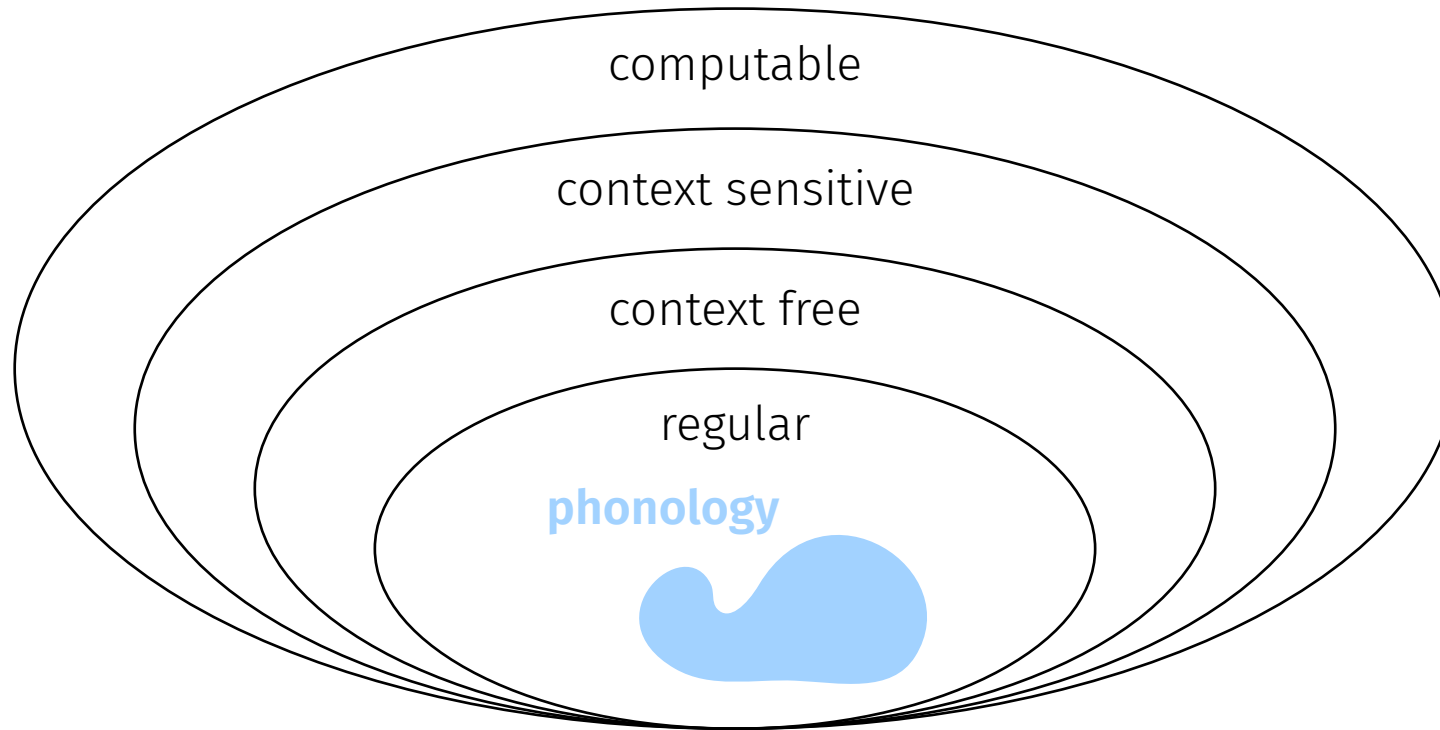
λ			φ		
NOM	$\mapsto$	/pəŋ/	/pəŋ/	$\mapsto$	[pəŋ]
CARRY	$\mapsto$	/bawa/	/bawa/	$\mapsto$	[bawə]
CLIMB	$\mapsto$	/daki/	/daki/	$\mapsto$	[daki]
:	:	:	:	:	:
NOMDIRECT	$\mapsto$	/pəŋarah/	/pəŋarah/	$\mapsto$	[pəŋarah]
NOMCARRY	$\mapsto$	/pəŋbawa/	/pəŋbawa/	$\mapsto$	[pəŋbawə]
NOMCLIMB	$\mapsto$	/pəŋdaki/	/pəŋdaki/	$\mapsto$	[pəŋdaki]
:	:	:	/əŋbə/	$\mapsto$	[əmbə]
			:	:	:



Target			Input	
λ			Finite sample of $\varphi \circ \lambda$	
		φ		
NOM	$\mapsto$ /pəŋ/	/pəŋ/	$\mapsto$ [pəŋ]	(DIRECT, [pəŋarah])
CARRY	$\mapsto$ /bawa/	/bawa/	$\mapsto$ [bawə]	(CARRY, [pəmbawə])
CLIMB	$\mapsto$ /daki/	/daki/	$\mapsto$ [daki]	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
NOMDIRECT	$\mapsto$ /pəŋarah/	/pəŋarah/	$\mapsto$ [pəŋarah]	(NOMDIRECT, [pəŋarah])
NOMCARRY	$\mapsto$ /pəŋbawa/	/pəŋbawa/	$\mapsto$ [pəŋbawə]	(NOMCARRY, [pəmbawə])
NOMCLIMB	$\mapsto$ /pəŋdaki/	/pəŋdaki/	$\mapsto$ [pəŋdaki]	(NOMCLIMB, [pəndaki])
$\vdots$	$\vdots$	/əŋbə/	$\mapsto$ [əmbə]	(NOMDIG, [pəŋgali])
		$\vdots$	$\vdots$	

The learning problem: From a set of meanings and surface forms exhibiting morphophonological alternations, return a set of underlying representations (URs) and a phonological grammar.

The path forward: What kind of *function* is the phonology?



Chandlee & Jardine (2025): Dataset from Johore Malay (Onn, 1976)

- **Input:** 14 stems × 5 affixes
- **Output:** Correct lexicon and grammar representing five processes
 - Including counter-feeding
- Generalizes theoretical results of Hua et al. (2021), Hua & Jardine (2021)
- **Inductive principles:**
 - Phonological functions are **input-local**
 - Phonological functions refer to **natural classes** (in the environment)

- This talk based on Chandlee & Jardine (2025)
- NSF Grant #2416184 with

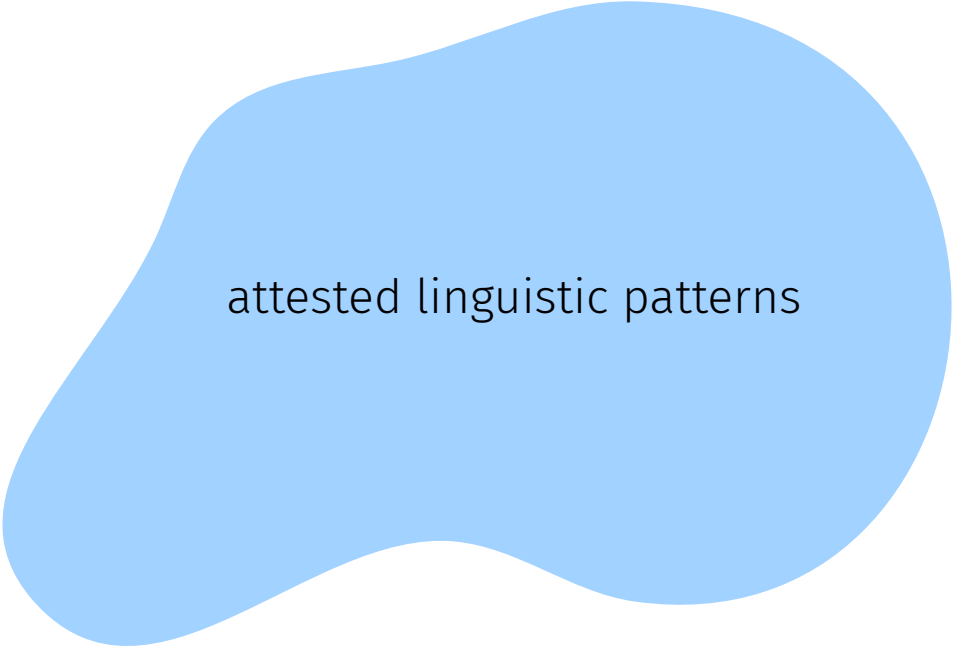


Jane Chandlee (Haverford) **Jeffrey Heinz** (Stony Brook)

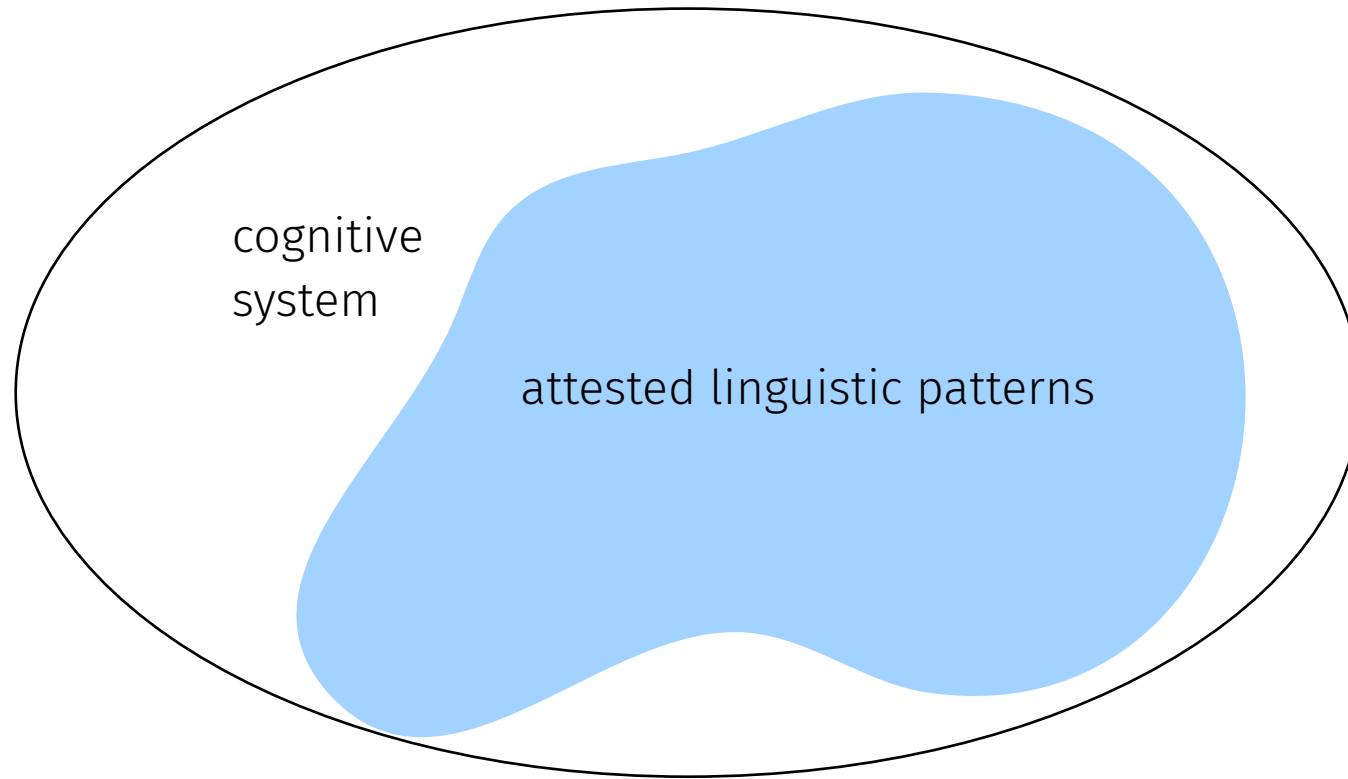


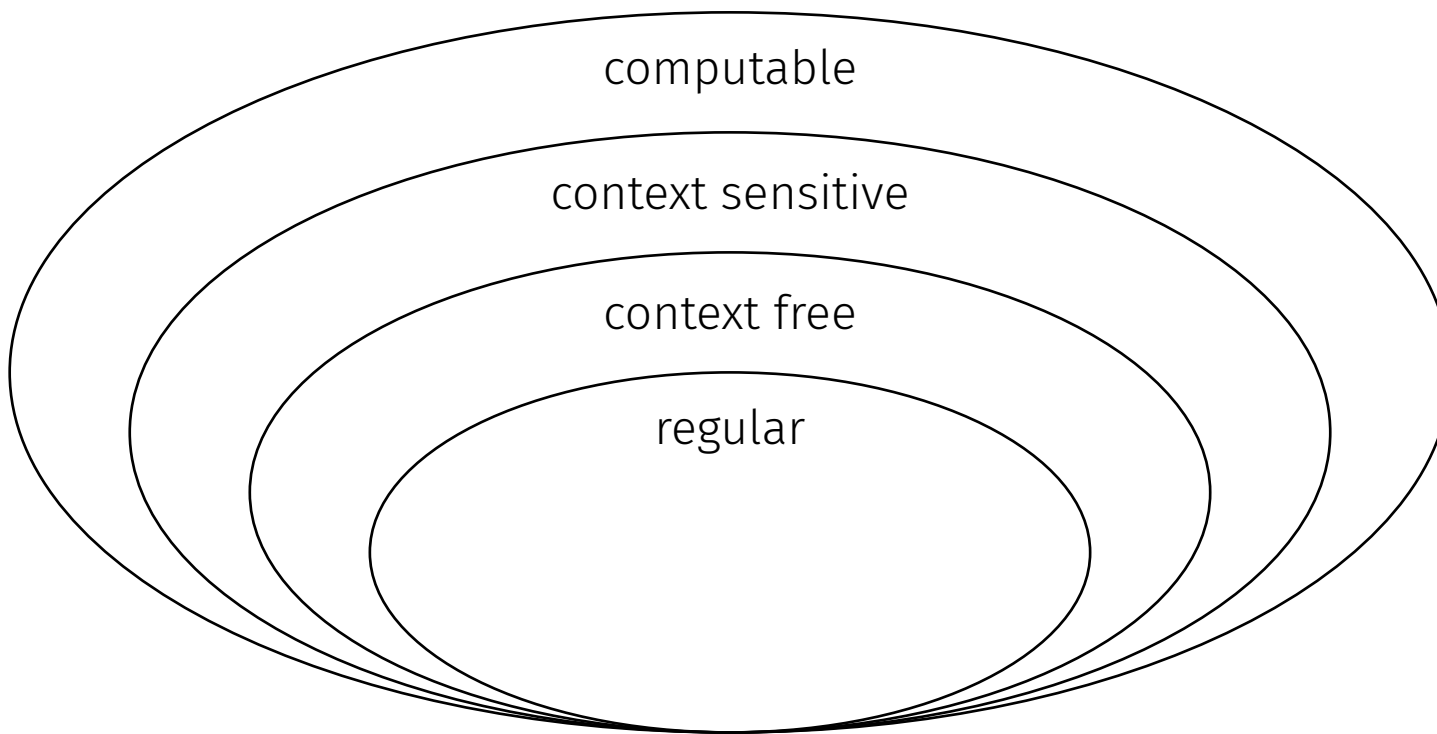
- **Grads:** Vincent Czarnecki Chenli Wang Eli Franz Zachary Helfand
Hyunjung Joo (now at Hanyang Univ.)
Huteng Dai (now at Univ. of Michigan)
Wenyue Hua (Ling MA, CS PhD, now at Microsoft Research)
- **UGs:** Ezra Anthony Neha Jeyaram Manish Kandepalli
Ben Evans (now at Stony Brook)

Phonology, linguistic universals, and computation

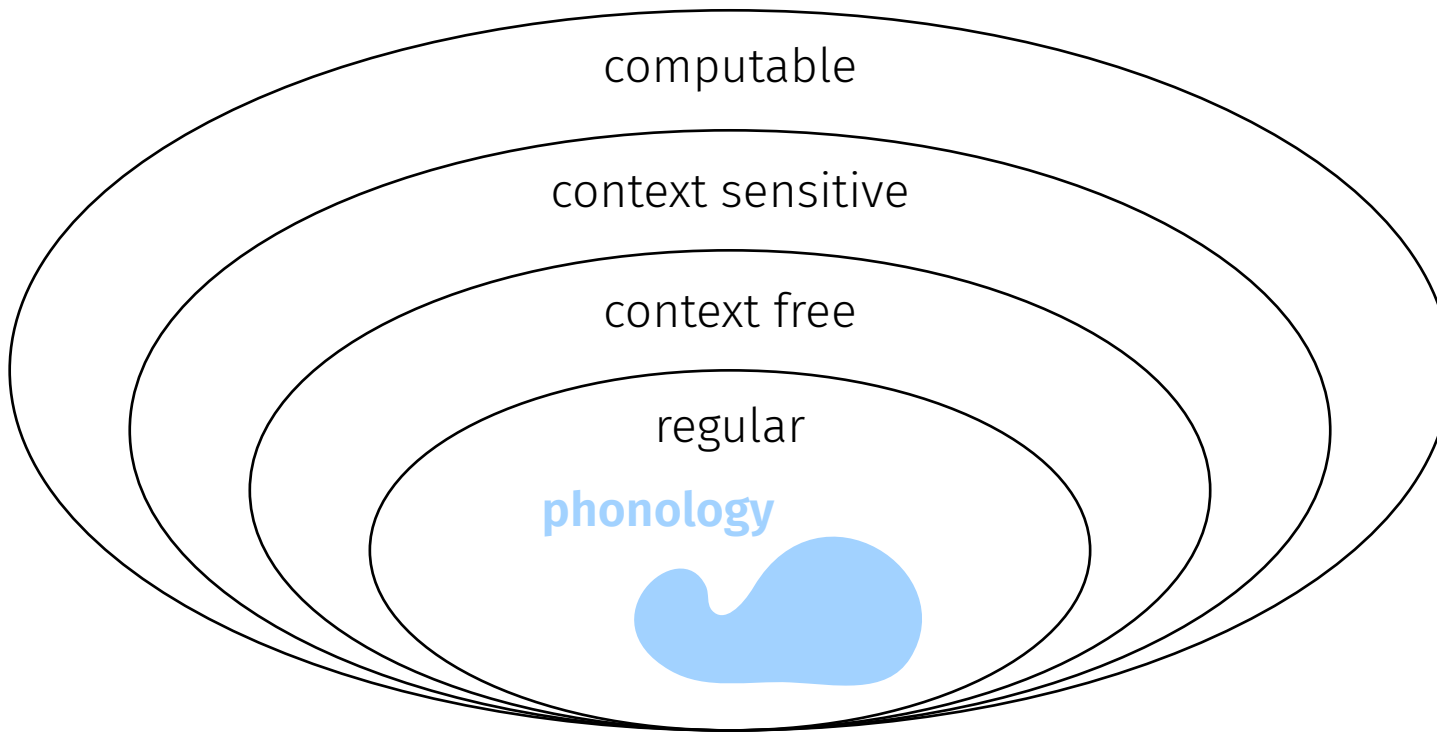
A light blue, irregular blob shape, resembling a cloud or a splash, is centered on the slide. It has a soft, organic outline with several rounded protrusions and indentations.

attested linguistic patterns

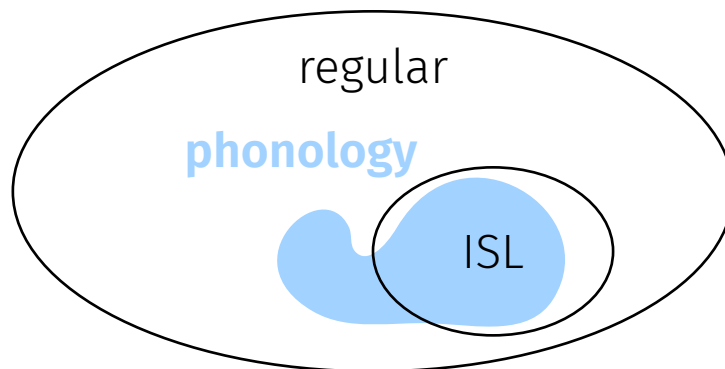




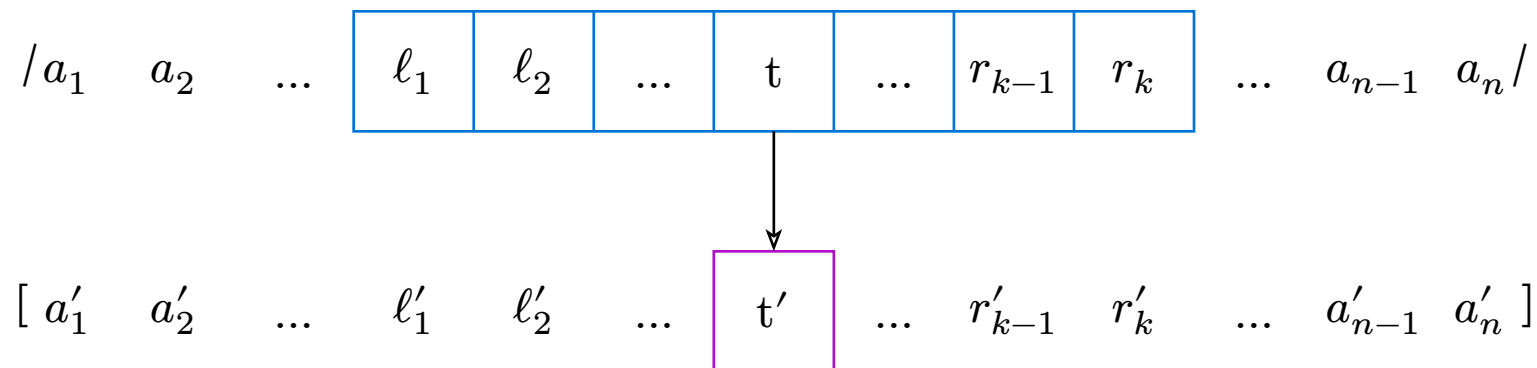
The Chomsky-Schützenberger Hierarchy (Chomsky 1957, Chomsky and Schützenberger 1963)



The Chomsky-Schützenberger Hierarchy (Chomsky 1957, Chomsky and Schützenberger 1963)
(Johnson 1972, Kaplan and Kay 1994, Heinz 2007, et seq.)



Input Strictly Local (ISL) functions
(Chandlee 2014, Chandlee and Heinz 2018)



Computation of an ISL_k function

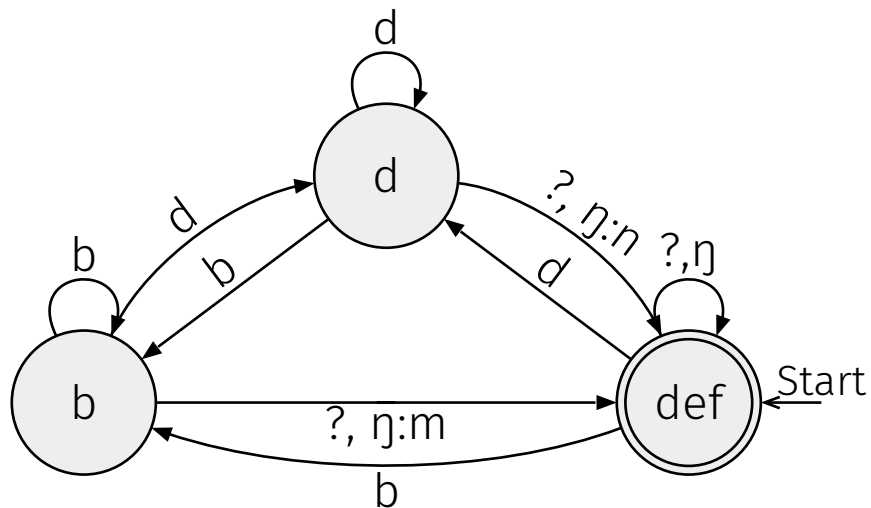
Nasal place assimilation in Johor Malay (Onn, 1976)

UR	SR	gloss
/pəŋarah/	[pəŋarah]	NOM-‘direct’
/pəŋbawə/	[pəmbawə]	NOM-‘carry’
/pəŋdaki/	[pəndaki]	NOM-‘climb’
/pəŋgali/	[pəŋgali]	NOM-‘dig’

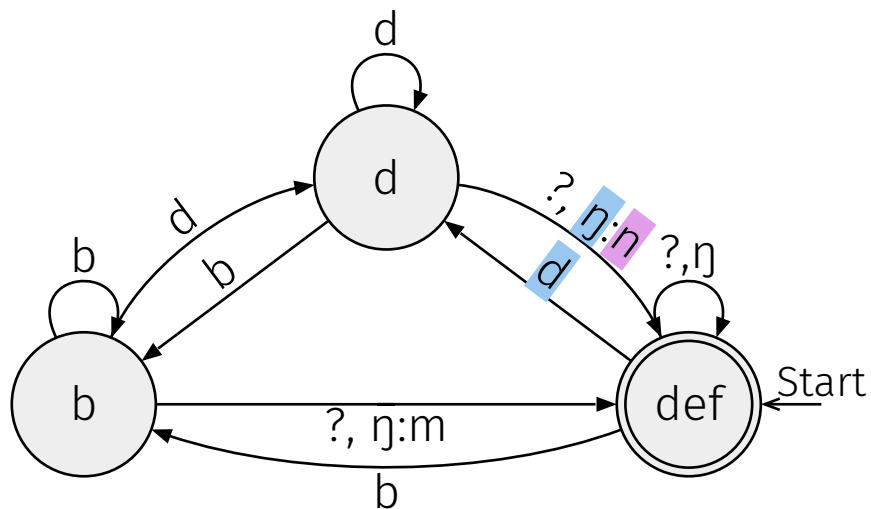
Nasal place assimilation in Johor Malay (Onn, 1976)

UR	SR	gloss
/pəŋarah/	[pəŋarah]	NOM-‘direct’
/pəŋbawə/	[pəmbawə]	NOM-‘carry’
/pəŋdaki/	[pəndaki]	NOM-‘climb’
/pəŋgali/	[pəŋgali]	NOM-‘dig’

(partial) **finite-state transducer (FST)** for nasal place assimilation

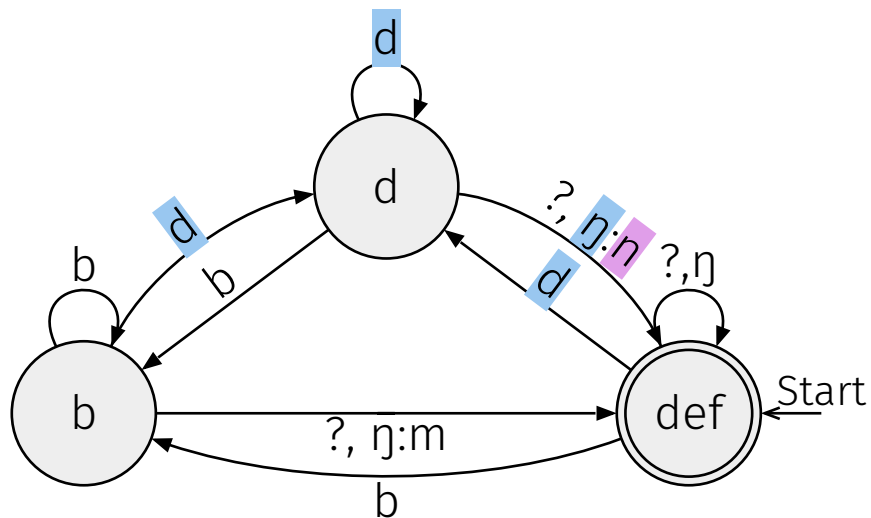


(partial) **finite-state transducer (FST)** for nasal place assimilation

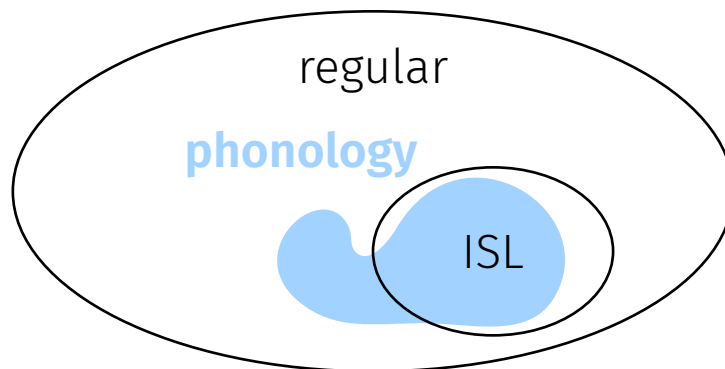


/pəɲdaki/ $\mapsto$ [pəndaki]

(partial) **finite-state transducer (FST)** for nasal place assimilation



/pəŋdaki/ $\mapsto$ [pəndaki]



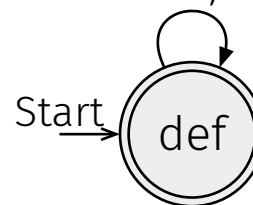
95% of P-Base (Mielke 2008) is ISL
(Chandlee 2014, Chandlee and Heinz 2018)

A proposal

Target			Input	
λ			Finite sample of $\varphi \circ \lambda$	
		φ		
NOM	$\mapsto$ /pəŋ/	/pəŋ/	$\mapsto$ [pəŋ]	(DIRECT, [pəŋarah])
CARRY	$\mapsto$ /bawa/	/bawa/	$\mapsto$ [bawə]	(CARRY, [pəmbawə])
CLIMB	$\mapsto$ /daki/	/daki/	$\mapsto$ [daki]	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
NOMDIRECT	$\mapsto$ /pəŋarah/	/pəŋarah/	$\mapsto$ [pəŋarah]	(NOMDIRECT, [pəŋarah])
NOMCARRY	$\mapsto$ /pəŋbawa/	/pəŋbawa/	$\mapsto$ [pəŋbawə]	(NOMCARRY, [pəmbawə])
NOMCLIMB	$\mapsto$ /pəŋdaki/	/pəŋdaki/	$\mapsto$ [pəŋdaki]	(NOMCLIMB, [pəndaki])
$\vdots$	$\vdots$	/əŋbə/	$\mapsto$ [əmbə]	(NOMDIG, [pəŋgali])
		$\vdots$	$\vdots$	

λ		
NOM	$\mapsto$	/pəŋ/
CARRY	$\mapsto$	/bawa/
CLIMB	$\mapsto$	/daki/
:	:	:
NOMDIRECT	$\mapsto$	/pəŋarah/
NOMCARRY	$\mapsto$	/pəŋbawa/
NOMCLIMB	$\mapsto$	/pəŋdaki/
:	:	:

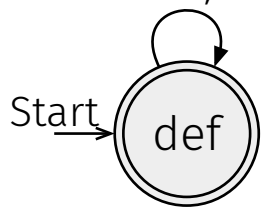
NOM:pəŋ, CARRY:bawa, CLIMB:daki, DIG:gali



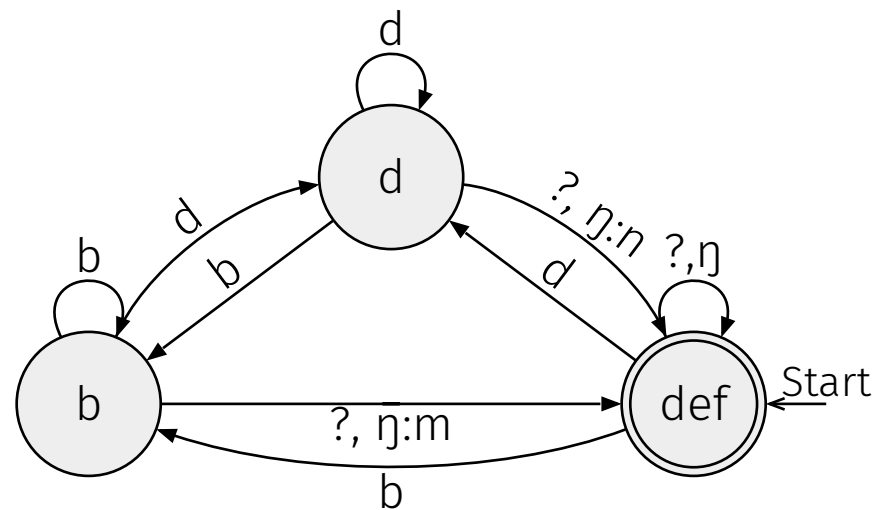
Finite-state machine for λ

- (Hua & Jardine, 2021), (Hua et al., 2021): λ is a **homomorphism**

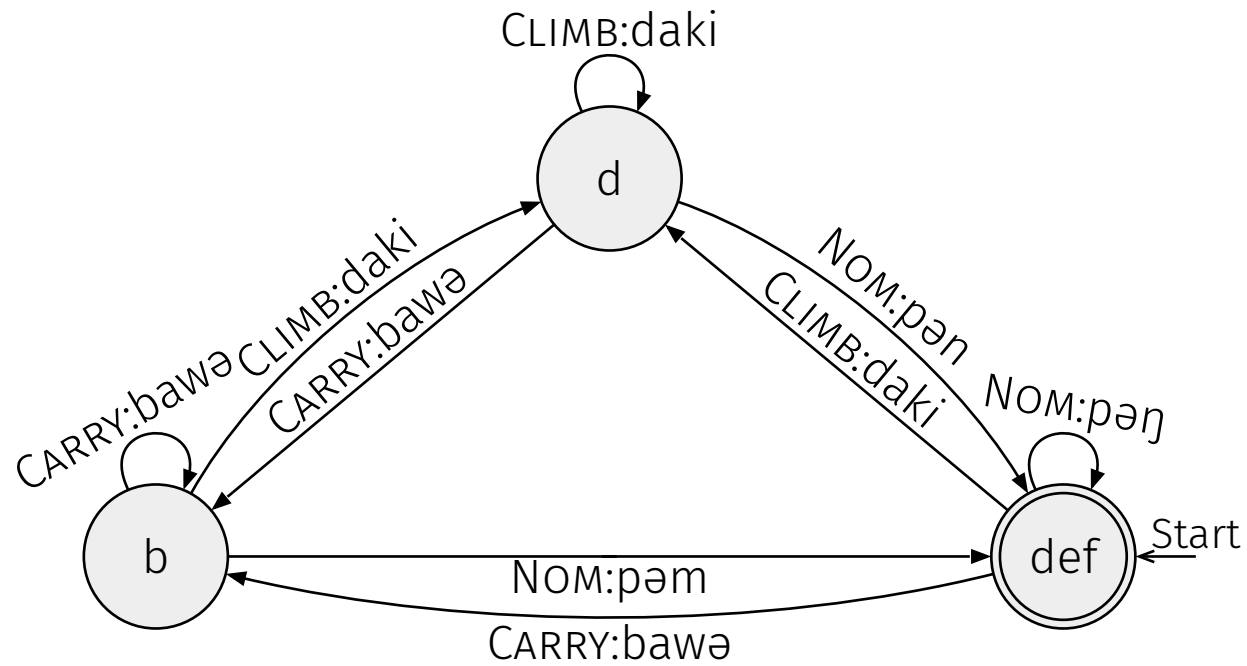
NOM:pəŋ, CARRY:bawa, CLIMB:daki, DIG:gali



Finite-state machine for λ



Finite-state machine for φ



FST for $\varphi \circ \lambda$

Hua & Jardine (2021), Hua et al. (2021)

- FST for $\varphi \circ \lambda$ is learnable from known techniques (Oncina et al., 1993)
- Learner then **decomposes** FST for $\varphi \circ \lambda$ into FSTs for φ and λ
- Learner **guaranteed** to identify φ and λ when:
 - λ is homomorphism
 - φ is ISL₂ and represents *one* process
 - Learner sees sufficient combinations of morphemes

Chandlee & Jardine (2025)

- **Natural classes** $\rightarrow \varphi$ can have multiple processes
- **Domain information** (Oncina & Varò, 1996) $\rightarrow$ we can learn from data with morphotactic restrictions

Chandlee & Jardine (2025)

• Johor Malay (Onn, 1976)

- ▶ Nasal place assimilation before obs.
- ▶ Post-nasal deletion of voiceless obs.
- ▶ Coda /r/ deletion
- ▶ Final /a/ raising
- ▶ Coda /k/ glottalization
- ▶ (other processes discussed later)

/pəŋborɔŋ/ ↦ [pəmborɔŋ] 'wholesaler'

/pəŋtali/ ↦ [pənali] 'dancer'

/kisar#/ ↦ [kisa] 'to answer'

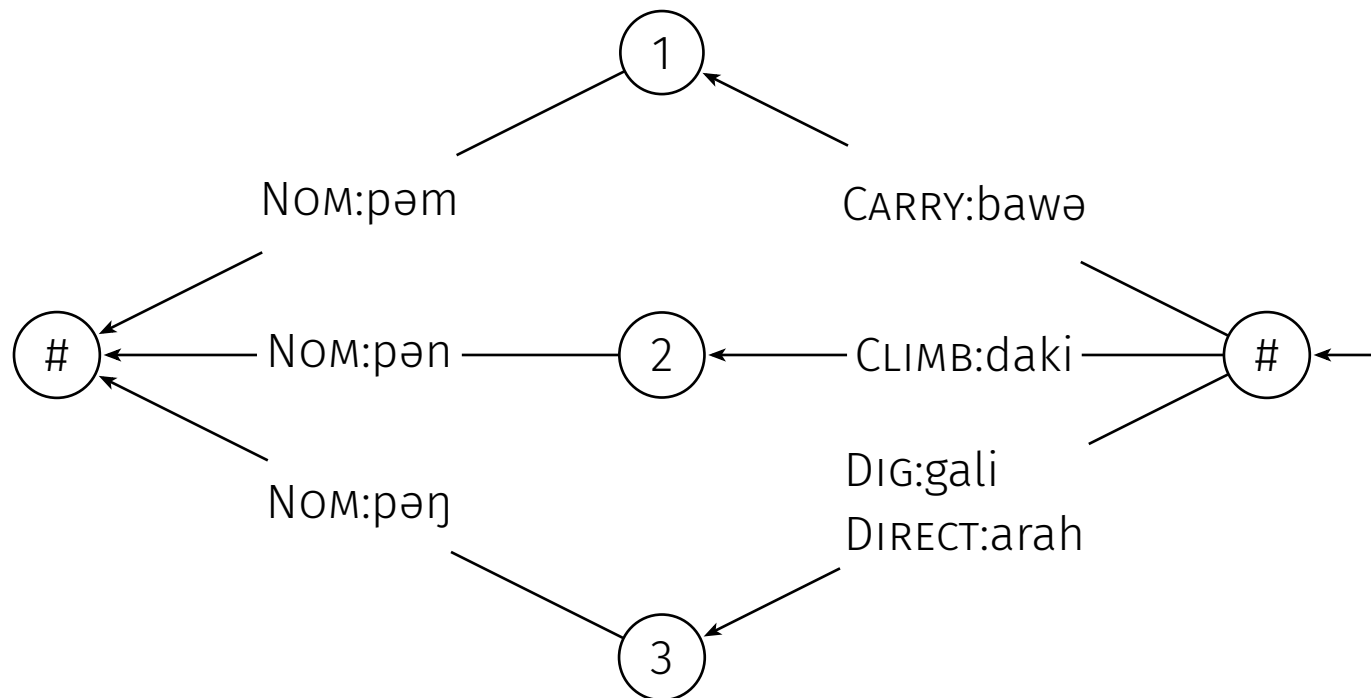
/bawa#/ ↦ [bawə] 'to answer'

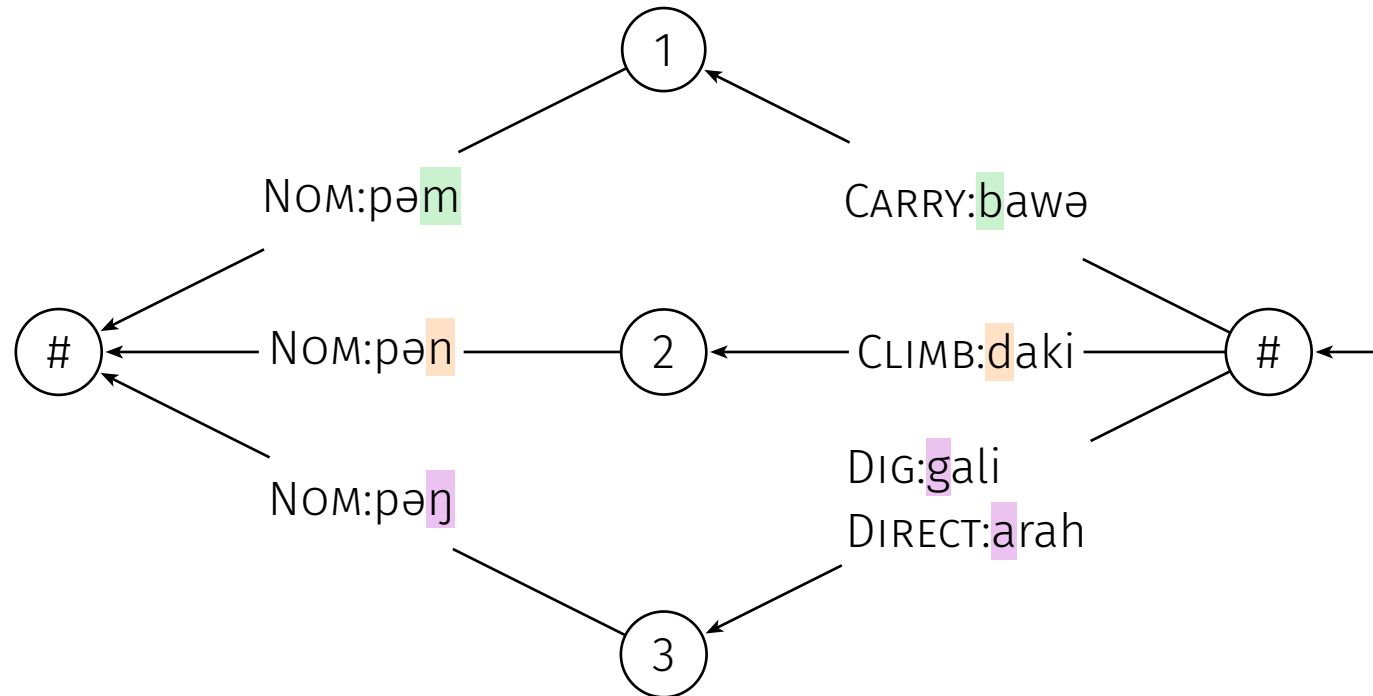
/masak#/ ↦ [masaʔ] 'to cook'

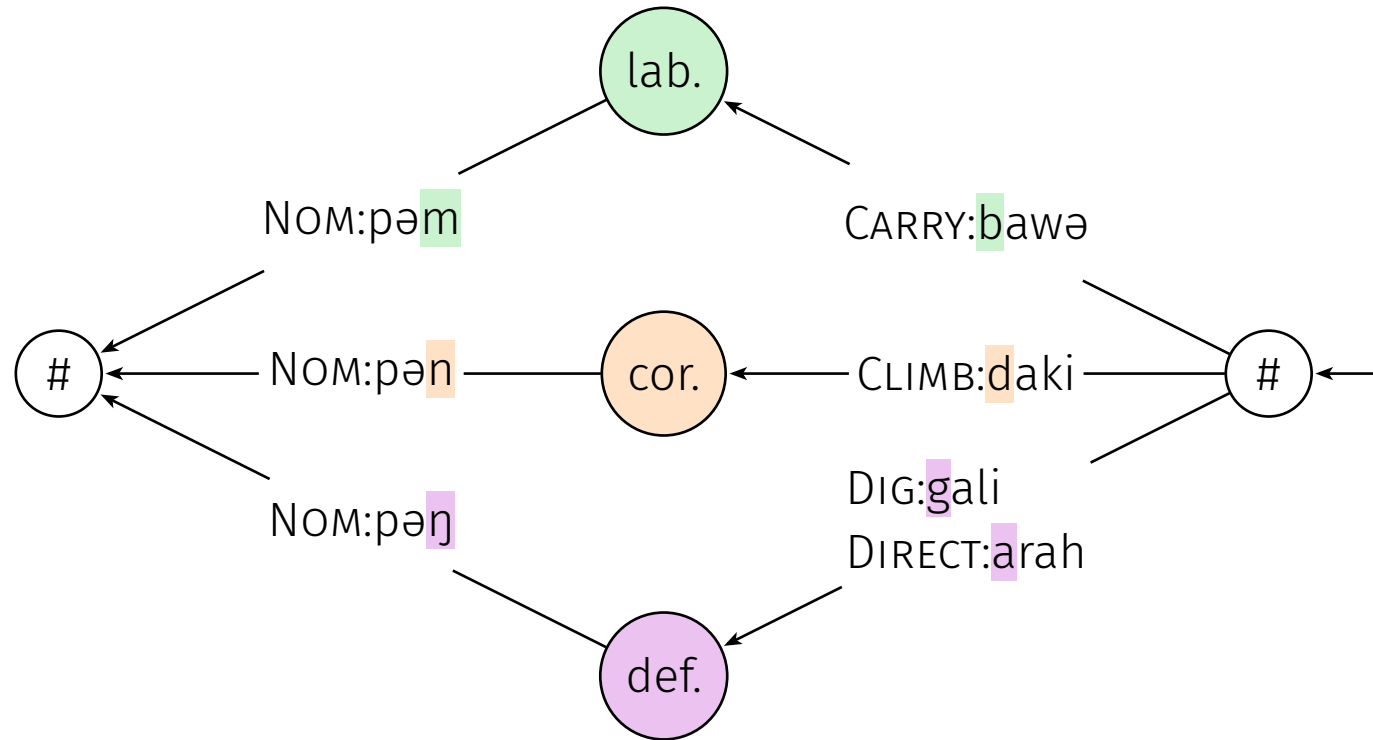
Input data

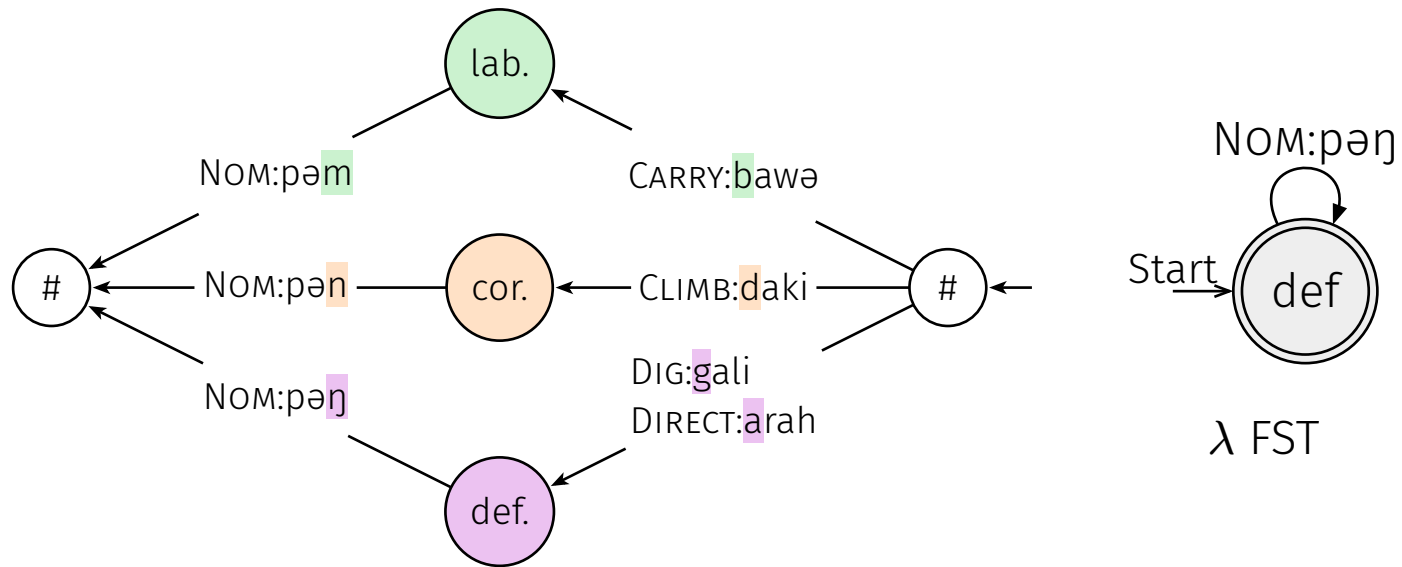
NOMDIRECT	pəŋarah
NOMCARRY	pəmbawə
NOMCLIMB	pəndaki
NOMDIG	pəŋgali

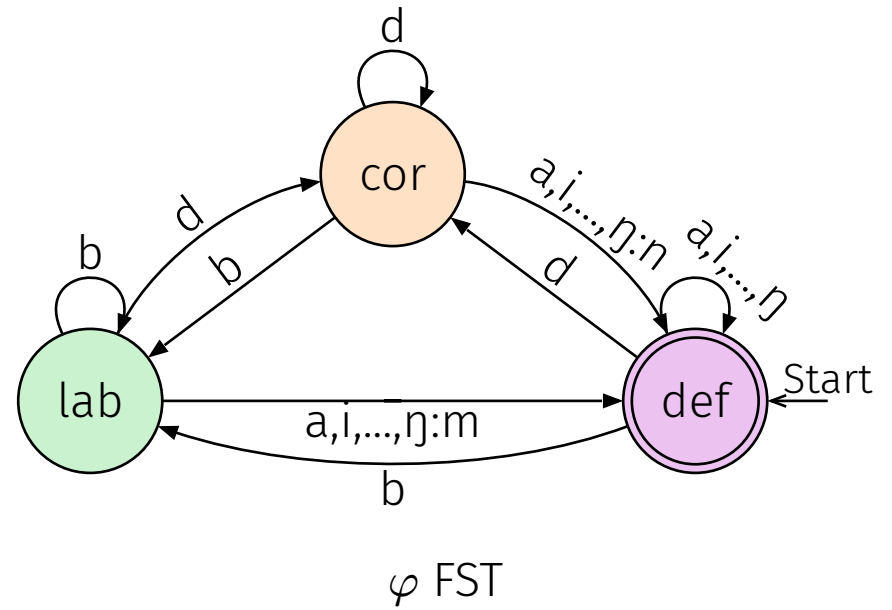
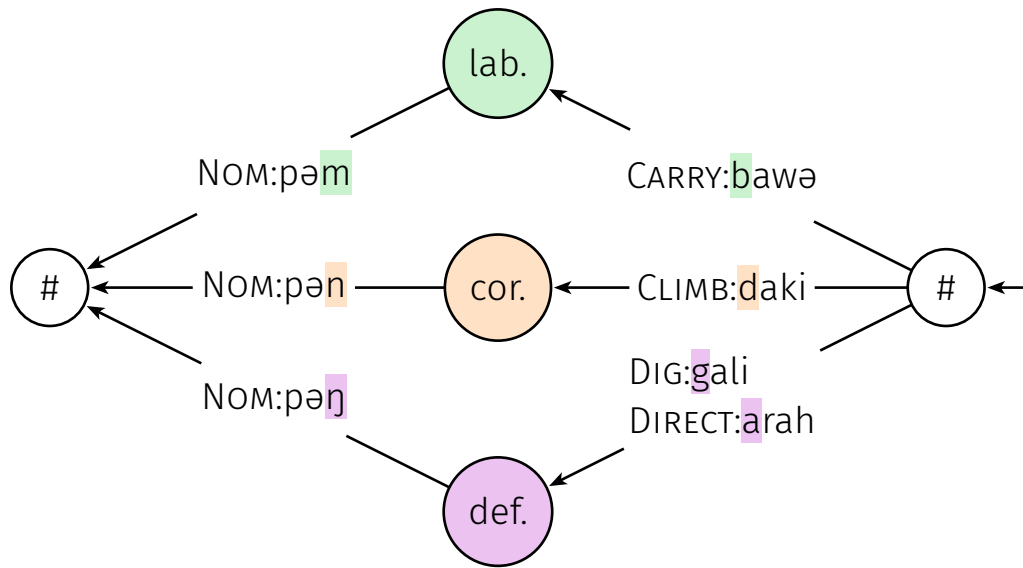
NOMDIRECT pəŋarah
NOMCARRY pəmbawə
NOMCLIMB pəndaki
NOMDIG pəŋgali











Chandlee & Jardine (2025): Dataset from Johore Malay (Onn, 1976)

- **Input:** 14 stems × 5 affixes:

ACT /məŋ-/

NOM /pəŋ-/

CAUSE /-i/

BEN /-kan/

NOMSUF /-an/

- **Output:** Correct lexicon and grammar representing all five processes
 - This includes counter-feeding of /r/-deletion and /a/-raising
/bakar/ ‘to burn’ $\mapsto$ [baka] *[bakə]

- **Inductive principles:**
 - Phonological functions are **input-local**
 - Phonological functions refer to **natural classes** (in the environment)
 - Error-checking to discover the coda generalizations

Limitations

- Bidirectional processes:
 - Nasal deletion before sonorants
 - Epenthesis for vowel hiatus

/pəŋrayu/ ⇨ [pərayu]

‘appeal’

/bantuan/ ⇨ [bantuwan]

‘aid, relief’

- **Other solutions**

- Constraint-based grammars (Tesar 2014; Jarosz 2006, 2013; Wang and Hayes 2022)
- Minimum-description length (Rasin et al., 2020, 2021)
- Bayesian approaches (Barke et al., 2019; Ellis et al. 2022)

- All require some **restriction** on hypothesis space

- Our learner is:

- Based on *provably correct* algorithms
- Efficient

Ongoing work

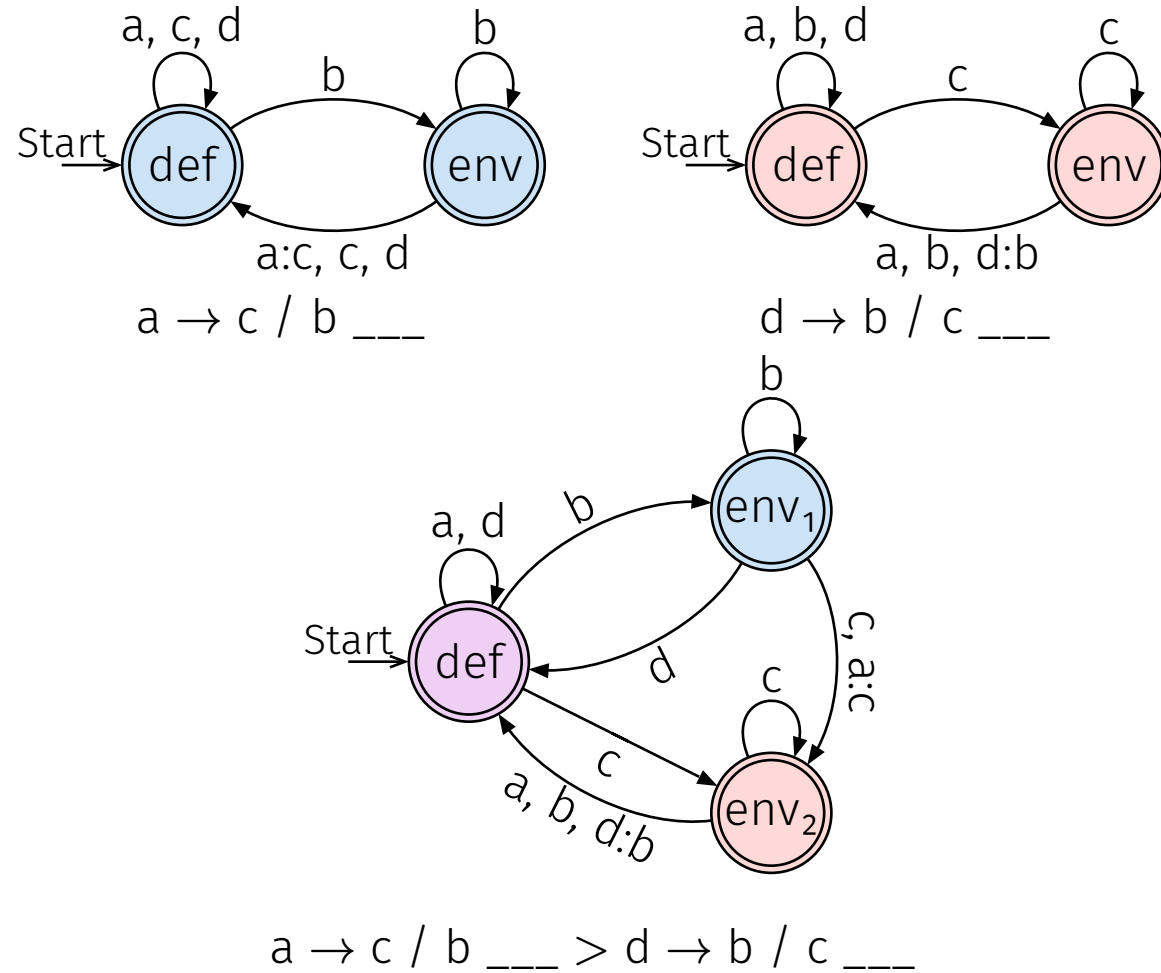
- Interacting processes (Wang)
- Two-sided contexts (Anthony, Jeyaram, Kandepalli)
- Long-distance processes (Czarnecki, Franz)
- Abstract lexical entries (Franz)
- Non-concatenative morphology (Helfand)
- Contexts longer than 2 (Evans)
- Testing on large corpus of datasets (Heinz et al.)
- ...

Learning Rule Interactions

(Wang and Jardine, 2026)



Chenli Wang



Tiers and UR learning

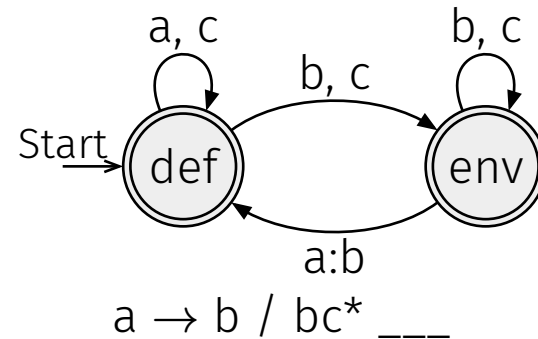
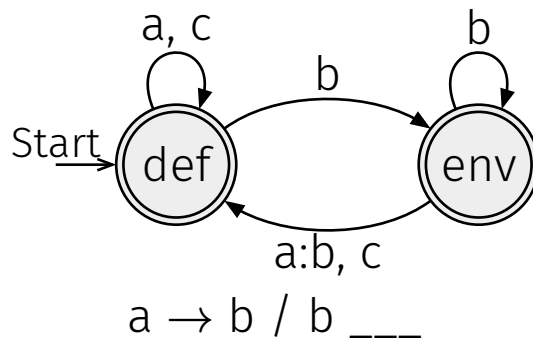
(in prep)



Vincent Czarnecki



Eli Franz



Implementation with graphical user interface

Neha Jeyaram

Manish Kandepalli

Ezra Anthony

Give it a spin!

 **README**  GPL-3.0 license 

ur_learning

Algorithms based on subregular hypothesis for induction of phonological grammars and sets of underlying forms from morphophonological paradigms. (In progress.)

`ostia.py`, `fst_object.py`, and `helper.py` courtesy of [@alenaks](#)'s SigmaPie package.

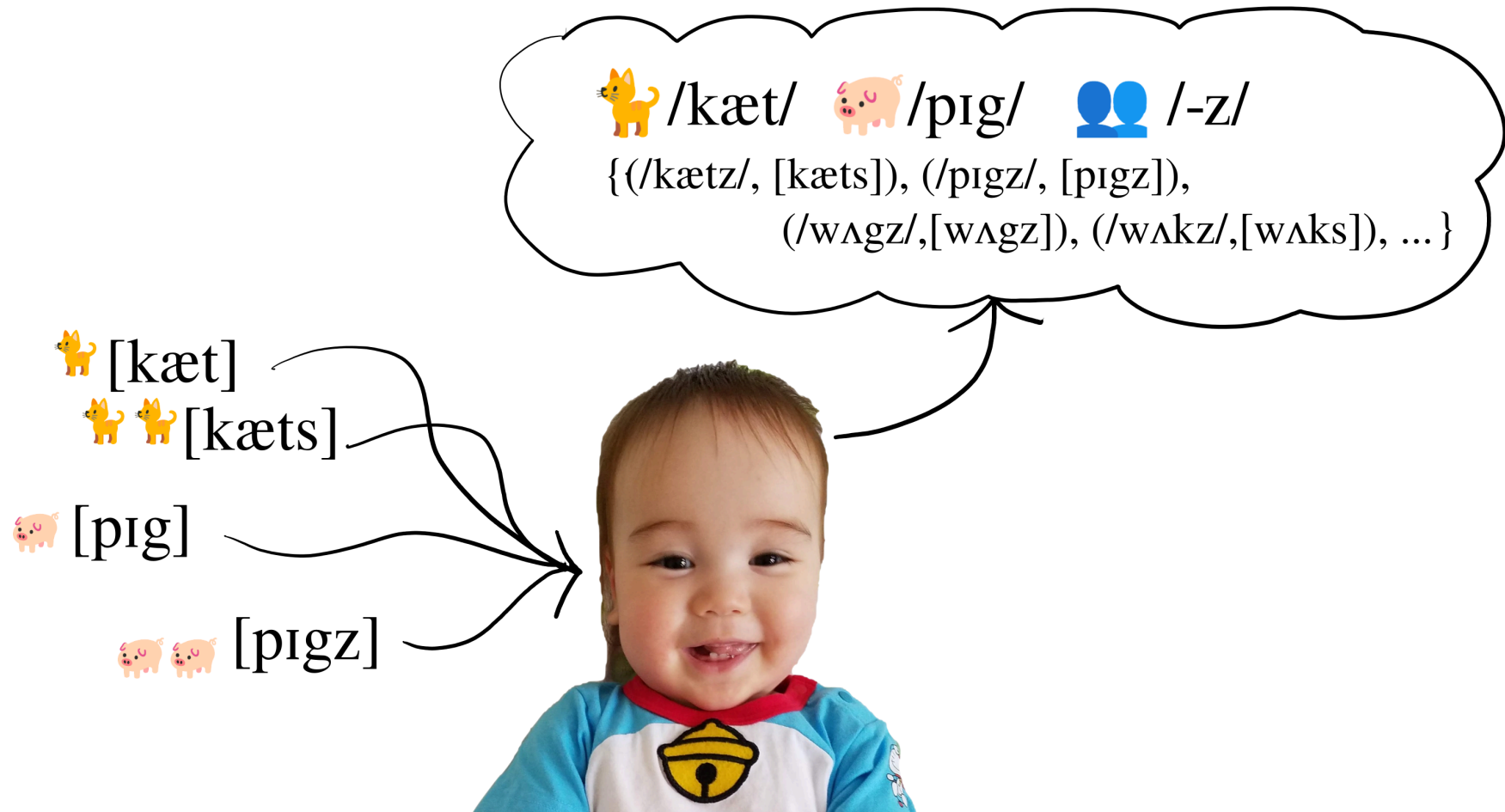
The Simplex Input Strictly 2-Local Decomposition Learning Algorithm of [Hua & Jardine 2021](#) is implemented in [si2dla.py](#).

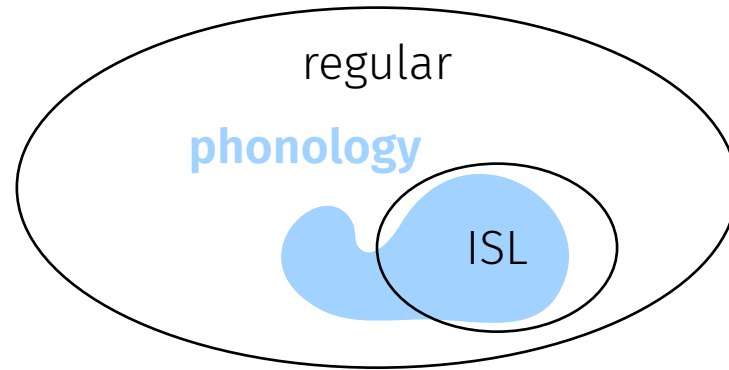
The Input (Strictly Local) Decomposition Learning Algorithm (To Appear) is implemented in [idla.py](#).

Other files are experimental variations on this algorithm:

File	Description
<code>so2dla.py</code>	An Output Strictly 2-Local version of the SI2DLA
<code>fsi2dla.py</code>	A featural version of the SI2DLA

To conclude





Thank you!