



Subregular Induction of Underlying Representations and a Phonological Grammar

Wenyue Hua, Adam Jardine, and Huteng Dai

{wenyue.hua,adam.jardine,huteng.dai}@rutgers.edu

Department of Linguistics, Rutgers University

Goals and Background

- Project goal: the **simultaneous inference** of URs and a grammar from SRs in a morphological paradigm.
- The **Input Strictly Local** functions provide a structure that can solve this problem.
(Chandlee and Heinz, 2018)

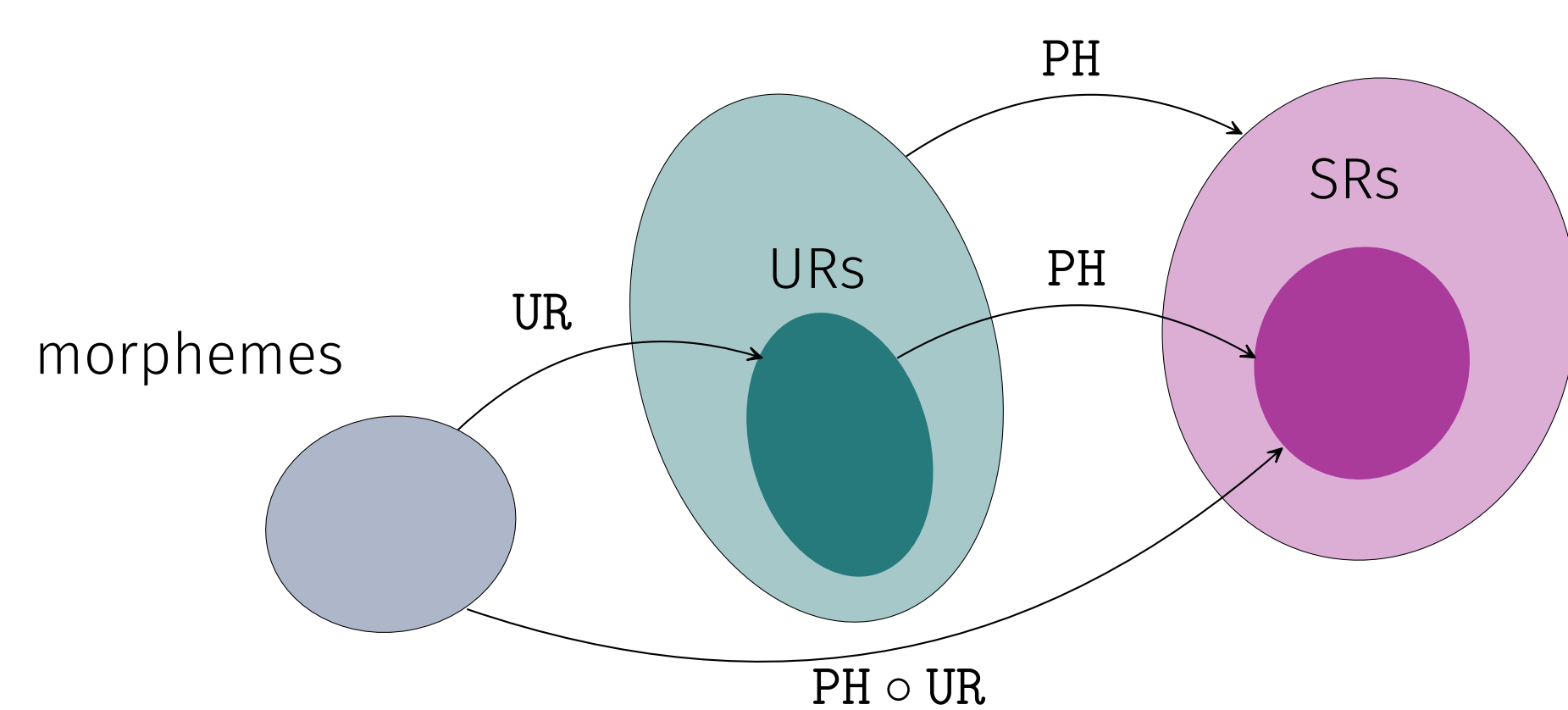
Primary result

- The learner induces UR and phonological grammar from a range of ISL₂ functions (ISL functions for $k = 2$).
e.g. progressive and regressive assimilation/dissimilation, deletion, epenthesis, metathesis, and opacity.

Learning Problem

- M : finite set of morphemes {CAT, DOG, ..., PL}
- Σ : finite set of segments {a, b, β, ..., z}
- UR function: maps one morpheme to one UR
 $UR : M^* \rightarrow \Sigma^*$
- PH function: maps URs to SRs
 $PH : \Sigma^* \rightarrow \Sigma^*$

UR (CAT)	= kæt	PH (kæt)	= kæt
UR (PL)	= z	PH (dɔgz)	= dɔgz
UR (CAT-PL)	= kætz	PH (kætz)	= kæts
...	...	PH (bnɪkz)	= bnɪks
...



- Given a finite sample of $PH \circ UR$, how do we identify PH and UR?

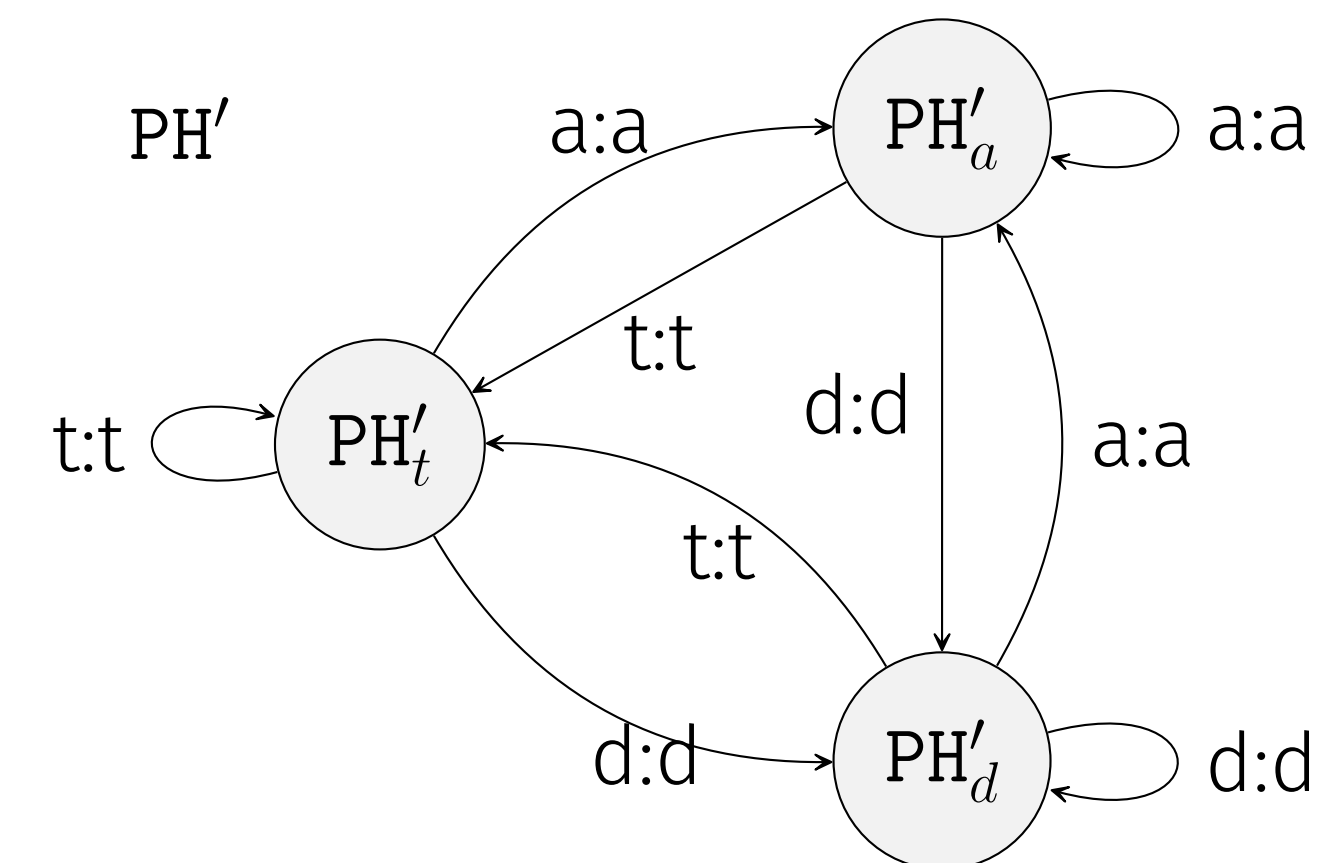
(CAT-PL, kæts), (DOG-PL, dɔgz), ..., (BOOK-PL, buks)

Initialization

- Running example $D \subset PH \circ UR(M^*)$

Sample of $PH \circ UR$					
w	$PH(UR(w))$	w	$PH(UR(w))$	w	$PH(UR(w))$
$r_1 s_1$	tatta	$r_2 s_1$	tadda	$r_3 s_1$	ata
$r_1 s_2$	tatda	$r_2 s_2$	tadda	$r_3 s_2$	ada
$r_1 s_3$	tata	$r_2 s_3$	tada	$r_3 s_3$	aa

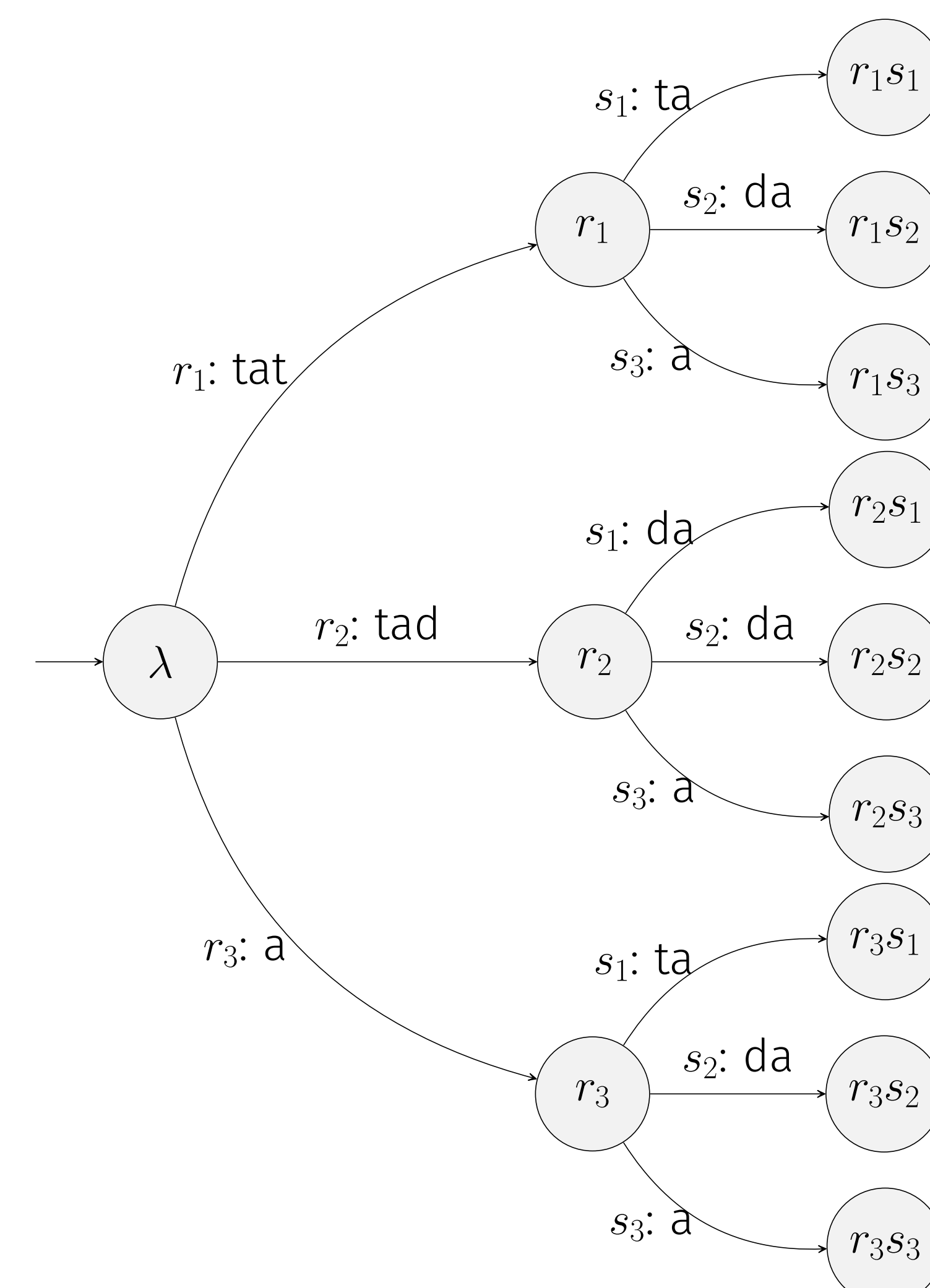
- Initialize PH' to the identity function
 $PH'(tatta) = tatta$, $PH'(tatda) = tatda$, etc.



- Initialize UR' to a **prefix tree transducer** representing D : segmentation based on **longest common prefix (lcp)**.

$$lcp(\{aab, aaba, aac\}) = aa$$

$$lcp(\{bac, abc\}) = \lambda$$



Inconsistency detection

If a morpheme is mapped to multiple SRs, the learner detects this inconsistency.

$$r_1: tat \quad r_2: tad \quad r_3: a$$

$$s_1: ta, da \quad s_2: da \quad s_3: a$$

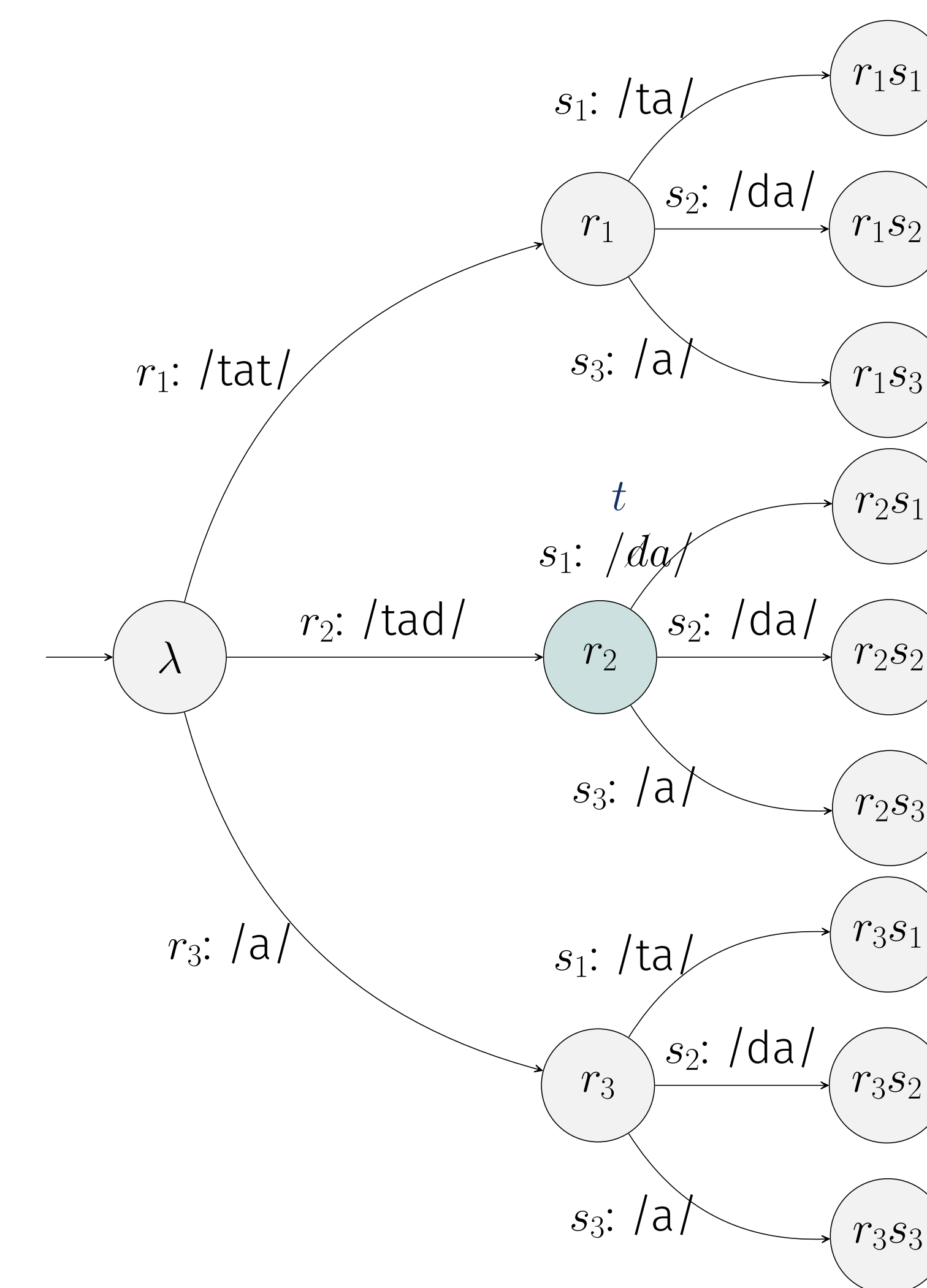
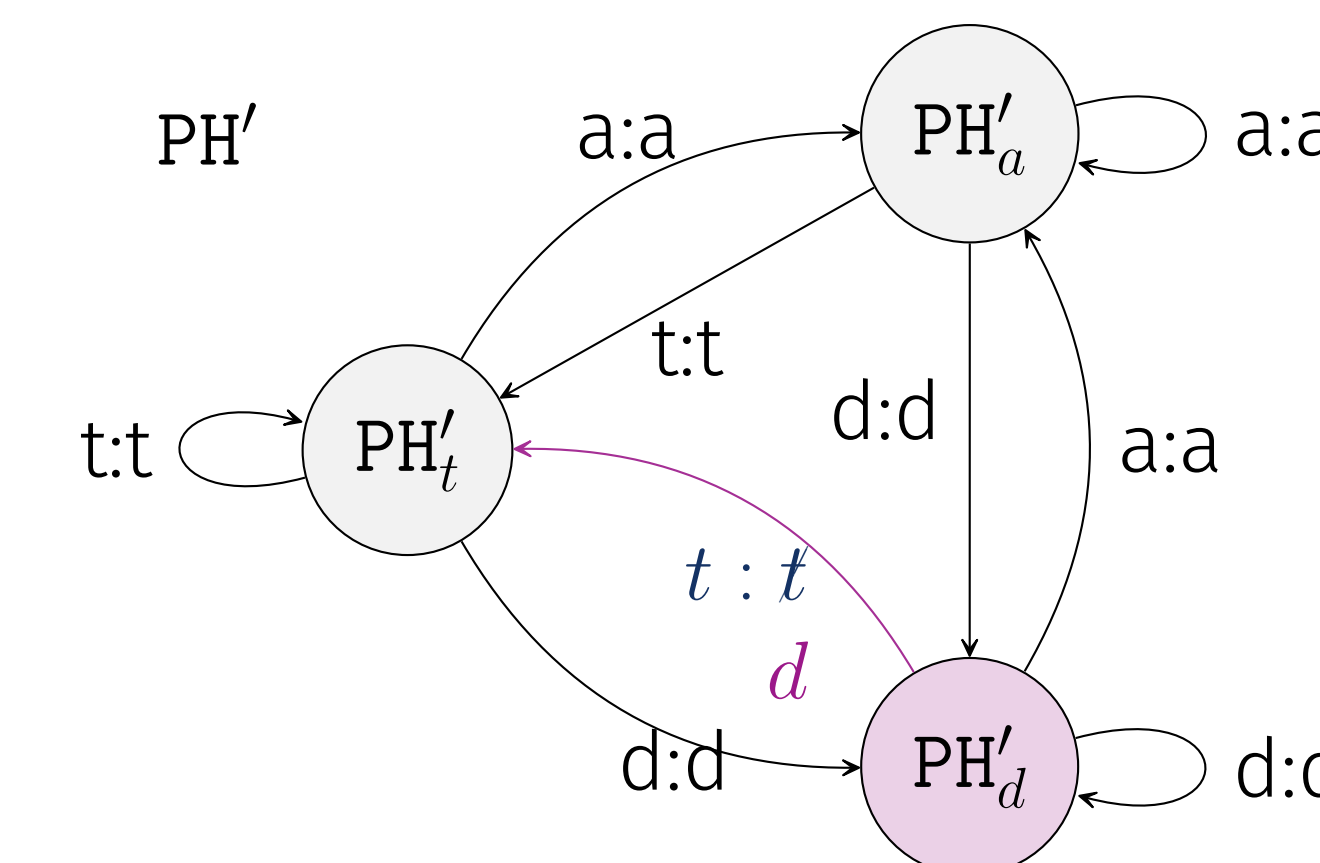
Environment collection

- ISL provides environment information
- t is the most **informative** form \rightarrow UR.

ws_1	env.	s_1
$r_1 s_1$	tat	ta
$r_3 s_1$	a	ta
$r_2 s_1$	tad	da

Modification

The modifications on PH' and UR' are opposite.



Take-home

From an **abstract** and **principled** perspective, learning is possible given the basic principles:

- Restrictive, structured hypothesis space
- Complementary distributed allomorphs
- A surface-driven set of URs
- Single form: one morpheme \rightarrow one UR
(Albright, 2002)

Future work

- Long-distance processes can be captured by different classes of subsequential functions with a similar structure;
- One example: output strictly-local class also has a restricted state structure;
(Chandlee et al., 2015)
- Abstract URs may be learnable when input alphabet is larger than output alphabet (and thus allows larger categories).

Selected References

Albright, Adam C (2002). *The identification of bases in morphological paradigms*. PhD thesis, University of California, Los Angeles.

Chandlee, Jane, Eyraud, Rémi, and Heinz, Jeffrey (2015). Output strictly local functions. In Kornai, Andras and Kuhlmann, Marco, editors, *Proceedings of the 14th Meeting on the Mathematics of Language (MoL 14)*, pages 52–63, Chicago, IL.

Chandlee, Jane and Heinz, Jeffrey (2018). Strict locality and phonological maps. *Linguistic Inquiry*, 49(1):23–60.

Acknowledgements: We thank attendees of the Rutgers MathLing group, NECPhon 2019, and in particular Jeff Heinz, Charles Reiss, Bruce Tesar, Adam McCollum, and Colin Wilson for their insightful comments.



Take a picture to download the extended abstract