# Spheres and Spaghetti: Generalization and Exceptionality in Phonotactic Acquisition

**Sarah Brogden Payne**
sarah.payne@stonybrook.edu

SYNC
March 4, 2023

# **Background:** Motivation

| | Attested | Unattested |
|---|---|---|
| **Licit** | *spot* | *wug* |
| **Illicit** | *sphere* | *bnick* |

- Suggests that ***sphere*** should pattern like ***bnick***
- ***sphere*** patterns like ***spot***
  - **Borrowings**
  - **New words**
  - **Production errors**

# Proposal

- ***sphere*** and ***spot*** are both **licit**
  - ***spot*** is **fully-licit**
  - ***sphere*** is **marginal**
- Illicit forms are ***always unattested***
- Licit forms can be attested or unattested

| | | Attested | Unattested |
|---|---|---|---|
| Licit | Fully-Licit | *spot* | *wug* |
| | Marginal | *sphere* | *spheal* |
| Illicit | | --- | *bnick* |

# **Proposal:** Degree of Specification

Fully-licit vs. marginal forms: **degree of specification**

## Underspecified: /#sp/

- Occurs before a ***wide range of vowels***
  - *spat, spell, spot, sputter*
- Belongs to **/#-[s]-[voiceless-stop]/**
  - {/#sp/, /#st/, /#sk/}

## Fully-Specified: /#sf/

- Occurs before a ***limited number of vowels***
  - *sphere, sphinx*
- Only similar onset = /#sv/
  - *svelte*

Evidence for early underspecification in phonological learning

# Proposal

- I propose a **recursive model of learning phonotactic generalizations** using the **Tolerance-Sufficiency Principle**
  - ***Increases the specification of sequences*** during learning
  - Contrasts ***fully-licit*** and ***marginal forms*** via ***degree of specification***
  - Learns ***positive grammar*** from ***positive data***
- Test this model on English complex onsets
  - Show that it learns ***plausible phonotactic sequences***

# **Evidence:** Marginal Forms are Licit

# **Evidence:** Borrowings & Repairs

- Illicit forms are repaired in borrowings:
  - Greek **/pneu̯mɔn/** → English **/njumoniə/**
  - German **/pfɪtsɐ/** → English **/faɪzɹ/**
- Spanish & Japanese: ***/#sC/**

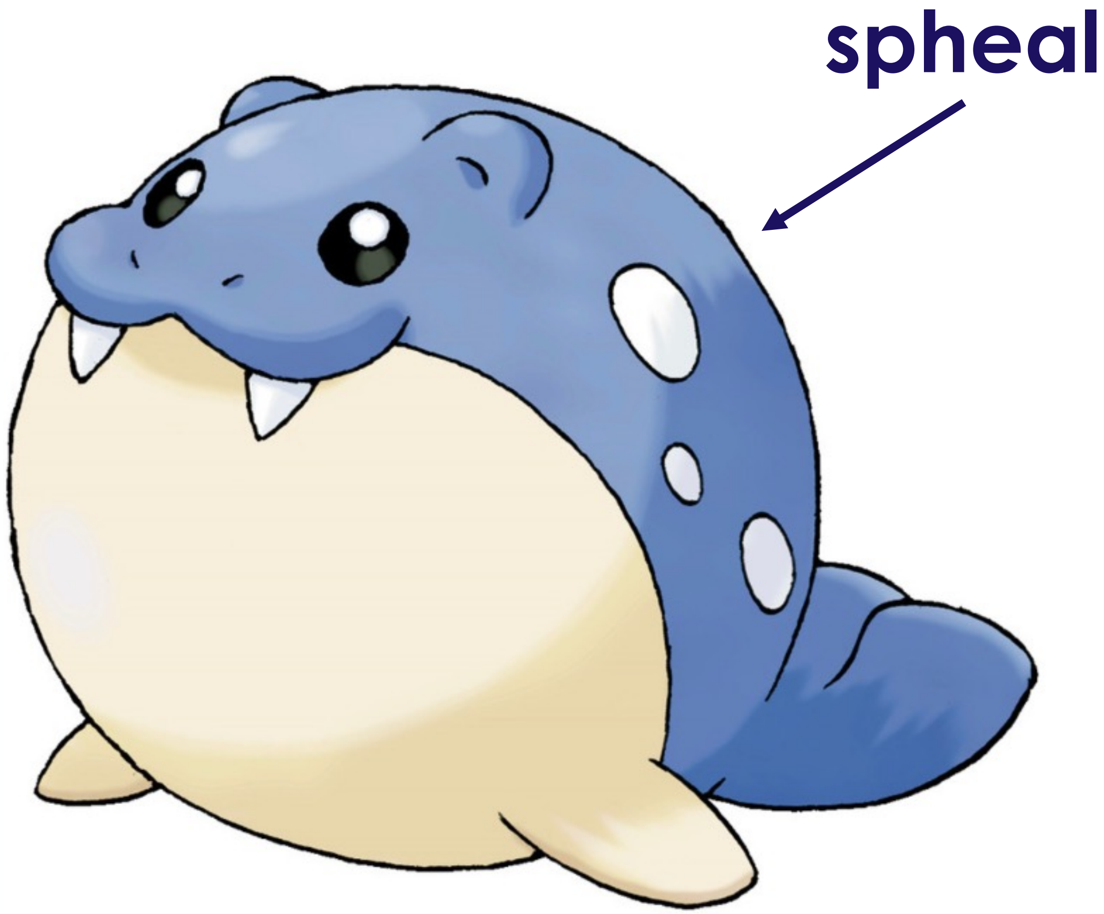
| | Spanish | Japanese |
|---|---|---|
| Italian: /spagetti/ | /espageti/ | /sɯpagetti/ |
| Greek: /sfiŋks/ | /esfinxe/ | /sɯɸinkɯsɯ/ |
| Greek: /sfaira/ | /esfera/ | (sɯɸia) |

# **Evidence:** Borrowings & Repairs

- Illicit forms are repaired in borrowings:
  - Greek **/pneu̯mɔn/** → English **/njumoniə/**
  - German **/pfɪtsɐ/** → English **/faɪzɹ/**
- Spanish & Japanese: ***/#sC/**

| | Spanish | Japanese | English |
|---|---|---|---|
| **Italian: /spagetti/** | /espageti/ | /sɯpagetti/ | /spəgɛti/ |
| **Greek: /sfiŋks/** | /esfinxe/ | /sɯɸinkɯsɯ/ | /sfinks/ |
| **Greek: /sfaira/** | /esfeɾa/ | (sɯɸia) | /sfɪɹ/ |

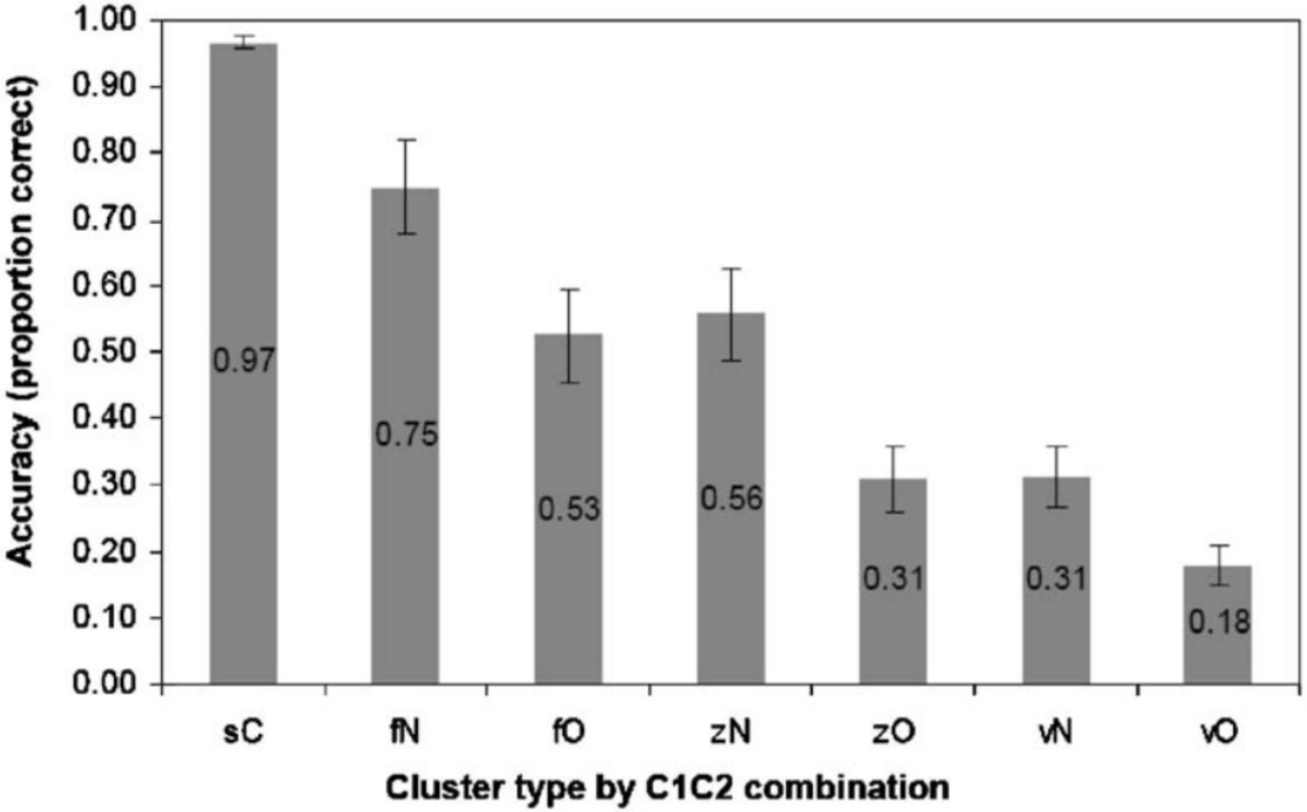
# **Evidence:** New Words



# **Evidence:** New Words

# **Evidence:** Production & Perception

- Speakers **have trouble producing illicit sequences**
- But they **don't have trouble producing /#sf/!**
    - 97% accuracy /#sC/ sequences where **C ∈ {f, p, t, k, m, n}**




(Davidson 2006)

# **Evidence:** Underspecification in Acquisition

# Underspecification in Early Phonology

- Early discrimination:
  - English–learning children at 1;2 (Yeung & Werker 2009):
    - ***Cannot discriminate*** **/bɪ/** and **/dɪ/** when ***lexical contrast*** implicated
    - ***Can discriminate*** **[b]** and **[d]** when ***phonetic contrast*** implicated
  - English-learning children (Gierut 1996):
    - Producing **/θ/** ***can discriminate*** **/s/** and **/θ/**
    - Not producing **/θ/** ***can not discriminate*** **/s/** and **/θ/**
    - Both ***can not discriminate*** **/f/** and **/ɸ/**

# Underspecification in Early Phonology

- “Mispronunciation” studies (Hallé & Boysson-Bardies 1966)
  - French-learning 11-month-olds:
    - Do not prefer **known words to alternants** with:
      - Different ***voicing*** (e,g. **[gato]** vs. **[kato]**)
      - Different ***manner*** (e.g. **[banan]** vs. **[vanan]** vs. **[balan]**)
- Suggests children have **knowledge of segments** but this knowledge is initially **featurally-underspecified**

# Previous Work

# Previous Work

## Maximum Entropy

(Hayes & Wilson 2008)

- **Negative grammar of markedness constraints**
- Weighted markedness constraints ⇒ **probability of output**
- Goal of learning = determine **constraints and ranking that maximize probability** of observed forms
- **Guaranteed to find global maximum**

## String Extension Learning

(Heinz 2010)

- **Positive grammar of *k*-factors**
- Accumulate ***k*-factors from the input**
  - ***k*-factors** = substrings of length *k*
- Add *k*-factors to the grammar as they are seen
- A string is licit if **all of its *k*-factors are licit**
- **Learnable in the Limit from Positive Data**

# **Previous Work:** Handling Marginal Forms

## Maximum Entropy

- Weight e.g. ***/#sf/** less than ***/#bn/**
  - Violating ***/#sf/** is ***less bad***
- Hayes & Wilson remove **"exotic onsets"** from train
  - Performance hit when they're included

## String Extension Learning

- If **all *k*-factors seen in input**, then string is licit
- **No distinction** between marginal and fully-licit inflected forms
- No **underspecification** in classic SEL
  - But see Chandlee et al (2019)

# Proposal

# **Proposal:** Measuring Generalizability

- **The Tolerance-Sufficiency Principle** (TSP, Yang 2016)
  - Threshold for generalization ***based on computational efficiency***
  - Given a rule *R* applicable to *N* types and seen applying to *M* of those types, ***generalize the rule iff:***

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

# **Proposal:** Measuring Generalizability

- Given a sequence of underspecified feature sets, **do a sufficient number of sequences fitting it occur?**
  - Let $N = \prod n_i$ where $n_i$ = **# segments that fit features at position $i$**
  - Let $M$ be the number of **distinct sequences observed that fit the entire feature set**
  - Check if $M - N \leq \frac{N}{\ln N}$

# **Proposal:** Recursive Learning

- Test feature-set sequence against the TSP
  - If passes, **productive sequence learnt!**
  - If not, **posit more specific sequence** by:
    - Finding **position $i$ with greatest difference between # observed segments and $n_i$**
    - Adding the most frequent feature at this position to the representation
    - **Subdivide & recurse**
- Recursion ends either when:
  - A **productive licit sequence** is learnt
  - **No more features** available to subdivide ⇒ **memorize**

# **Proposal:** Recursive Learning

- Example: **English complex onsets**
  - $N$(**[+sibiliant] [-son, -cont]**) = |**{z, s}** x **{p, t, k, b, d, g}**| = **12**
  - $M$ = number of distinct sequences that fit **[+sibiliant] [-son, -cont]**
    - Seen **{sp, st, sk}** ⇒ $\boldsymbol{M = 3}$
  - $\boldsymbol{N - M = 12 - 3 = 9 > \theta_{12} \approx 4.8}$ ✗
  - **Subdivide:** find position with **greatest difference** between number of **observed** & number of **possible** segments
    - **First position:** 2 possible, 1 observed ⇒ **1 difference**
    - **Second position:** 6 possible, 3 observed ⇒ **3 difference**
  - Add most frequent feature occurring at this position: $\boldsymbol{\pm\text{voice}}$
  - Recurse: **[+sibiliant] [-son, -cont, -voi]** vs. **[+sibiliant] [-son, -cont, +voi]**

# **Experiment:** English Complex Onsets

- We apply the model to a sample of **child-directed speech**
  - 5584 forms from the ***CHILDES Brown corpus***
  - Transcribed using the ***CMU Pronouncing Dictionary***
  - ***Distinctive features*** encoded for ARPABET based on those in Hayes & Wilson (2008)
    - Features can be **positive, negative, or unspecified**

# **Results:** English Complex Onsets

| Complex Onset | Example |
|---|---|
| {+cont, +cons, +strident, +coronal, -son, +anterior, -approx, -voi, -V}<br>{+son, +cons, -approx, +labial, +nasal, -V}<br>{+V, -cons, +approx} | small, smell |
| {+cont, +cons, +strident, +coronal, -son, +anterior, -approx, -voi, -V}<br>{+cons, -son, -cont, -approx, -voi, -V}<br>{+approx} | skip, spatter, spray |
| {+cons, -son, +voi, -cont, -approx, -V}<br>{+son, +cons, +anterior, +coronal, +approx, -strident, -V}<br>{+V, -cons, +approx} | break, drab, black |
| {+cont, +cons, +strident, +coronal, -son, +anterior, -approx, -voi, -V}<br>{+cons, +coronal, +anterior, -son, -cont, -approx, -strident, -voi, -V}<br>{+son, +cons, +anterior, +coronal, +approx, -strident, -V} | stress, strike |
| {+cont, +cons, +strident, +coronal, -son, +anterior, -approx, -voi, -V}<br>{+cons, +coronal, +anterior, -son, -cont, -approx, -strident, -voi, -V}<br>{+V, -cons, +approx} | still, stem |
| {+cons, -son, -approx, -voi, -V}<br>{+son, +cons, +anterior, +coronal, -strident, -V}<br>{+V, -cons, +approx} | plank, throw, floor |

# **Results:** Productive English Complex Onsets

- Onsets that **don't start with /s/:**
  - ***Voiced stops and voiceless stops and fricatives*** can precede liquids
    - e.g. **/#bl/, /#tr/, /#sl/**
  - ***Voiced fricatives*** cannot
    - e.g. ***/#zl/**
- Onsets that **do start with /s/:**
  - Second position can be a ***voiceless stop*** & third can be ***vowel or liquid***
    - e.g. **/#str/, /#spl/**
  - Second position can be a nasal
    - Only sees **/#sm/** so does not generalize to **/#sn/** or **/#sŋ/**

# Conclusion & Future Directions

- Model of **phonotactic acquisition** that uses **recursive search & the Tolerance-Sufficiency Principle**
  - Learns ***positive grammar*** from ***positive data***
  - ***Increasing specification*** of licit sequences
  - ***Fully-licit*** vs. ***marginal*** vs. ***illicit*** forms
- Future directions:
  - Apply to **more languages**
  - Incorporate **syllable structure**
  - **Long-distance** dependencies

# Thank you!!

I am grateful to Jeff Heinz, Jordan Kodner, and Charles Yang for their mentorship; Kyle Gorman, Scott Nelson, and Huteng Dai for helpful discussion; and Logan Swanson and Salam Khalifa for support throughout this project.

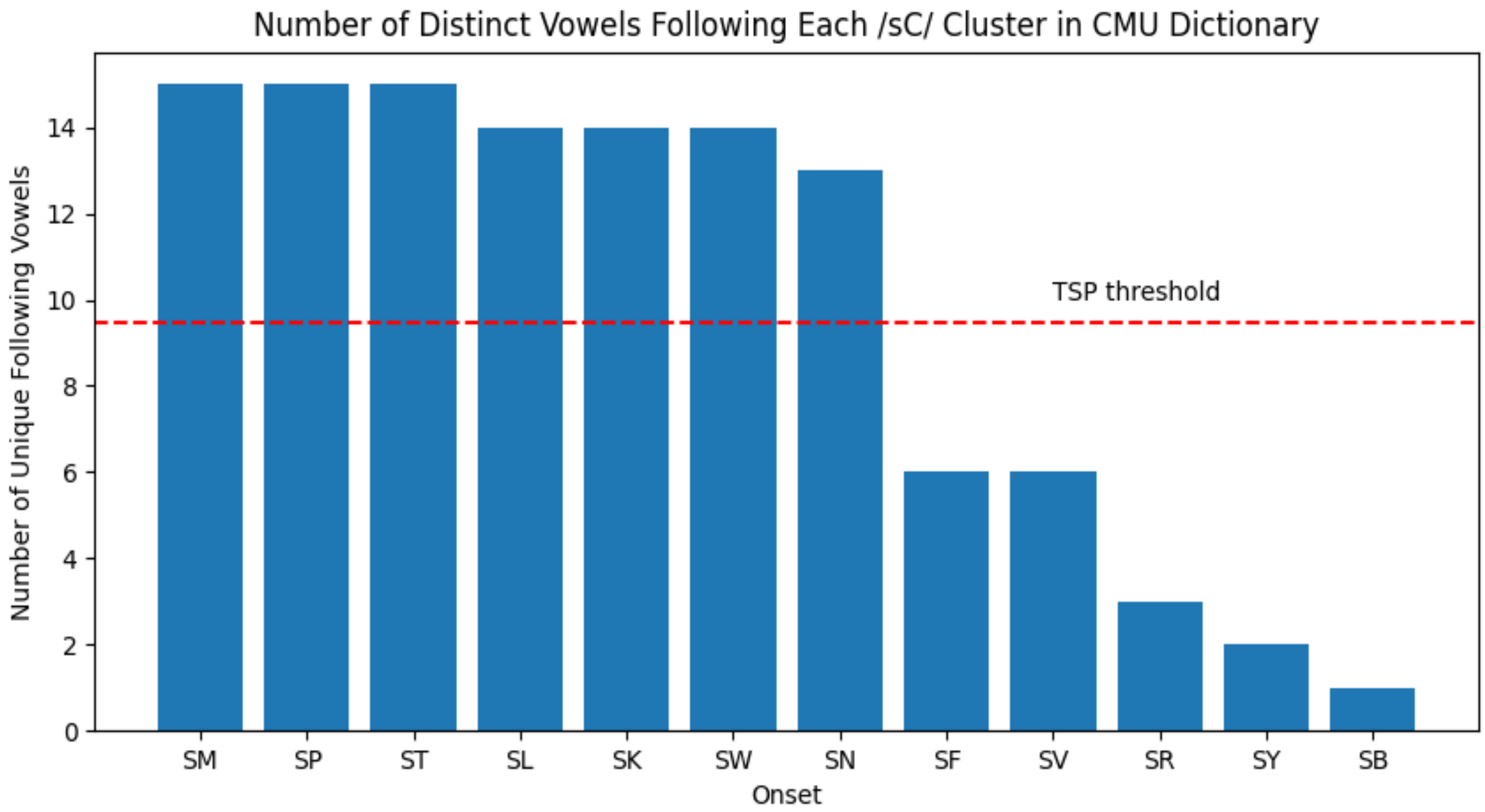
I am grateful for funding by the Institute for Advanced Computational Science and National Science Foundation.

NSF

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

DEPARTMENT OF LINGUISTICS STONY BROOK

# **Proposal:** Degree of Specification



# **Previous Work:** Gradient Models

- **MaxEnt** (Hayes & Wilson 2008): ***well-formedness = probability***
  - **Weighted markedness constraints** ⇒ probability of output
  - Goal of learning = determine **constraints and ranking that maximize probability** of observed forms
    - ***Guaranteed to find global maximum***

# **Previous Work:** Categorical Models

- **String-Extension Learning** (SEL, Heinz 2010): accumulate ***k*-factors from the input** to form a positive grammar
  - Initial grammar = ∅
  - For some input $t[i]$, the output of the learner $\phi$ is:
    $$\boldsymbol{\phi(t[i]) = \phi(t[i-1]) \cup \{x \in \Sigma^k : \exists\, u, v \in \Sigma^*, w = uxv\}}$$
  - The language of the resulting grammar is given by:
    $$\boldsymbol{L(G) = \{w \in \Sigma^* : fac_k(w) \subseteq G\}}$$
  - Strictly Local languages are ***Learnable in the Limit from Positive Data***