Marginal Sequences as a Window into Phonotactic Learning

Sarah Brogden Payne

sarah.payne@stonybrook.edu | paynesa.github.io

Stony Brook University

NSF

iACS INSTITUTE FOR ADVANCED COMPUTATIONAL SCIENCE

MIT LingLunch

April 18, 2024

Attestation vs. Licitness

- **Subcomponent attestation** is closely related to **licitness**
- **Subcomponents: syllable-based** or **linear *k*-factors**
  - Toy example: given **[can]** and **[dab]**, is **[cab]** acceptable?

Syllable sub-components

- **[can]** ⇒ **[c]** is a licit **onset**
- **[dab]** ⇒ **[ab]** is a licit **rime**
- **[cab]** = **[c]** + **[ab]**

✅

Linear *k*-factors

- 2-factors of **[can]** = **{#c, ca, an, n#}**
- 2-factors of **[dab]** = **{#d, da, ab, b#}**
- 2-factors of **[cab]** = **{#c, ca, ab, b#}**

✅

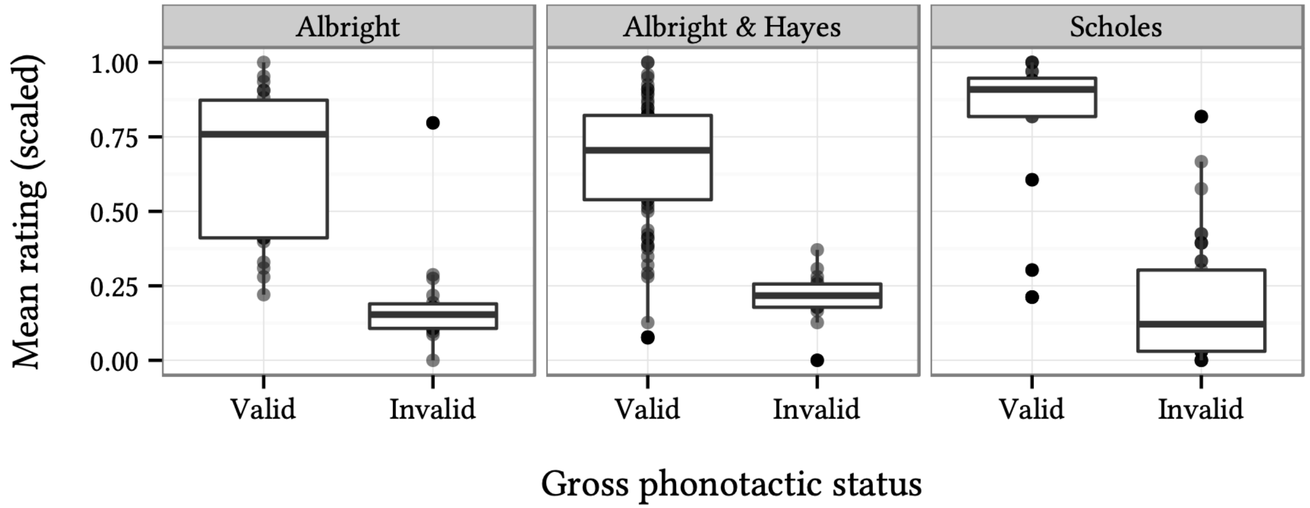
Attestation vs. Licitness

- **Subcomponent attestation** is closely related to **licitness**

Gorman (2013)

Syllable-based attestation vs. English nonce-word judgments

***r* = 0.73** ***r* = 0.6** ***r* = 0.8**



Kostyszyn & Heinz (2022)

2-factor attestation for Polish word-initial complex onsets

Pearson's ***r* = 0.73**

What's the causal relationship between attestation and licitness?

Attestation vs. Licitness: Traditional View

	Attested	Unattested
Licit	spot	blick
Illicit	sphere	bnick

Subcomponents are attested
e.g. *blip, sick*

Traditional view: licitness ⇒ subcomponents are attested

Attestation vs. Licitness: Traditional View

	Attested	Unattested
Licit	spot	blick
Illicit	sphere	bnick

Some subcomponent is unattested
*[#bn] or *bn-onset

Traditional view: illicit ⇒ unattested subcomponent

Attestation vs. Licitness: Traditional View

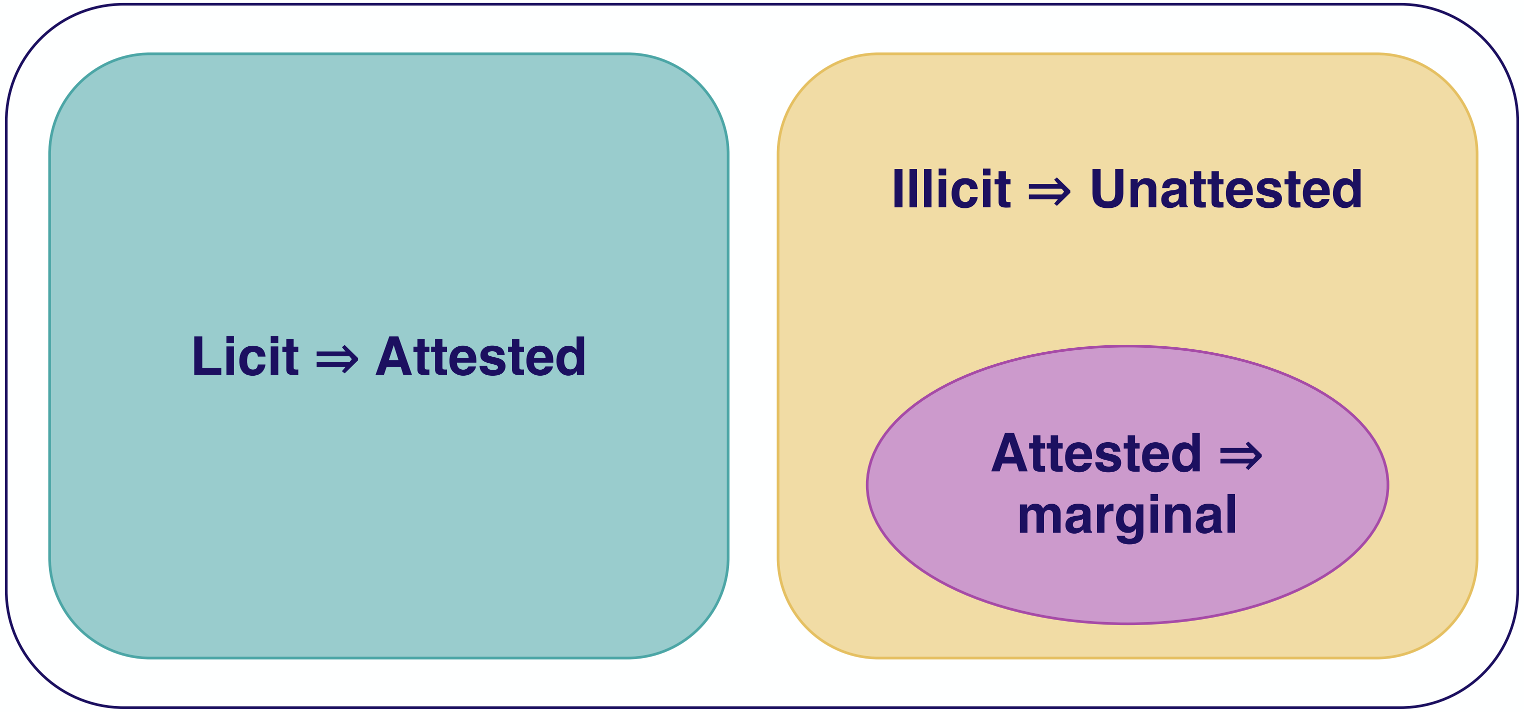
	Attested	Unattested
Licit	spot	blick
Illicit	sphere	bnick

⇛ **All subcomponents are attested but rated poorly**

Traditional view: marginal = **exceptional subclass of illicit**

→ Illicit **but contain no unattested subcomponent**

Attestation vs. Licitness: Traditional View



Attestation vs. Licitness Revisited

	Attested	Unattested
Licit	spot	blick
Illicit	sphere	bnick

Subcomponents are attested
e.g. *blip, sick*

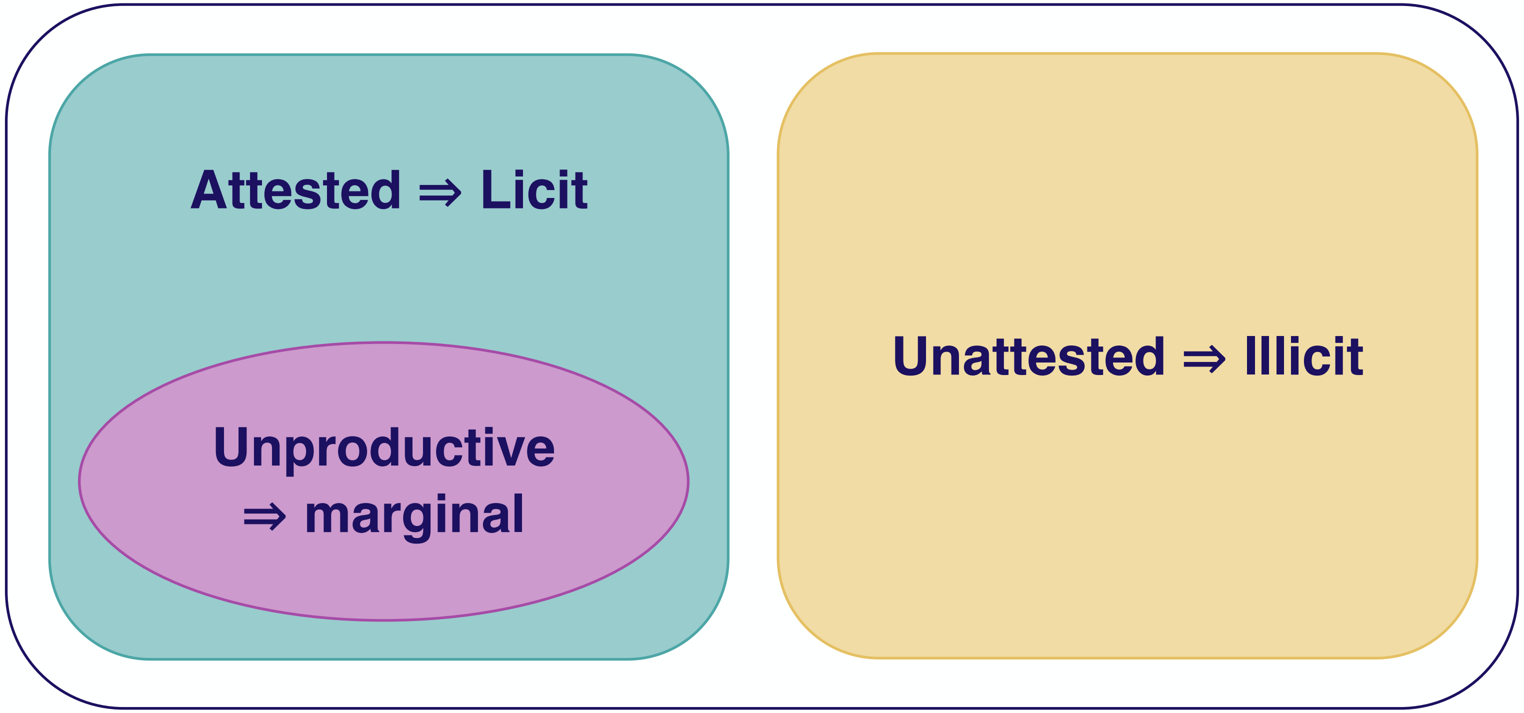
Subcomponents attested ⇒ **licit**

Unattested subcomponent ⇒ **illicit**

Marginal = exceptional subclass of attested

Subcomponents attested but **not licit**

Attestation vs. Licitness: Proposal



The phonotactic grammar is **positive**, **syllable-based**, and **categorical**, with forms being either **licit**, **marginal**, or **illicit**

Outline

- Re-thinking the phonotactic grammar
  - Motivating observations
  - What's (not) in the phonotactic grammar
  - Phonotactic knowledge is non-linear
  - A positive phonotactic grammar
  - Phonotactic representations may be categorical
- Working Proposal
  - Proposal: Sequence-Wise Generalization Learner
  - Evaluation: English complex onsets
- Future work

Outline

- Re-thinking the phonotactic grammar
  - **Motivating observations**
  - What’s (not) in the phonotactic grammar
  - Phonotactic knowledge is non-linear
  - A positive phonotactic grammar
  - Phonotactic representations may be categorical
- Working Proposal
  - Proposal: Sequence-Wise Generalization Learner
  - Evaluation: English complex onsets
- Future work

Difference in Predictions

Traditional View

- Licit ⇒ Attested
- Illicit ⇒ Unattested
  - Attested ⇒ marginal

Marginal sequences are an **exceptional subclass of illicit** ones, so we expect **marginal sequences to pattern like illicit ones**

Proposal

- Attested ⇒ Licit
  - Unproductive ⇒ marginal
- Unattested ⇒ Illicit

Marginal sequences are an **exceptional subclass of licit** ones, so we expect **marginal sequences to pattern like licit ones**

Evidence: Borrowings and 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/	/esfeɾa/	(sɯɸia)

Evidence: Borrowings and Repairs

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

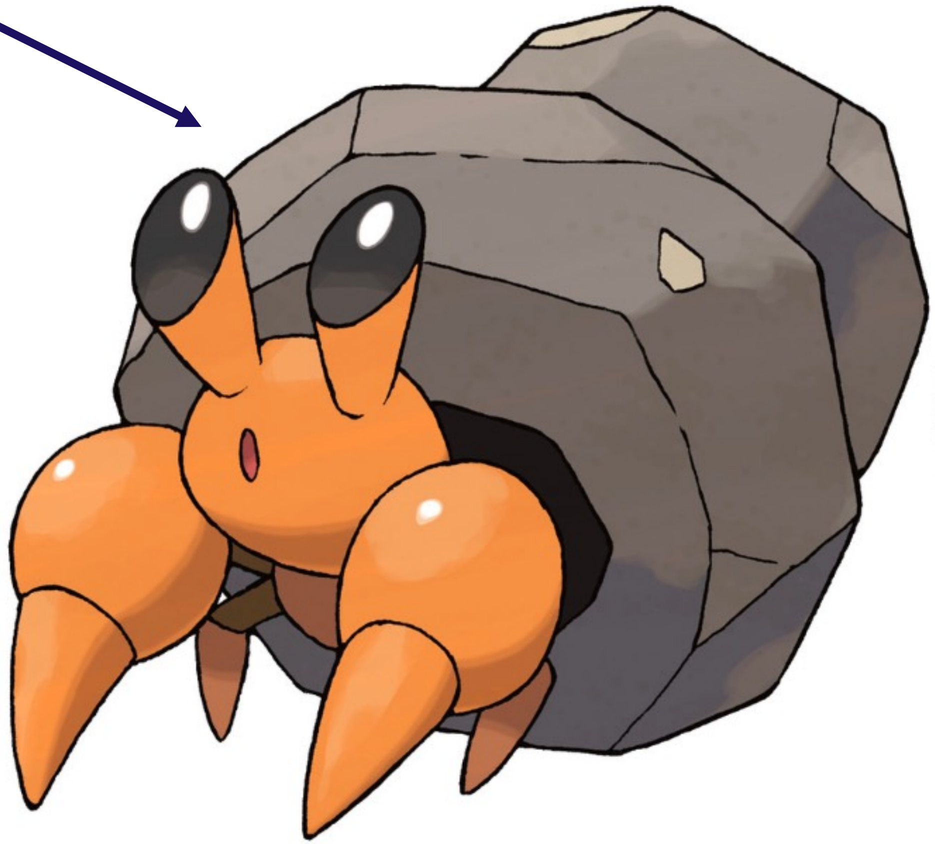
	Spanish	Japanese	English
Italian: /spagetti/	/espageti/	/sɯpagetti/	/spəgɛti/
Greek: /sfiŋks/	/esfinxe/	/sɯɸinkɯsɯ/	/sfinks/
Greek: /sfaira/	/esfeɾa/	(sɯɸia)	/sfɪɹ/

Evidence: New Words

spheal

Evidence: New Words

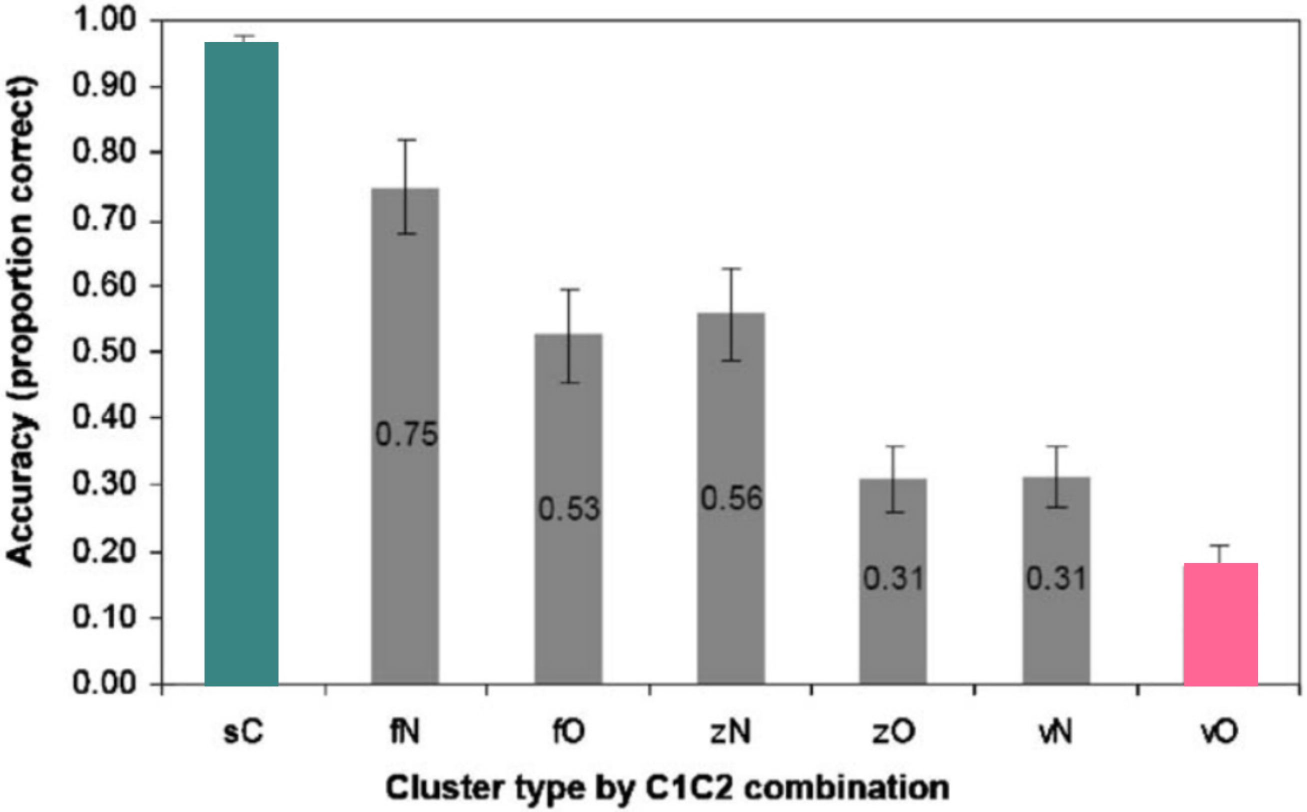
dwebble



spheal

Evidence: Production and 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)

Difference in Predictions

Traditional View

Licit ⇒ Attested

Illicit ⇒ Unattested

Attested ⇒ marginal

Marginal sequences are an **exceptional subclass of illicit** ones, so we expect **marginal sequences to pattern like illicit ones**

Proposal

Attested ⇒ Licit

Unproductive ⇒ marginal

Unattested ⇒ Illicit

Marginal sequences are an **exceptional subclass of licit** ones, so we expect **marginal sequences to pattern like licit ones**

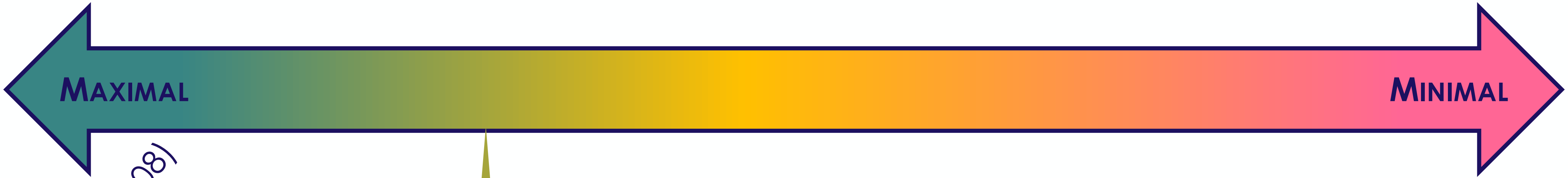
Outline

- Re-thinking the phonotactic grammar
  - Motivating observations
  - **What's (not) in the phonotactic grammar**
  - Phonotactic knowledge is non-linear
  - A positive phonotactic grammar
  - Phonotactic representations may be categorical
- Working Proposal
  - Proposal: Sequence-Wise Generalization Learner
  - Evaluation: English complex onsets
- Future work

What’s in the Phonotactic Grammar?

Any surface-true generalization that holds based on **statistical inference over the lexicon**, not necessarily resulting from phonological alternations or restrictions on the prosodic system
Hayes & Wilson (2008)

What's in the Phonotactic Grammar?



Hayes & Wilson (2008)

Any surface-true generalization that holds based on **attestation in the lexicon**, not necessarily resulting from phonological alternations or restrictions on the prosodic system, but **restricted to certain computational classes**
Heinz (2010), Chandlee et al. (2019), Rawski (2021)

MaxEnt and SEL: A Closer Look

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** = subcomponents 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**

MaxEnt and SEL: Handling Marginal Forms

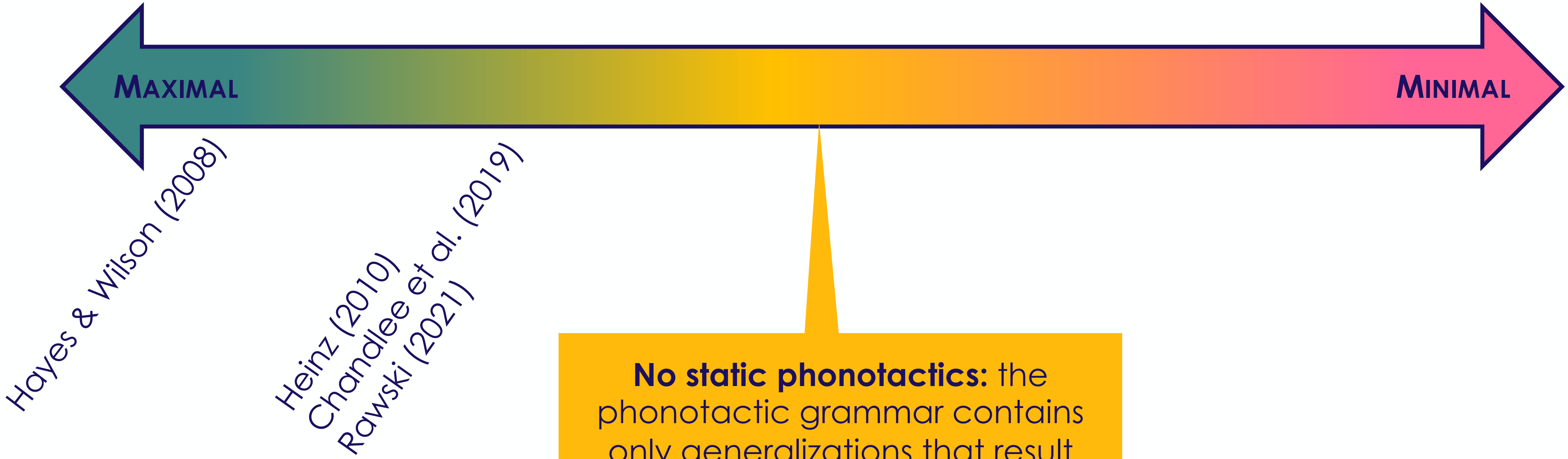
Maximum Entropy

- 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

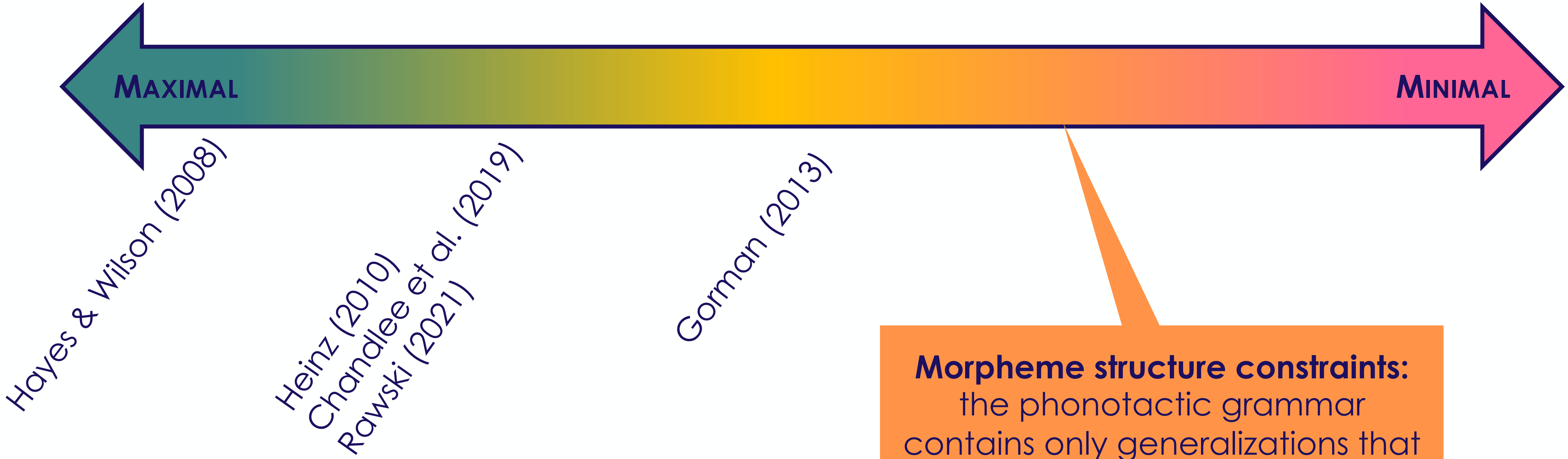
- If **all *k*-factors seen in input**, then string is licit
- **No distinction** between marginal and licit forms

What's in the Phonotactic Grammar?



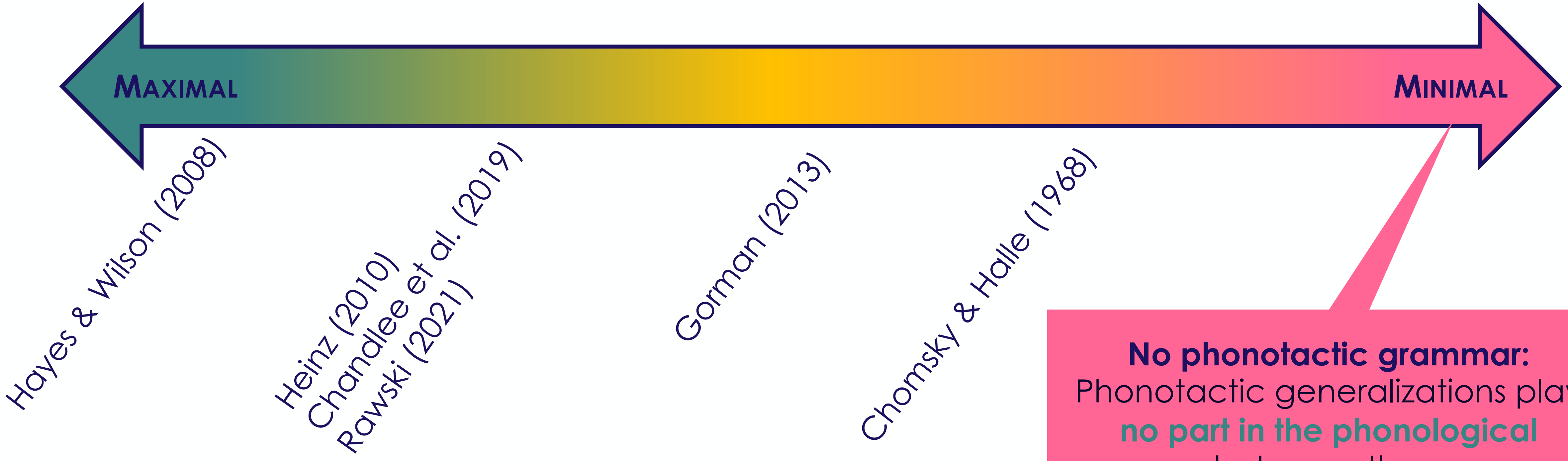
No static phonotactics: the phonotactic grammar contains only generalizations that result from **phonological alternations** or **restrictions on the prosodic inventory**
Gorman (2013)

What's in the Phonotactic Grammar?



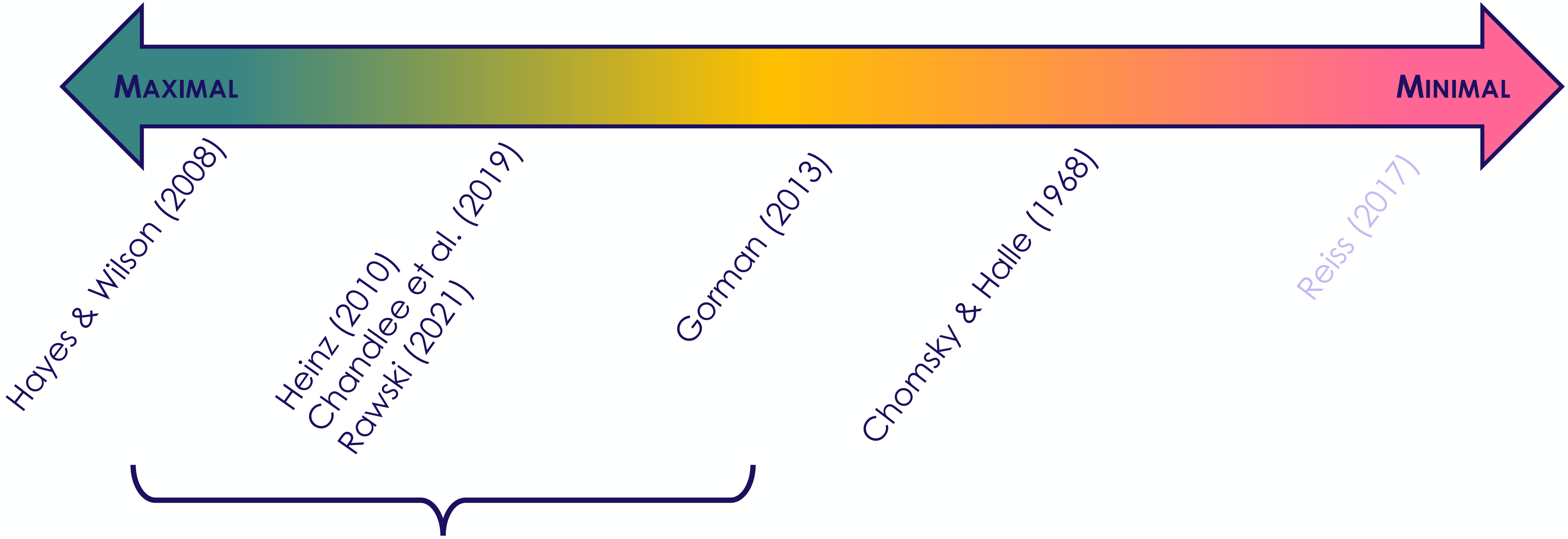
Morpheme structure constraints: the phonotactic grammar contains only generalizations that result from **restrictions on the prosodic inventory of underlying representations**
Chomsky & Halle (1968)

What's in the Phonotactic Grammar?



No phonotactic grammar: Phonotactic generalizations play **no part in the phonological grammar** but are rather **emergent, metalinguistic knowledge**
Reiss (2017)

What's in the Phonotactic Grammar?



Are there cases of **purely static restrictions** that are **synchronically active** in speakers' grammars?

Inactive Static Restrictions: Turkish

- Backness Harmony
  - **61%** of roots conform
- Roundness Harmony
  - Applies to high vowels
  - **83%** of roots conform
- Labial Attraction
  - High back vowels tend to be round after *a*-labial consonant sequences
  - **Not reflected in alternations**
  - **69%** of roots conform

Nom.Sg	Nom.Pl	Dat.Sg
pelür	pelür**ler**	pelür**ü**
boğaz	boğaz**lar**	boğaz**ɪ**
ip	ip**ler**	ip**i**

Gorman (2013)

Inactive Static Restrictions: Turkish

- **Zimmer (1969)** paired word-likeness task: which is better?
- Goodman-Kruskall ɣ measured for each restriction:
  - **Backness Harmony**: ɣ = **0.694** ✅
  - **Roundess Harmony**: ɣ = **0.68** ✅
  - **Labial Attraction**: ɣ = -0.043 ❌

  } **Reflected in alternations** (Backness Harmony, Roundess Harmony)

  } **Purely static** (Labial Attraction)
- Suggests a more **minimal** view of the phonotactic grammar
  - **Not all surface-true generalizations** will be grammaticalized
  - Current work: focus on **restrictions on prosodic inventory**

Gorman (2013)

Outline

- Re-thinking the phonotactic grammar
  - Motivating observations
  - What's (not) in the phonotactic grammar
  - **Phonotactic knowledge is non-linear**
  - A positive phonotactic grammar
  - Phonotactic representations may be categorical
- Working Proposal
  - Proposal: Sequence-Wise Generalization Learner
  - Evaluation: English complex onsets
- Future work

Subcomponents Revisited

Syllable sub-components

- [can] ⇒ [c] is a licit **onset**
- [dab] ⇒ [ab] is a licit **rime**
- [cab] = [c] + [ab]

✅

Gorman (2013)

Linear *k*-factors

- 2-factors of [can] = {#c, ca, an, n#}
- 2-factors of [dab] = {#d, da, ab, b#}
- 2-factors of **[cab]** = {#c, ca, ab, b#}

✅

Hayes & Wilson (2008)
Heinz (2010)
Chandlee et al. (2019)
Rawski (2021)

Equivalence?

- Linear representations can be augmented with syllable boundaries
  - **2-factors** of **[hæ.pi] = {#h, hæ, æ., .p, pi, i#}**
- Inherent generalization power is still different:

Linear + Syllable Boundaries

- Will need ***k* > 3** to capture clusters
- **[d.n]** and **[b.m]** in the observed *k*-factors
- **[d.m]** and **[b.n]** will **not** be accepted

Syllable-Based Representations

- Observe **[d]** and **[b]** as licit codas
- Observe **[m]** and **[n]** as licit onsets
- **[d.m]** and **[b.n]** will be accepted

Which do humans do?

Evidence for Non-Linear Representations

- **Bernard & Onishi (2023):** infants & children spontaneously represent phonotactic restrictions **over syllables**
- **Kabak & Idsardi (2007):** adult Korean speaker's illusory vowel perception is governed by syllable-position restrictions
- Extremely **early sensitivity** to syllables

Evidence for Non-Linear Representations

- **Bernard & Onishi (2023):** infants & children spontaneously represent phonotactic restrictions **over syllables**
  - **Children** (5;0) and **infants** (0;11)
    (55.5–65.8 months) **(10.6–12.1 months)**
  - Distinguish sensitivity to **linear co-occurrence** vs. **syllable position**

Spontaneous Representation of Restrictions

Restricted consonants (p,z,d,f) vs. **unrestricted** (b,k,t,v)

[f] can only occur as a coda &
[p] can only occur as an onset
[f] can only occur before [p]

`cvC.Cvc` words displaying both **syllable-position** and **consonant co-occurrence** restrictions word-medially

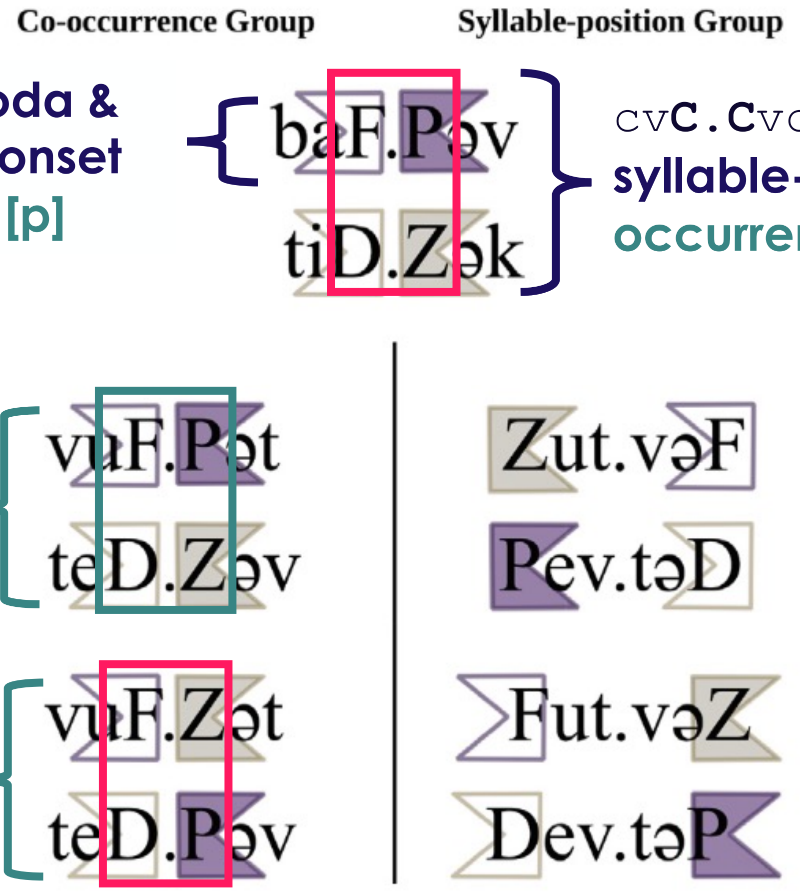
Restricted consonants in **same syllable positions** as training

Test: Same

The `c.c` sequence **is attested** in train

Different

The `c.c` sequence **is not attested** in train

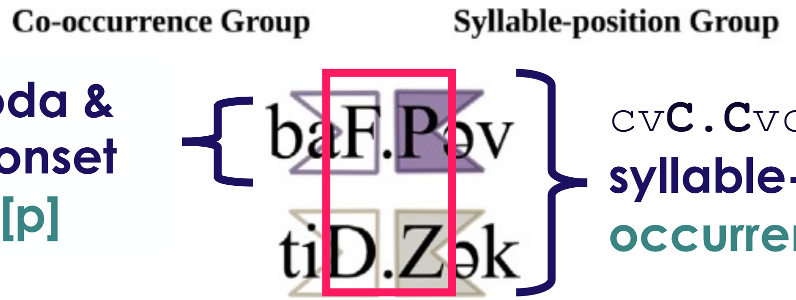


Bernard & Onishi (2023)

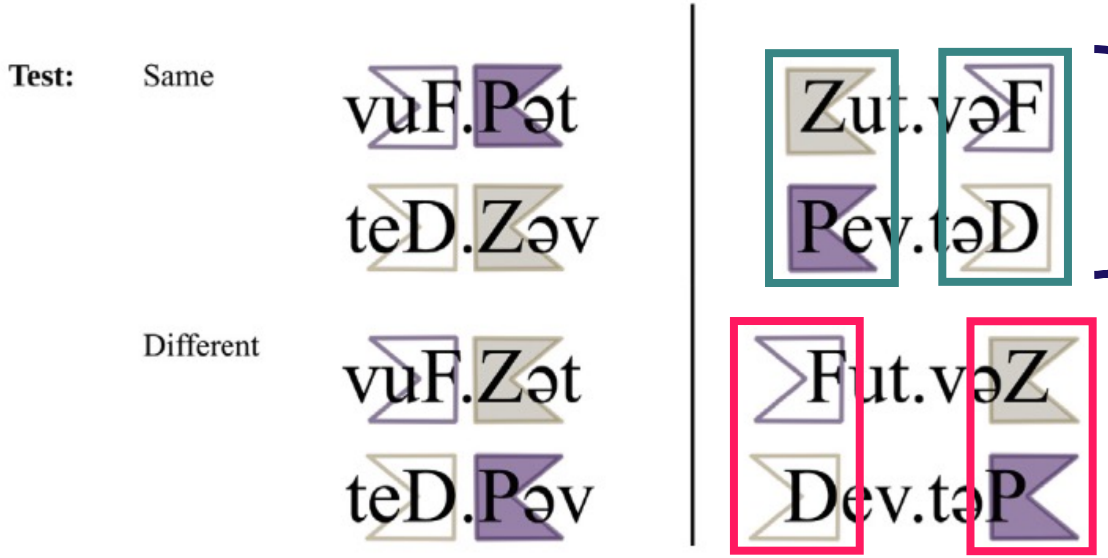
Spontaneous Representation of Restrictions

Restricted consonants (p,z,d,f) vs. **unrestricted** (b,k,t,v)

[f] can only occur as a coda &
[p] can only occur as an onset
[f] can only occur before [p]



cvC.Cvc words displaying both **syllable-position** and **consonant co-occurrence** restrictions word-medially

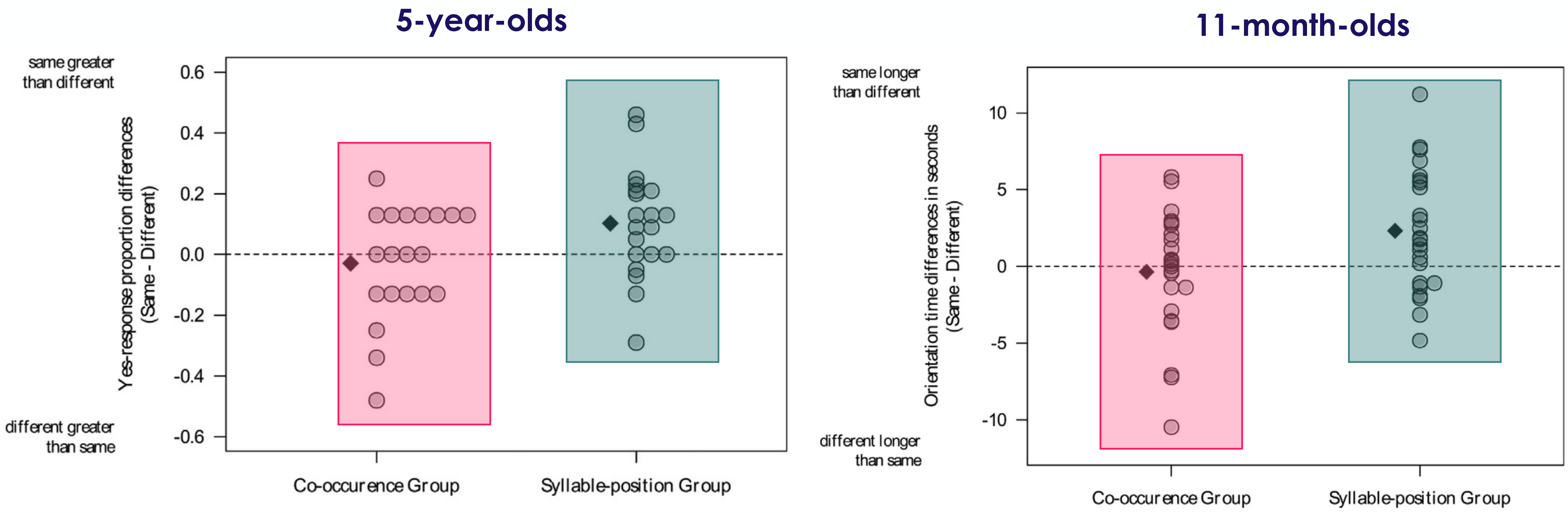


The restricted consonants are in the **same syllable position** as in train

The restricted consonants are in **different syllable positions** as in train

Bernard & Onishi (2023)

Spontaneous Representation of Restrictions



Children & infants exploit syllable structure in phonotactic learning even when other information is available

Bernard & Onishi (2023)

Evidence for Non-Linear Representations

- **Bernard & Onishi (2023):** infants & children spontaneously represent phonotactic restrictions **over syllables**
- **Kabak & Idsardi (2007):** adult Korean speaker's illusory vowel perception is governed by syllable-position restrictions
  - $VC_1.C_2V$ sequences are generally ok in Korean but **some unattested**
  - **Contact:** C_1 is a licit coda and C_2 is a licit onset, but $C_1.C_2$ **unattested**
    - ***[k.m]** because **[k]** undergoes nasalization to **[ŋ.m]**
  - **Syllable-position:** C_1 is **unattested** as coda or C_2 **unattested** as onset
    - ***[c.] *[ɾ.]** for codas and ***[.l] *[.ŋ]** for onsets
  - Korean-speaking adults can discriminate $VC_1.C_2V$ from $V.C_1V.C_2V$ in the **contact** case but struggle in the **syllable-position** case
    - **Syllable-based** account predicts this **asymmetry**

Evidence for Non-Linear Representations

- **Bernard & Onishi (2023):** infants & children spontaneously represent phonotactic restrictions **over syllables**
- **Kabak & Idsardi (2007):** adult Korean speaker's illusory vowel perception is governed by syllable-position restrictions
- Extremely **early sensitivity** to syllables
  - **Bijeljac-Babic et al (1993): 4-day-old infants** discriminate words based on number of syllables but not number of phonemes
  - **Bertocini & Mehler (1981):** infants can discriminate syllable-like stimuli better than non-syllable stimuli **before 0;2**
  - **Peters (1983):** word **segmentation errors** align with syllable boundaries

Outline

- Re-thinking the phonotactic grammar
  - Motivating observations
  - What's (not) in the phonotactic grammar
  - Phonotactic knowledge is non-linear
  - **A positive phonotactic grammar**
  - Phonotactic representations may be categorical
- Working Proposal
  - Proposal: Sequence-Wise Generalization Learner
  - Evaluation: English complex onsets
- Future work

Positive & Negative Grammars: Equivalence

- **Dominant View:** phonotactic grammar made up of **negative constraints** (e.g., *[#bn])
  (Prince & Smolensky 1993, Hayes & Wilson 2008, Dai 2024, i.a.)
- Why not store sub-components that **are allowed?**
- **Model Theory** tells us:
  - **Over segments:** straightforward **conversion** between grammar types
  - **Over feature bundles:** the same **algorithm** can be used to learn both
- From a computational perspective, **no a-priori reason to favor a negative grammar**

Positive & Negative Grammars: Equivalence

- **Model Theory** tells us:
  - **Over segments:** straightforward **conversion** between grammar types
    - **Toy example:**
      2-factor grammar, $\Sigma = \{\texttt{V}, \texttt{C}\}$
    - Positive grammar:
      $G^+ = \{\texttt{VC}, \texttt{CV}\}$
    - Negative grammar:
      $G^- = \Sigma^2 \setminus G^+ = \{\texttt{VV}, \texttt{CC}, \texttt{CV}, \texttt{VC}\} \setminus \{\texttt{VC}, \texttt{CV}\} = \{\texttt{VV}, \texttt{CC}\}$
    - Banning **VV** and **CC** or only allowing **VC** and **CV** ⇒ same language!

Heinz (2010)

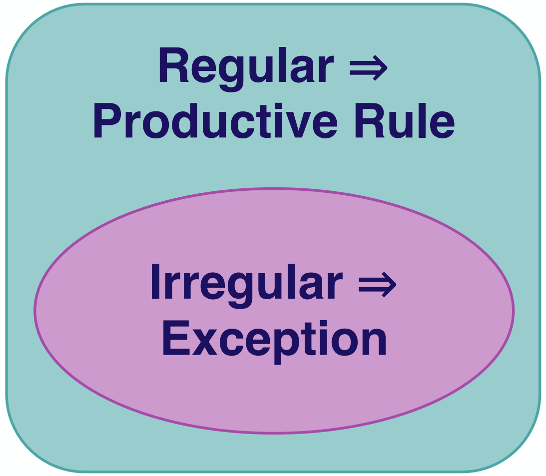
Positive & Negative Grammars: Equivalence

- **Model Theory** tells us:
  - **Over segments:** straightforward **conversion** between grammar types
  - **Over feature bundles:** the same **algorithm** can be used to learn both
    - **Chandlee et al (2019) & Rawski (2021):** algorithm to learn **only negative grammars** over sequences of feature bundles
      - ***[+Nas][-Son, -Voi]** instead of ***nt, *mp, *ŋk**, etc. separately
    - **Prohibitively costly** to convert between negative & positive grammars of feature bundles
    - **Payne (2024):** if we **fix *k*** (the size of the elements in the grammar), we can adapt this algorithm to **learn positive and negative grammars with the same guarantees**

Computationally, **no advantage to a negative grammar**

Is Phonotactic Learning Really so Different?

- Syntax: **positive grammar**
  Chomsky (1957, 1992); Liang et al. (2022); Li & Schuler (2023) i.a.
- Morphology: **positive grammar**
  Pinker (1998); Yang (2016); Belth et al. (2021) i.a.
- Phonology:
  - Rule-based view: **positive grammar**
    Chomsky & Halle (1968); Belth (2023, 2024), i.a.
  - Optimality Theory: **negative grammar**
    Prince & Smolensky (1993); McCarthy (2007, 2008), i.a.
- Phonotactics: **is it different?**



Is Phonotactic Learning Really so Different?

- Syntax: **positive grammar**
  Chomsky (1957, 1992); Liang et al. (2022); Li & Schuler (2023) i.a.
- Morphology: **positive grammar**
  Pinker (1998); Yang (2016); Belth et al. (2021) i.a.
- Phonology:
  - Rule-based view: **positive grammar**
    Chomsky & Halle (1968); Belth (2023, 2024), i.a.
  - Optimality Theory: **negative grammar**
    Prince & Smolensky (1993); McCarthy (2007, 2008), i.a.
- Phonotactics: **is it different?**

Expand the learning approaches from these subfields to phonotactics

Outline

- Re-thinking the phonotactic grammar
  - Motivating observations
  - What's (not) in the phonotactic grammar
  - Phonotactic knowledge is non-linear
  - A positive phonotactic grammar
  - **Phonotactic representations may be categorical**
- Working Proposal
  - Proposal: Sequence-Wise Generalization Learner
  - Evaluation: English complex onsets
- Future work

Gradient Representations?

- Experimental studies on **phonotactic acceptability judgments** generally report **gradient results**
  (Scholes 1966, Frisch et al. 2000, Albright 2009, Daland et al. 2011, i.a.)
- **Dominant view:** gradient acceptability judgment results should be accounted for by a **gradient phonotactic grammar**
  (Albright 2009, Frisch et al. 2000, Hayes & Wilson 2008, Shademan 2006, Daland et al. 2011, i.a.)
  - Equate **probabilistic likelihood** with **phonotactic well-formedness**
- Gradience could also result from **experimental methodology**

Eliciting binary judgments but reporting averaged results

Eliciting Likert-scale judgments

Averaged Binary Judgments

- **Scholes (1966):** "could this be a word of English?" (**yes**/**no**)
- Report **number of participants** who gave **yes** judgment
- **Toy example:** 8/10 participants give **yes** judgment

Gradient Interpretation:

The word is **80% acceptable** in **any given speaker's grammar**

Erases possibility of individual variation

Gradience in averaged binary judgments ⇏ gradience in phonotactic representations

Categorical Interpretation:

The word was **completely licit** for 8 speakers and **completely illicit** for 2

Individual variation causes gradience when averaged over speakers

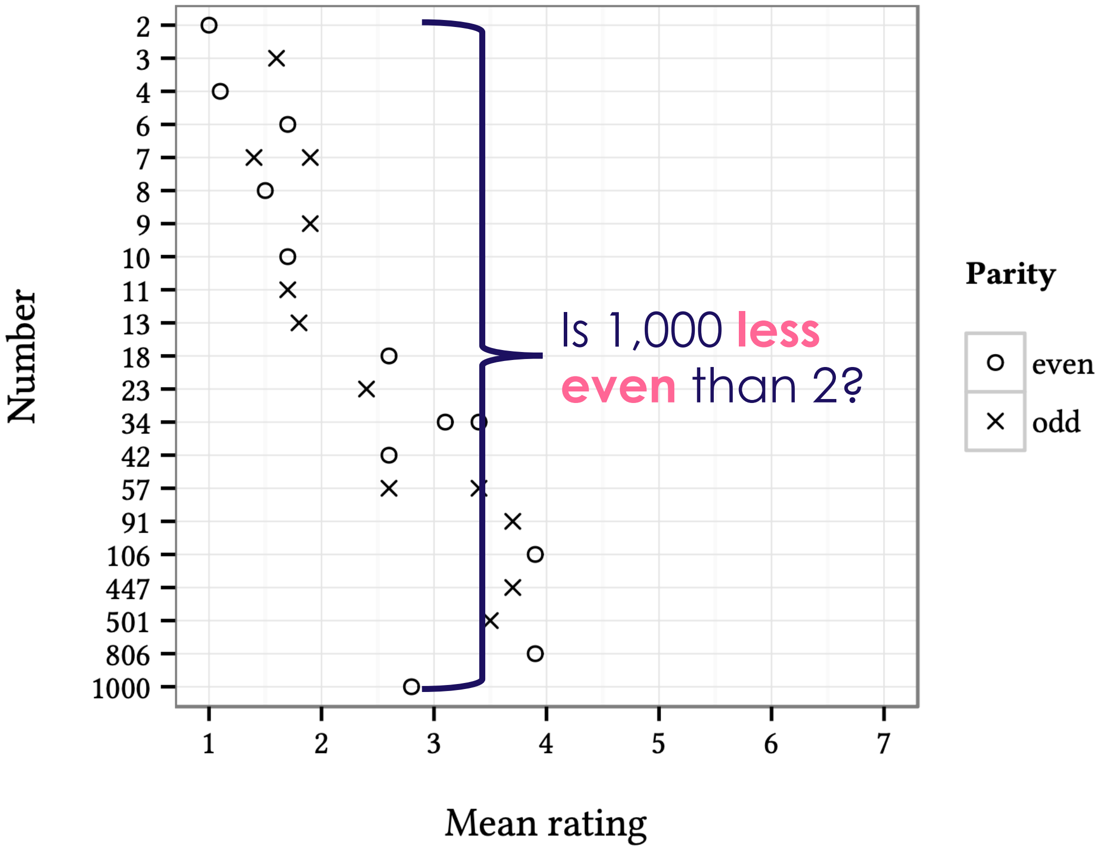
Unfortunately, by-speaker results not available for most studies

Likert-Scale Judgments

- **Daland et al. (2011):** "how likely is this word to become a word of English in the 21st century, on **a scale of 1-6?**"
- Report **average rating** of each word
- Likert scales are **known to produce task effects**

Likert-Scale Judgments

- Likert scales are **known to produce task effects**
- **Armstrong, Gleitman & Gleitman (1983):** how representative are numbers of **even** or **odd**?
  - **Gorman (2013):** similar **task effect** may occur for acceptability judgments
  - **Schütze (2011):** gradience may emerge when subjects try to **reconcile categorical grammar with gradient task**



Likert-Scale Judgments

- Likert scales are **known to produce task effects**
- **Armstrong, Gleitman & Gleitman (1983):** how representative are numbers of **even** or **odd**?
  - **Gorman (2013):** similar **task effect** may occur for acceptability judgments
  - **Schütze (2011):** gradience may emerge when subjects try to **reconcile categorical grammar with gradient task**

Gradience in Likert scale judgments ⇏ gradience in phonotactic representations

Gradient Representations?

Gradience in averaged binary judgments ⇏ gradience in phonotactic representations

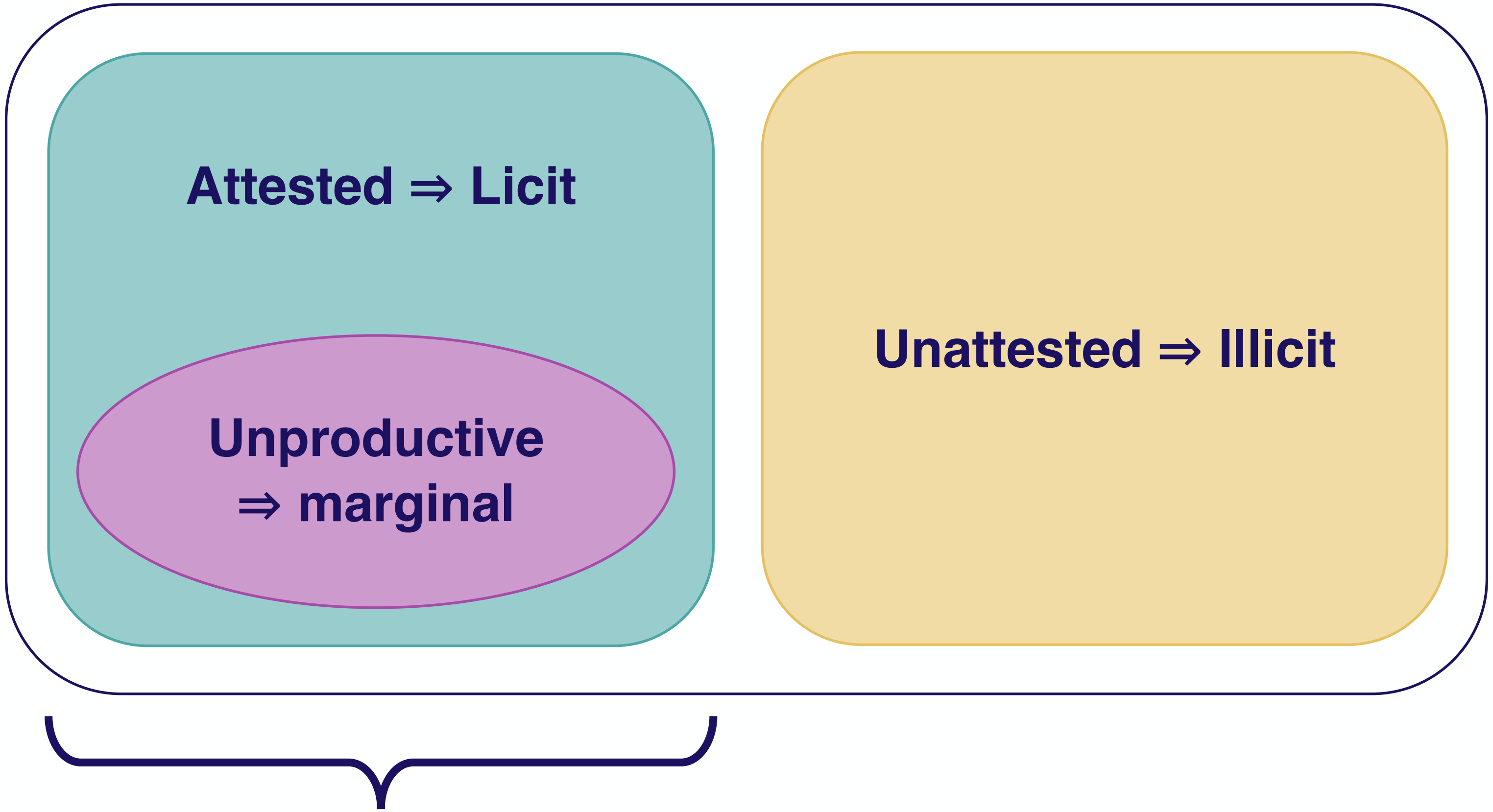
Gradience in Likert scale judgments ⇏ gradience in phonotactic representations

- **Gradient judgments ⇏ gradient** phonotactic grammar
- Possibility of **task effects ⇏ categorical** phonotactic grammar
- Some reasons to favor a **categorical** approach
  - We can successfully elicit **categorical judgments**
  - Binaries are simpler and don't require **scalar computation**
  - Other parts of the grammar (e.g., syntax) are generally considered categorical ⇒ **internal consistency**

Outline

- Re-thinking the phonotactic grammar
  - Motivating observations
  - What's (not) in the phonotactic grammar
  - Phonotactic knowledge is non-linear
  - A positive phonotactic grammar
  - Phonotactic representations may be categorical
- **Working Proposal**
  - **Proposal: Sequence-Wise Generalization Learner**
  - Evaluation: English complex onsets
- Future work

Attestation vs. Licitness: Proposal



How do we learn whether a subcomponent is licit or marginal?

Motivating Observations

Licit: [sp]-onset

- Occurs before a **wide range of vowels**
  - *spat, spell, spot, sputter*
- Belongs to **[s]-[voiceless-stop] onsets**
  - **[sp]**, **[st]**, **[sk]** all licit

Marginal: [sf]-onset

- Occurs before a **limited number of vowels**
  - *sphere, sphinx*
- Only similar onset = **[sv]**
  - ***svelte*** – also marginal

Working Proposal: "combinatorial power" of syllable sub-components related to licitness

Proposal: Measuring Combinatorial Power

- **The Tolerance-Sufficiency Principle**
  - Threshold for generalization based on **computational efficiency**
    - Children will generalize a rule when it's more efficient to
  - Given a rule ***R*** applicable to ***N*** types and seen applying to ***M*** of those types, **generalize the rule iff:**

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

Yang (2016)

Proposal: Measuring Combinatorial Power

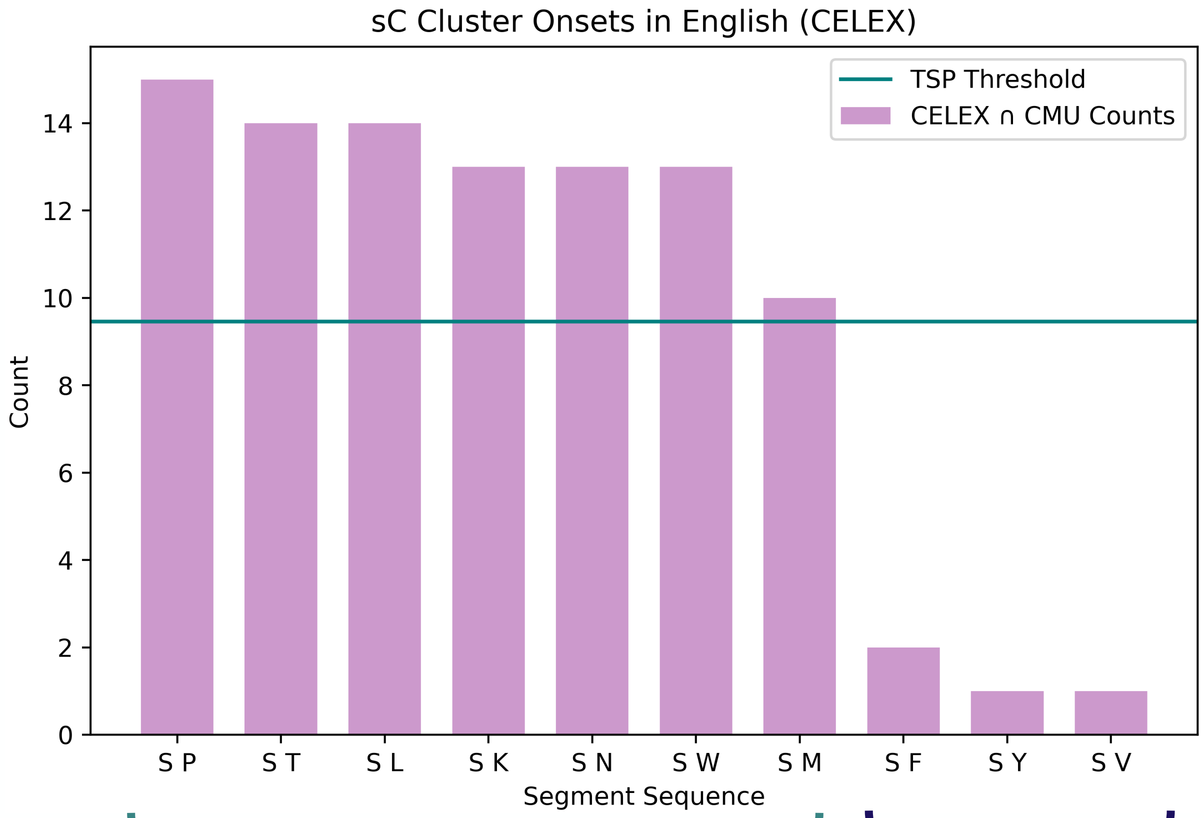
- **The Tolerance-Sufficiency Principle**
  - In a language with ***N* possible nuclei**, an attested onset/coda is licit iff it occurs with **at least *M*** of those nuclei and

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

Yang (2016)

Illustration: English **[sC]** Onsets

Number of distinct nuclei an onset is attested with



Tolerance Principle Threshold

Licit

Marginal

Sequence-Wise Generalization Learner

- **Recursive, feature-based subdivision** to learn phonotactics as **increasingly-specific sequences of feature sets**
  - Parallel to **Belth, Payne et al. (2021)** for morphological learning
- At each step, intersect all subcomponents in the current input to give some **underspecified sequence *S***
  - If **sufficiently many syllable subcomponents matching *S*** are licit, **add *S* to the set of licit subcomponents**
  - Otherwise, **subdivide the input** based on the most frequent feature set at the index in the string with greatest difference between ***N*** and ***M***
- If no generalization & no more features to subdivide on, ***S* is marginal**

Proposal: Measuring Generalizability

- Given some S, **are a sufficient number of subcomponents fitting it licit?**
  - Let $N = \prod n_i$ where n_i **= # segments that fit features at position** i
  - Let M be the number of **distinct syllable subcomponents observed that fit the entire feature set & are licit**
  - Check if $M - N \leq \frac{N}{\ln N}$

Proposal: Illustration

- Example: **English complex onsets**
  - N([+SIBILIANT] [-SON, -CONT]) = |{z, s} x {p, t, k, b, d, g}| = 12
  - M = number of licit subcomponents that fit [+SIBILIANT] [-SON, -CONT]
    - **{sp, st, sk} are licit ⇒ $M = 3$**
  - $N - M = 12 - 3 = 9 > \theta_{12} \approx 4.8$ ✕
  - **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: ±VOICE
  - Recurse: [+SIBILIANT] [-SON, -CONT, -VOI] vs. [+SIBILIANT] [-SON, -CONT, +VOI]

Outline

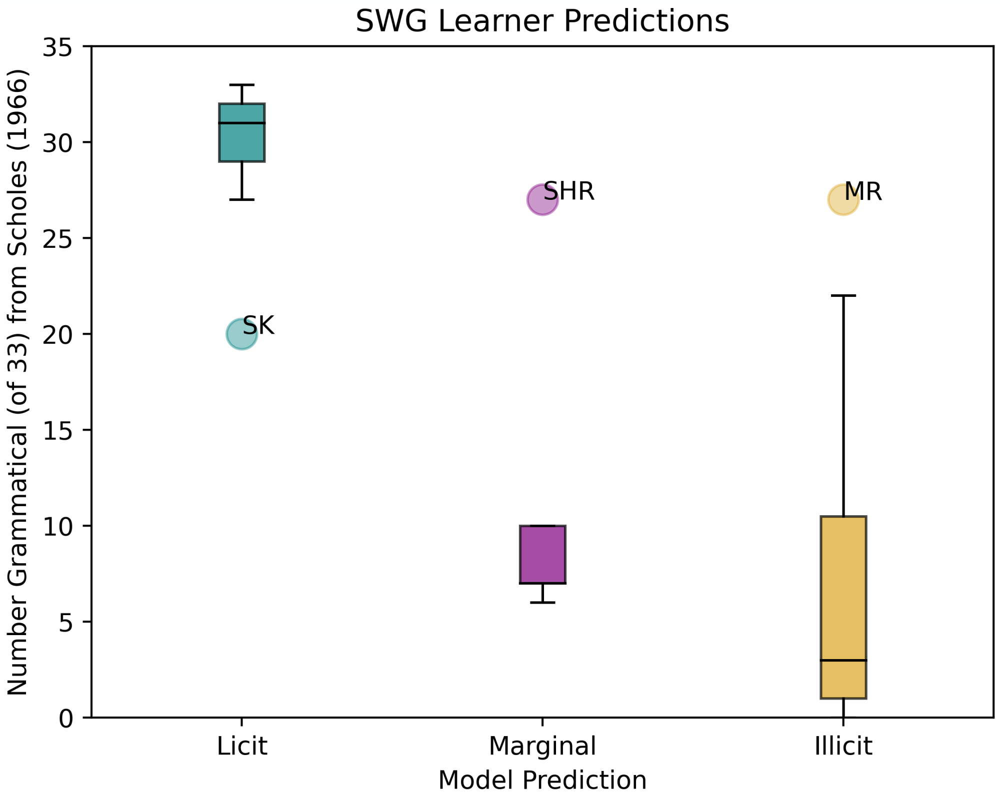
- Re-thinking the phonotactic grammar
    - Motivating observations
    - What’s (not) in the phonotactic grammar
    - Phonotactic knowledge is non-linear
    - A positive phonotactic grammar
    - Phonotactic representations may be categorical
- Working Proposal
    - Proposal: Sequence-Wise Generalization Learner
    - **Evaluation: English complex onsets**
- Future work

Experiment: English Complex Onsets

- Apply the model to real data: **English complex onsets**
  - **CELEX ∩ CMU**: ~41k words
  - Transcribed using the **CMU Pronouncing Dictionary**
  - Syllabified using the tool from **Gorman (2013)**
  - **Distinctive features** encoded for ARPABET based on those in Hayes & Wilson (2008)
    - Features can be **positive, negative, or unspecified**

Experiment: English Complex Onsets

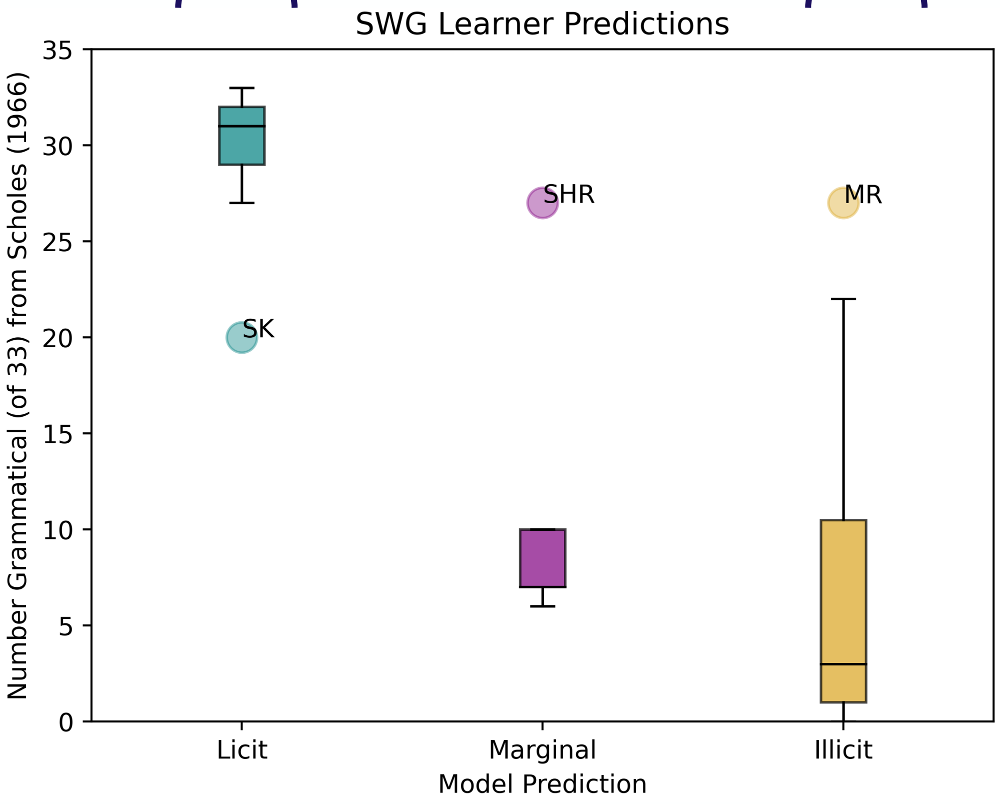
- **Scholes (1966):** complex onsets in **monosyllabic nonce words**
  - Binary decisions by **33 seventh graders**



Experiment: English Complex Onsets

Sequences predicted to be licit rated highly

Sequences predicted to be illicit rated poorly



One outlier per group

Experiment: English Complex Onsets

	Attestation Baseline	MaxEnt	SWG
Pearson's r	0.78	0.84	0.86
Spearman's TR ρ	0.74	0.79	0.78
Goodman-Kruskal γ	0.89	0.65	0.89
Kendall's τ_b	0.62	0.61	0.66

Gorman (2013): Attestation Baseline, MaxEnt

Our model!: SWG

Doesn't penalize ties: Pearson's r, Spearman's TR ρ, Goodman-Kruskal γ

Penalizes ties: Kendall's τ_b

Outline

- Re-thinking the phonotactic grammar
  - Motivating observations
  - What’s (not) in the phonotactic grammar
  - Phonotactic knowledge is non-linear
  - A positive phonotactic grammar
  - Phonotactic representations may be categorical
- Working Proposal
  - Proposal: Sequence-Wise Generalization Learner
  - Evaluation: English complex onsets
- **Future work**

Testing Model Predictions

- **Model predictions**
  - Initial stage of **conservatism**
  - Accumulate **sufficient evidence**
- Further **testing & comparison**
  - **Polish** complex onsets
  - More **judgments** (e.g., Daland et al. 2011)
  - Comparison with more **other models**
- **Experimental** investigation
  - Languages with **smaller vowel spaces**
  - **Artificial language** studies

Features or Segments?

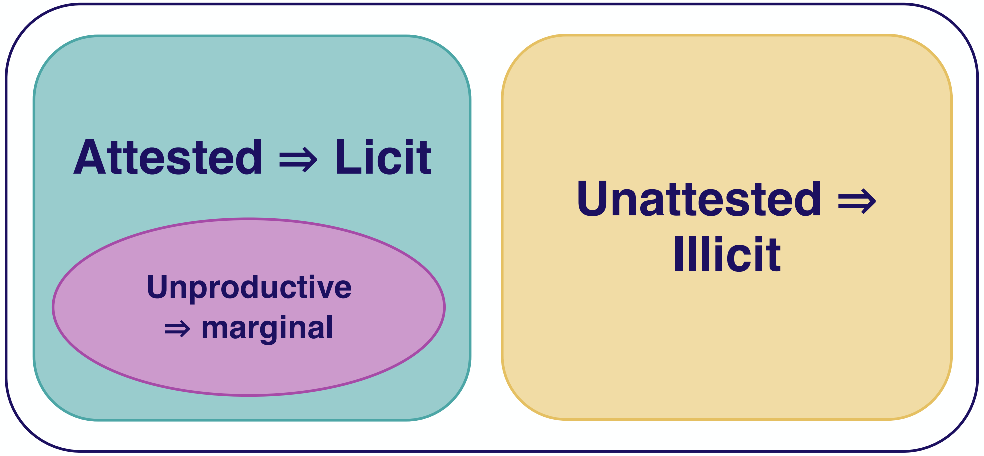
- Some evidence for **early underspecificaition**
  - English-learning children **cannot discriminate /bɪ/ and /dɪ/ when lexical contrast is implicated** but **can discriminate [b] and [d] when phonetic contrast is implicated** (Stager & Werker 1997)
  - French-learning 11-month-olds **do not prefer known words to alternates with different voicing or manner** (Hallé & Boysson-Bardies 1996)
- In practice, recursion almost always leads to **maximally-specified feature set sequences**
  - **No measurable differences** between segments & features in terms of correlation with human judgments **on full training**
- **Is phonotactic knowledge underspecified?**

Features or Segments?

- Can we make **phonotactic generalizations** based on features?
  - sp, st → sk ✅
  - sm, sn → sŋ ❌
- Is there something **special about [ŋ]** or is what's allowed/disallowed **too arbitrary** to allow for feature-based generalization?

Conclusions

- The **phonotactic grammar** is:
  - **Positive**
  - **Categorical**
  - **Syllable-based**
  - **Minimal:** contains no static restrictions



- Preliminary **learning model** in this framework
  - Uses **recursive search** with the **Tolerance-Sufficiency Principle**
  - Categorizes attested subcomponents as **licit** or **marginal**
  - Matches better with the judgments of **Scholes et al.** than MaxEnt

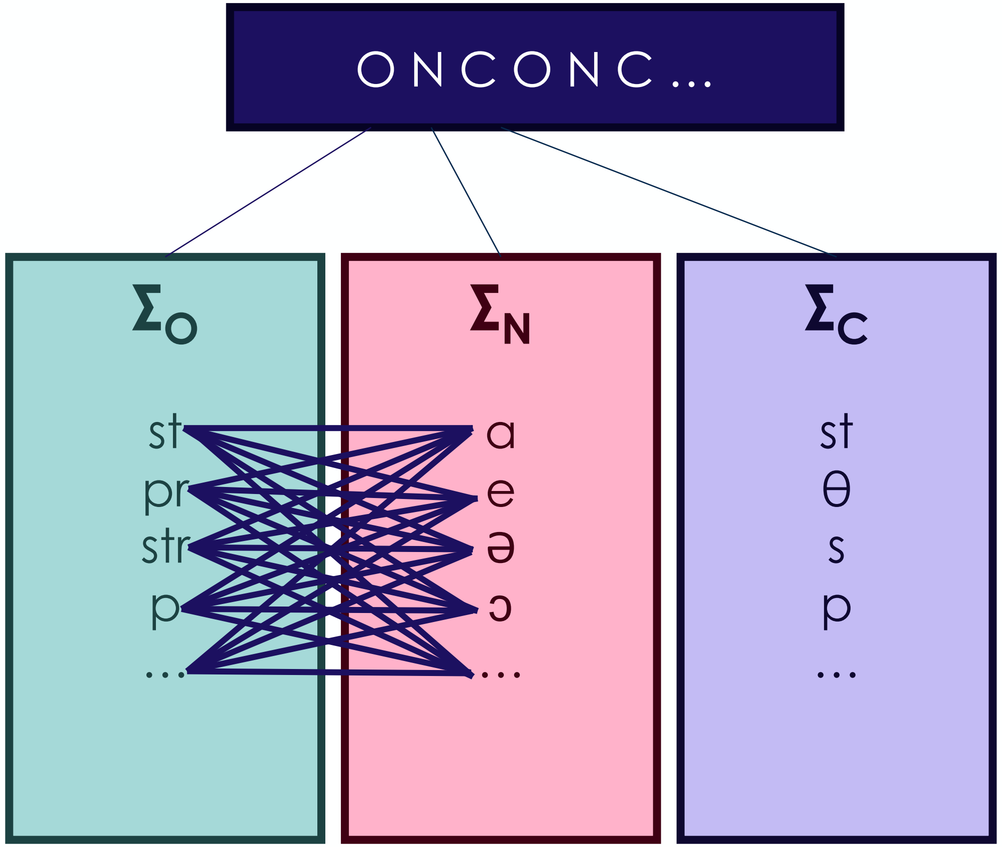
Thank you!!

I am grateful to **Jeff Heinz, Jordan Kodner**, and **Charles Yang** for their mentorship and **Kyle Gorman, Scott Nelson, Caleb Belth, Logan Swanson** and **Huteng Dai** for helpful discussion.

This work was funded by the **Institute for Advanced Computational Science** and **National Science Foundation**.

Extra Slides

Linear vs. Syllable-Based Representations



- We can think of the syllable-based representation being **SL over 3 alphabets**
- Can convert this to a **single, linear SL grammar** straightforwardly
  - For each transition, **add all possible combinations** except those that are disallowed (i.e. marginal)
  - The grammars will generate the same language but the **linear one doesn't build in generalization**

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 ϕ 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***

What's in the Phonotactic Grammar?

3 Arguments:

- G adient judgments incompatible with categorical grammar
  - Task effects are possible!
- Ph onotactic judgments are colored by orthography, lternations, experience with other languages, etc.
  - Is this not true of all linguistic judgments?
- S bjects have accurate judgments for languages they don't now
  - [pumehana] vs. [bɛzvzglɛndni]: Polish vs. Hawaiian
  - Just need to know Hawaiian doesn't allow CC
  - What about more nuanced judgments: [sfɪn] vs. [stɪn]?

MINIMAL

(1968)

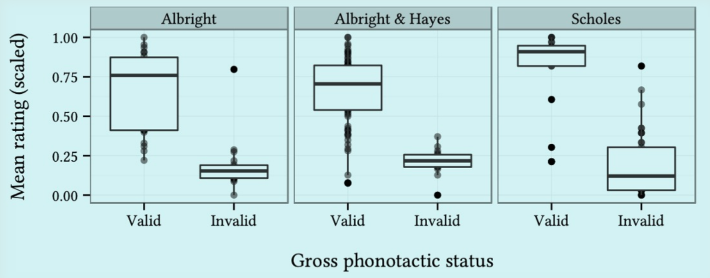
No phonotactic grammar:
Phonotacti ene izations play
no part in th honological
grammar but ther emergent,
metalin istic owledge
Reiss (2017)

Gradient vs. Categorical: Previous Work

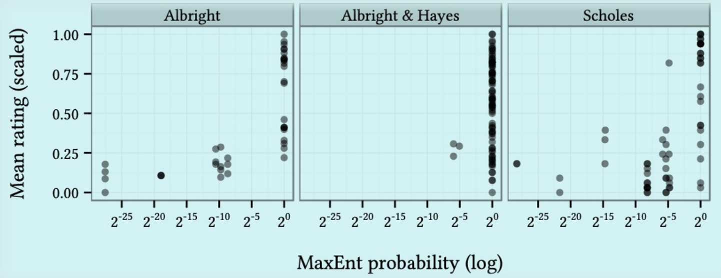
Gorman (2013)

Onsets & rimes are well-formed if they appear in a representative sample

***r* = 0.73** ***r* = 0.6** ***r* = 0.8**



***r* = 0.70** ***r* = 0.21** ***r* = 0.53**



Durvasula (2020)

Attestation-based categorical baselines perform at least as well as MaxEnt

When applied to the Scholes (1966) judgment data, type frequency of the onset sequence does not affect model fit, raising questions about where gradience in acceptability comes from.

Kostyszyn & Heinz (2020)

2-factor attestation for Polish word-initial complex onsets predicts acceptability better than the MaxEnt model:

2-factor Pearson's ***r* = 0.73**

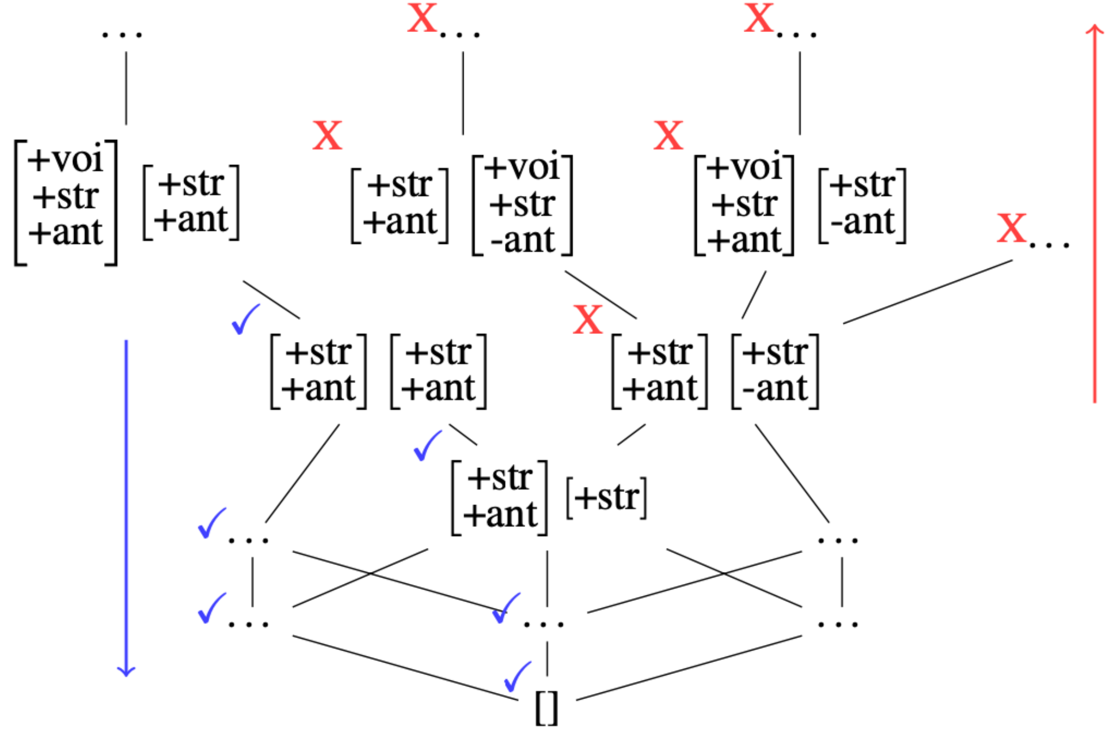
MaxEnt Pearson's ***r* = -0.07**

The Cost of Underspecification

- Far **more possible *k*-factors** when we allow for underspecification
  - Model with ***n* binary features**: $s < 2^n$ segments
    - **$s^k < (2n)^k$ possible *k*-factors**
  - **Underspecification ⇒ ternary** features: **$(3n)^k$ possible *k*-factors**
- **Interdefinition algorithm** less straightforward:
  - To determine if a *k*-subfactor **should be added to G+:**
    - Check if it's **in G-**
    - Also check if **any of its sub-factors or super-factors are in G-**

Positive & Negative Grammars: BUFIA

- Chandlee et al. (2019) & Rawski (2021):
  - Traversal that **exploits partially ordered hypothesis space**
  - Only continue to search if **som** ***k*****-factor matching the description is attested**
  - Otherwise, **learn constraint**
- Constraints of length **≤ *k***



Positive & Negative Grammars: BUFIA

- Payne (2024): positive grammars require all factors be of ***exactly* size *k*** in order to tile
  - Extend BUFIA to learn both positive & negative grammars:
  - A factor is **allowed** if:
    - All sequences matching it are **attested**
    - None of the sequences matching it are **unattested**
  - A factor is banned if:
    - All sequences matching it are **unattested**
    - None of the sequences matching it are **attested**
  - **Same learning guarantees as BUFIA!**

	$\in G$	$\notin G$
$\forall$	**Positive Grammar** (Equation 13)	**Negative Grammar** (Equation 10)
$\exists$	**Negative Grammar** (Equation 11)	**Positive Grammar** (Equation 14)