The Interaction Between Learning Algorithms and Formal Language Theory

Caleb Belth

Computational and Approach to Learning

Formal-Language-Theoretic analysis of linguistic generalizations

- Starts with **computational-level analyses** of linguistic structures
- Puts **constraints** on learning algorithms, allowing **efficient learning**

Algorithmic Approach

Algorithmic Approach** to **Phonological** ***Rules** & **Representations

1) Identify **tools** available to a learner
2) Propose **learning algorithm(s)** that use these
3) Eval **accuracy** & developmental + experimental **predictions**
4) **Rules** and **representations** constructed along the way gain **algorithmic**, **learning-based** support

Belth 2023. *Towards an Algorithmic Account of Phonological Rules and Representations*. PhD Dissertation.

Complementary Approaches

Corroboration through convergence of approaches

- *Example*: iterative removal to form tiers

Background

Which, of attested classes, will a learner **construct**?

- *Example*: will a learner construct a tier-local or a string-local generalization if both are compatible with learning data?

Today

How differences in **distributional properties** matter

- *Example*: Different behavior for computationally-equivalent processes in Germanic voicing alternations

Future

Background

Complementary Approaches

Corroboration through convergence of approaches

- *Example*: iterative removal to form tiers

Background

Which, of attested classes, will a learner **construct**?

- *Example*: will a learner construct a tier-local or a string-local generalization if both are compatible with learning data?

Today

How differences in **distributional properties** matter

- *Example*: Different behavior for computationally-equivalent processes in Germanic voicing alternations

Future

Alternations and Tier Locality

Phonological segments **alternate** in a **predictable** way

[dɑl-lɑr-ɯn] branch-PL-GEN (Kabak, 2011, p. 3)

[jer-ler-in] place-PL-GEN (Kabak, 2011, p. 3)

[ip-ler-in] rope-PL-GEN (Nevins, 2010, p. 28)

Dependencies cross **intervening** consonants

Turkish Language

Back Vowels: {ɑ, ɯ}

Front Vowel: {e, i}

Kabak, Bariş. (2011). Turkish vowel harmony. *The blackwell companion to phonology*, 1–24.
Nevins, Andrew. (2010). *Locality in vowel harmony*. Vol. 55. Mit Press.

Alternations and Tier Locality

Phonological segments **alternate** in a **predictable** way

/baʔ-s-e/ ⎓ [baʔse] ‘he bought’

/ʔuʃ-s-it/ ⎓ [ʔuʃʃit] ‘I cooked’

/ʒaʔ-s-it/ ⎓ [ʒaʔʃit] ‘I arrived’

/ʃed-er-s-it/ ⎓ [ʃederʃit] ‘I was seen’

Dependencies cross **intervening** non-sibilants

Aari Language

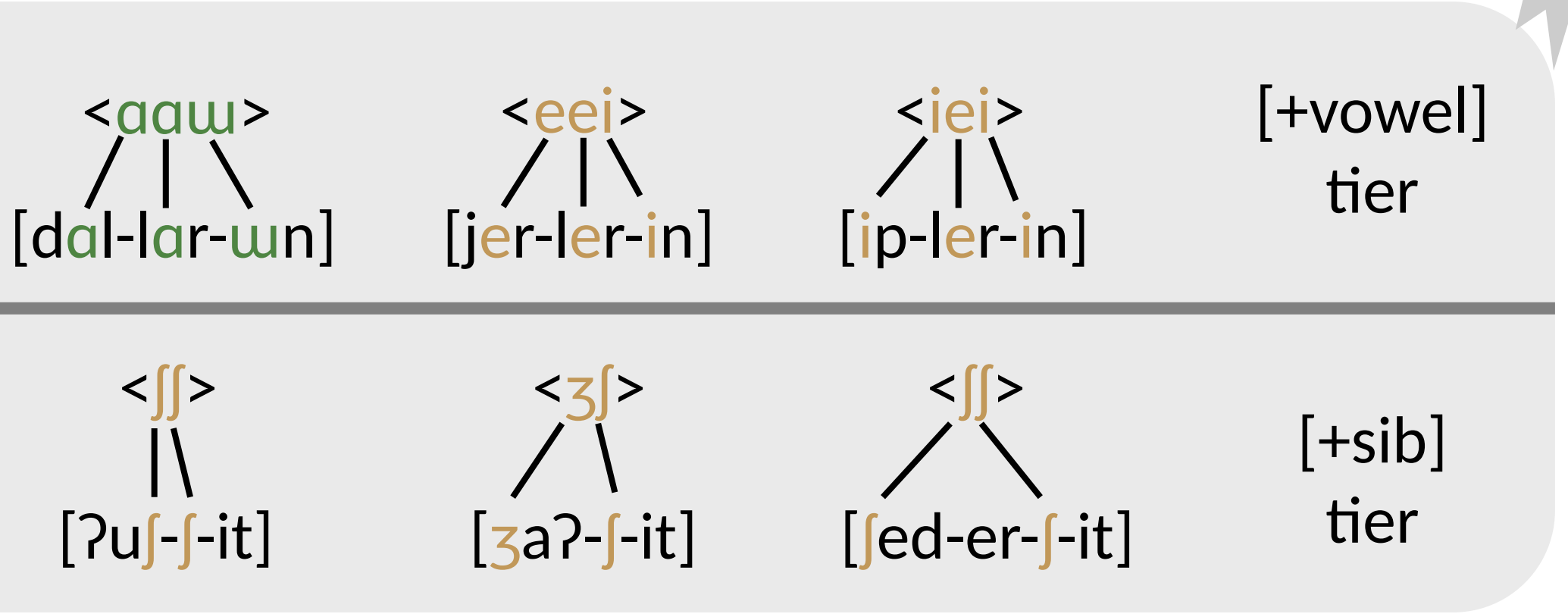
McMullin, Kevin James. (2016). *Tier-based locality in long-distance phonotactics: learnability and typology*. Ph.D. thesis, University of British Columbia.

Hayward, Richard J. (1990). Notes on the aari language. *Omotic language studies*, 425–493.

Alternations and Tier Locality

2TSL

Dependent segments are **adjacent** on some **tier**



Goldsmith, John Anton. (1976). *Autosegmental phonology*. Ph.D. thesis, Massachusetts Institute of Technology.
Clements, George N. (1976). *The autosegmental treatment of vowel harmony*.

Statistical Learning Studies

Infants & adults can track **adj dependencies** across many types of items:

- Syllables (Saffran et al., 1996, 1997; Aslin et al., 1998)
- Words (Gómez, 2002)
- Morphemes (Santelmann & Jusczyk, 1998)
- Non-linguistic tones (Saffran et al., 1999)
- Shapes (visual domain) (Fiser & Aslin, 2002)

Independent Mechanism
Tracking Adj. Dep.

Ability to track non-adj dependencies **develops later** and is **secondary resort** (Gomez, 2002; Gomez & Maye, 2005)

"It is as if learners are attracted by adjacent probabilities long past the point that such structure is useful."
- Gomez & Maye (2005, p. 199)

Background

Hypothesis: in alternation learning, learners track adjacent dependencies before…

- Changing representations (Belth, Accepted *LI*)
- Looking further (Belth, In Press *Phonology*)

Hypothesis as Learning Algorithm: If an alternation ***cannot be predicted*** from adjacent segs, **delete** and **try again**

Uses Tolerance Principle (Yang, 2016) to handle **sparsity** and **exceptions**

Caleb Belth. In Press. A learning-based account of local phonological processes. *Phonology*.
Caleb Belth. Accepted. A learning-based account of phonological tiers. *Linguistic Inquiry*.

Corroboration

Burness & McMullin (2019; 2021)

- Properties of 2TSL functions provide conditions under which **removing a segment from the tier** is provably correct
- Assuming target function is 2TSL **guarantees** *efficient* & *correct* learning if characteristic sample present

Iterative Removal From Representation 

Belth (Accepted, *LI*)

- Restricting attention to **adj dependencies** & **deleting when generalization fails**
- >0.98 **accuracy** on natural language despite **sparsity and exceptionality**
- Predicts **human behavior** in art. lang. exps. (Finley 2011; McMullin & Hansson 2016)
- Consistent with **developmental patterns** (Altan, 2009)

Corroboration

For the **extinction of dinosaurs via an asteroid** to go from ***reasonable conjecture*** to ***fact of natural history***, it took

- Recognition of a **mass-extinction** much larger than just the dinosaurs
- Discovery of **iridium** world-wide at the right layer of rock
- Discovery of **crater** of appropriate age and size

Complementary Approaches

Corroboration through convergence of approaches

- *Example*: iterative removal to form tiers

Background

Which, of attested classes, will a learner **construct**?

- *Example*: will a learner construct a tier-local or a string-local generalization if both are compatible with learning data?

Today

How differences in **distributional properties** matter

- *Example*: Different behavior for computationally-equivalent processes in Germanic voicing alternations

Future

Do learners really track adjacency first?

Learning **algorithm** first tracks **only adjacency**

if adjacent and non-adjacent dependencies are **equally robust**, learners will **not track non-adjacent dependencies**

Proposal: design artificial language where this is the case

Do learners really track adjacency first?

Stem (SG) ends with:

(1) Voiced Consonant {b, d} and back vowel {u, ɔ}

PL takes [-f]

E.g. [bibu-f], [bɔtbɔ-f]

(2) Voiceless Consonant {p, t} and front vowel {i, ɛ}

PL takes [-ʃ]

E.g., [pɔti-ʃ], [dubtɛ-ʃ]

Do learners really track adjacency first?

1) Train / Exposure Phase:

{b, d}{u, ɔ}-final and {p, t}{i, ɛ}-final SG, PL pairs — CV**CV**-[**f**] / CV**CV**-[**ʃ**]

- ([bɔtbɔ], [bɔtbɔ-f]), …, ([pɔti], [pɔti-ʃ]), …

2) Test Phase: (SG, followed by 2AFC between PL forms)

- SG: [dupu] PL: [dupuf] OR [dupuʃ]? …
- Some follow **training** pattern to confirm generalization learned

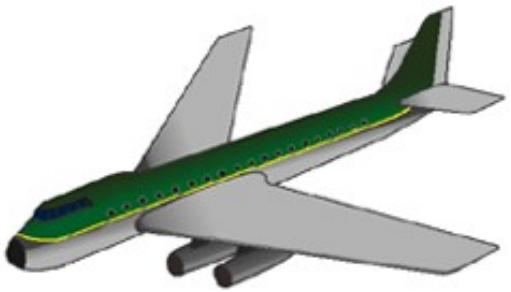
  CV**CV**-[**f**] CV**CV**-[**ʃ**]
- Others **flip** so that {b, d} goes with {i, ɛ} and {p, t} with {u, ɔ}

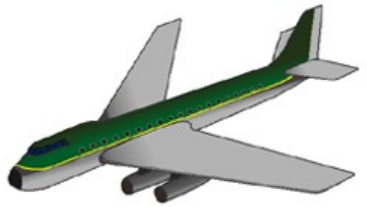
  CV**CV**-? CV**CV**-?

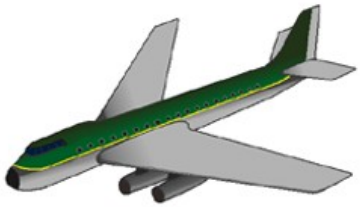
Experiment Example

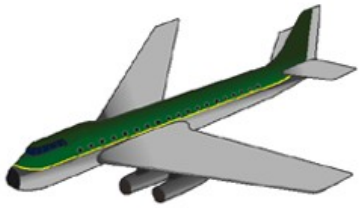
Training Phase

CVCV-[f]
CVCV-[ʃ]









[bɔtbɔ]

[bɔtbɔ-f]

Experiment Example

Training Phase

CVCV-[f]
CVCV-[ʃ]

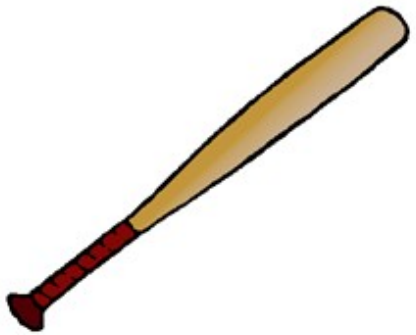
[pɔti]

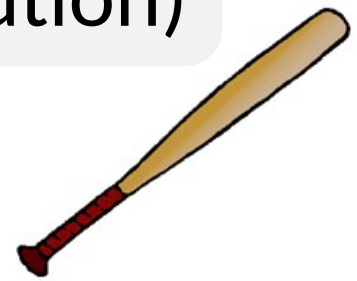
[pɔti-ʃ]

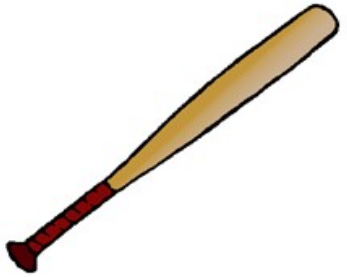
Experiment Example

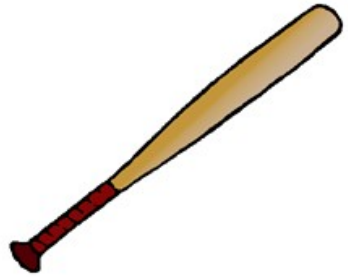
Testing Phase
(**Training** Distribution)

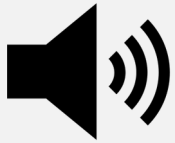
CV**CV**-[**f**]
CV**CV**-[**ʃ**]











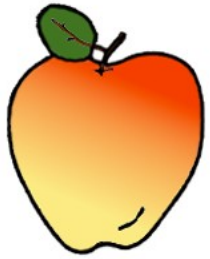
[bibu]

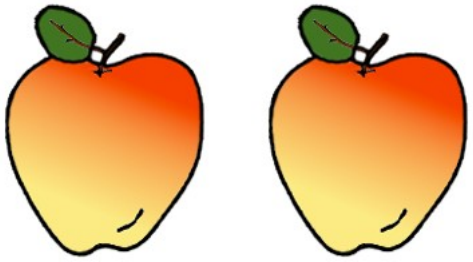
[bibu-f] or [bibu-ʃ]

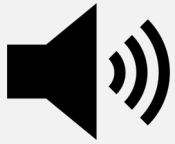
Experiment Example

Testing Phase
(New Distribution)

CVCV-[f]
CVCV-[ʃ]







[dupu]

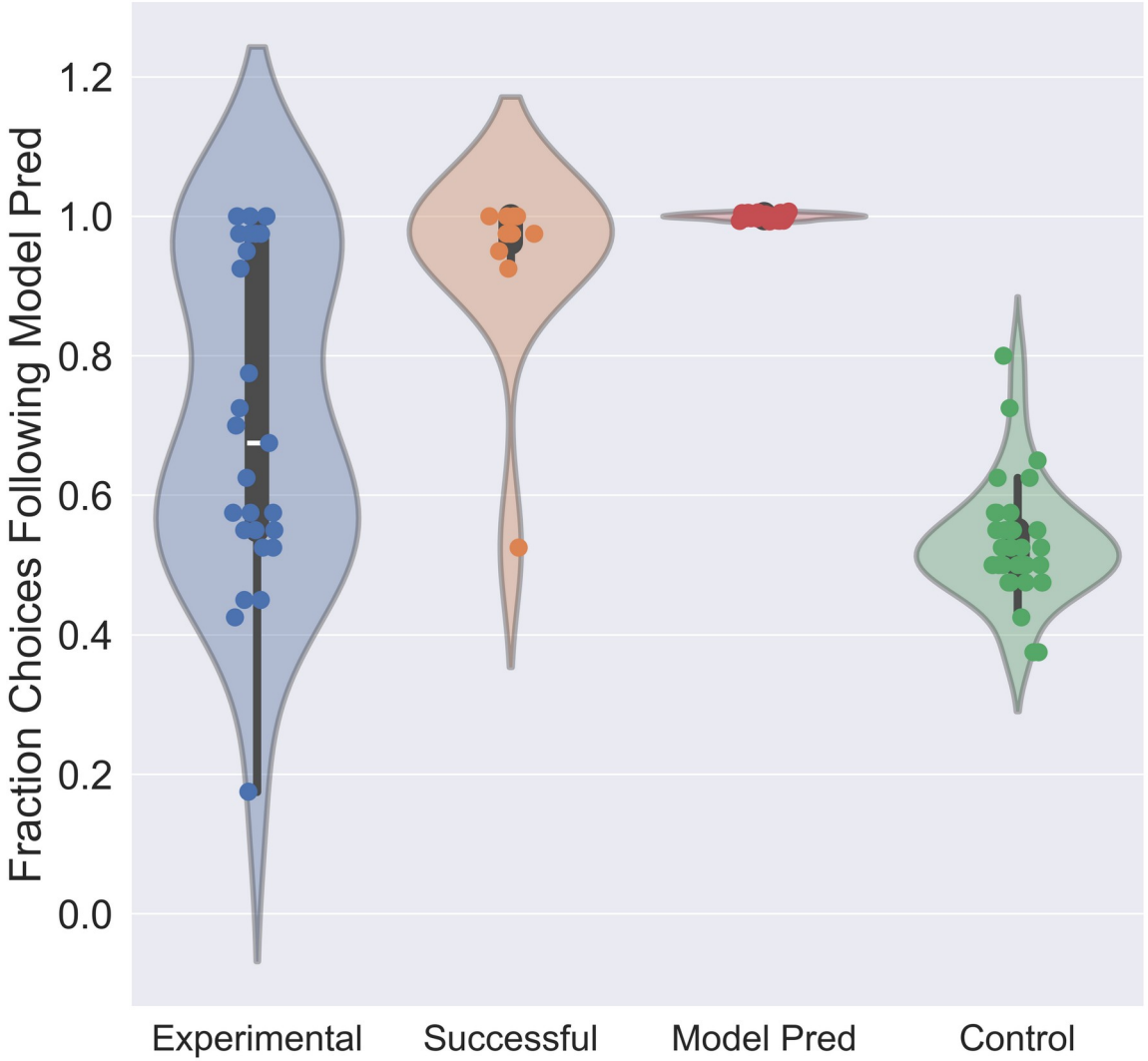
[dupu-f] or [dupu-ʃ]

Do learners really track adjacency first?

	Generalization	Training-Distribution	New Distribution
ISL ()	{u, ɔ}# [-f] Elsewhere [-ʃ]	[bibu] [bibu-f] [dubtɛ] [dubtɛ-ʃ]	[dupu] [dupuf] *Model Prediction*
ISL ()	{b, d}{u, ɔ}# [-f] Elsewhere [-ʃ]	[bibu] [bibu-f] [dubtɛ] [dubtɛ-ʃ]	[dupu] [dupuʃ]
TSL ()	{b, d}[*]# [-f] Elsewhere [-ʃ]	[bibu] [bibu-f] [dubtɛ] [dubtɛ-ʃ]	[dupu] [dupuʃ]

ISL & **TSL** are both ***lower bounds*** on phonology

Do learners really track adjacency first?



Exp Group made choice consistent with ***adjacent segment*** (e.g., [dupuf] over [dupuʃ]) much more than **Control Group**

Statistical analysis (mixed-effects logistic regression) corroborates

Effect is nearly categorical for **successful** learners

"Successful" = training-distribution accuracy cannot be attributed to chance

Complementary Approaches

Corroboration through convergence of approaches

- *Example*: iterative removal to form tiers

Background

Which, of attested classes, will a learner **construct**?

- *Example*: will a learner construct a tier-local or a string-local generalization if both are compatible with learning data?

Today

How differences in **distributional properties** matter

- *Example*: Different behavior for computationally-equivalent processes in Germanic voicing alternations

Future

Final Devoicing

- Final devoicing Dutch & German [bɛt] “bed” ~ [bɛdən] “beds”
  - **ISL ()**
- Yet, **Dutch** children show **no evidence of productivity** (Zamuner *et al.*, 2006, 2012; Kerkhoff 2007) while **German kids do** (van de Vijver & Baer-Henney 2014)…
- …and German learners appear to develop knowledge of alternation **more quickly** (Buckler & Fikkert, 2016)
- Distributional properties quite different
- Belth (2023) & ongoing work provides potential explanation for how distributional differences yield developmental differences

van de Vijver, R., & Baer-Henney, D. (2014). Developing biases. Frontiers in psychology, 5, 634.
Buckler, H., & Fikkert, P. (2016). Dutch and German 3-year-olds’ representations of voicing alternations. Language and Speech, 59(2), 236-265.