Marginal Sequences are Licit but Unproductive

Sarah Payne
sarah.payne@stonybrook.edu
Stony Brook University
iACS Institute for Advanced Computational Science

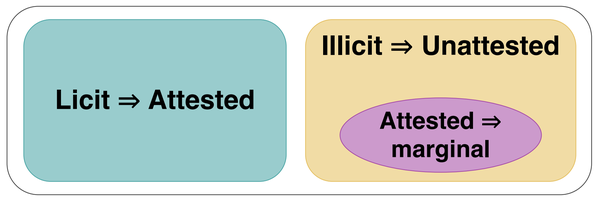
Marginal Sequences in Phonotactic Theory

How are Attestation and Licitness related?

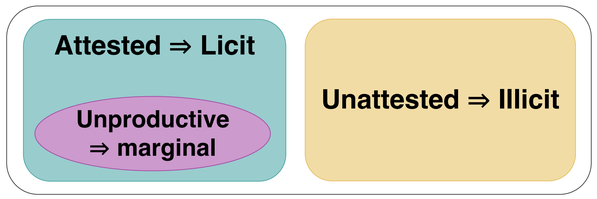
- Attested subsequences are generally Licit
- Unattested subsequences are generally Illicit

Where do Marginal Forms fit in?

- Previous approaches: Illicit but Attested (Hyman 1975)



- Our approach: Attested but Unproductive



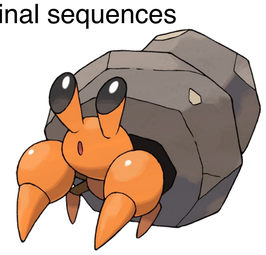
Evidence for Our Model

- **Borrowings:** not repaired

	Spanish	Japanese	English
German: /pfɪtsɐ/	/fajser/	/ɸaid͡za/	/faɪzɹ/
Italian: /spagetti/	/espageti/	/sɯpagetti/	/spəgɛti/
Greek: /sfiŋks/	/esfinxe/	/sɯɸinkɯsɯ/	/sfɪŋks/
Greek: /sfaira/	/esfera/	(sɯɸia)	/sfɪəɾ/

- **New Words:** may contain marginal sequences


spheal


dwebble

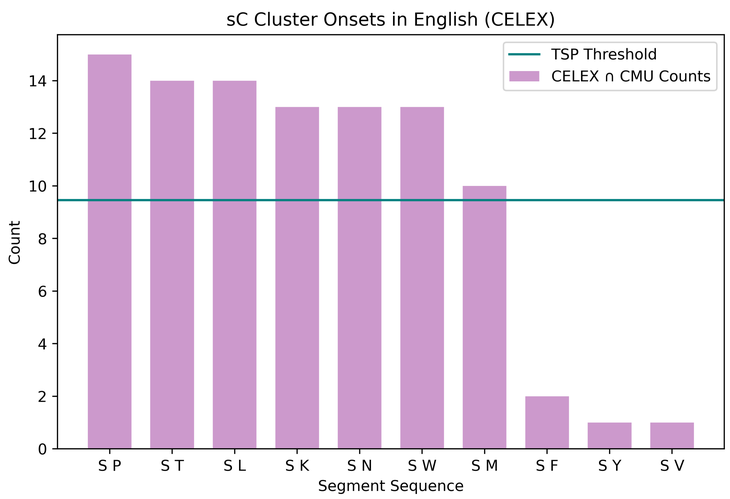
- **Production & Perception Errors**
  - Speakers **struggle to produce illicit sequences**
  - **97% production accuracy** on **/#sC/** sequences by English speakers
    - **C** ∈ {**f, p, t, k, m, n**} (Davidson 2006)

Selected References:
Hayes & Wilson 2008. A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry.*
Davidson 2006. Phonology, phonetics, or frequency: Influences on the production of non-native sequences. *Journal of Phonetics.*
Gorman 2013. *Generative Phonotactics.* UPenn Dissertation.
Hyman 1975. *Phonology: Theory and Analysis.* Harcourt Press.
Kabak & Idsardi 2007. Perceptual distortions in the adaptation of English consonant clusters: Syllable structure or consonantal contact constraints? *Language and Speech.*
Yang 2016. *The Price of Linguistic Productivity: How Children Learn to Break the Rules of Language.* MIT Press.

Formalizing Marginal vs. Licit with the TSP

- **Licit vs. Marginal as a Difference in Productivity**
  - **Licit Onsets/Codas:** occur with a **sufficiently diverse set of nuclei**
    - Occurs with **most nuclei** ⇒ can occur with all
  - **Marginal Onsets/Codas:** can occur with only a few, **lexicalized** nuclei
    - Occurs with a **few nuclei** ⇒ memorize those nuclei
- **The Tolerance Principle (TSP, Yang 2016):**
  - In a language with ***N*** possible nuclei, an attested onset/coda is **Licit** if it occurs with at least ***M*** of these possible nuclei and:

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



Model: Sequence-Wise Generalization Learner (SWG)

- **Motivation & Assumptions:**
  - Phonotactic knowledge **represented over syllables**
  - Representations initially **featurally-underspecified** during acquisition

We present a Syllable-Based computational model that learns a Positive Phonotactic Grammar categorizing forms as Licit, Marginal, or Illicit.

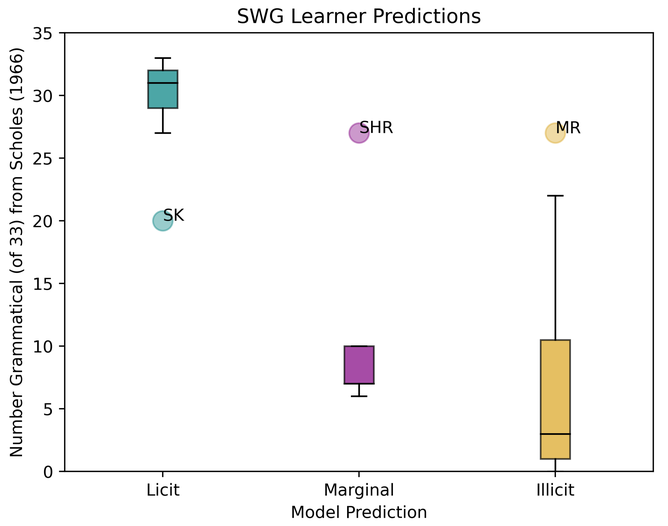
- **Learning Algorithm: recursive, feature-based subdivision** to learn phonotactics as **increasingly-specific sequences of feature sets**
  - At each step, intersect all sequences in current input to give underspecified sequence ***S***
  - If **sufficiently many sequences** matching ***S*** are licit, **add *S* to set of licit sequences**
  - Otherwise, **subdivide the input** based on the most frequent feature at the index in the string with the greatest difference between ***N*** and ***M,*** and recurse
  - If **no generalization & no more features to subdivide on**, then ***S* is marginal**

Acknowledgements:
I am grateful to Jeff Heinz, Jordan Kodner, Charles Yang, Scott Nelson, Salam Khalifa, Felix Fonseca, Kyle Gorman, and Huteng Dai for helpful discussion. This work was supported by the Institute for Advanced Computational Science (IACS) Graduate Research Fellowship and the National Science Foundation (NSF) Graduate Research Fellowship Program under NSF Grant No. 2234683. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the IACS or the NSF.

Data

- **Training:**
  - **CELEX ∩ CMU Pronouncing Dictionary:** ~41k words
  - Syllabify and extract syllable constituents (Gorman 2013)
  - Phonological Features from Hayes & Wilson 2008
- **Judgments:**
  - **Scholes:** complex onsets in monosyllabic nonce words
  - Binary decisions by 33 seventh graders

Results



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

Future Work

- **Developmental Implications**
  - Model predicts **initial stage of conservatism**
  - Must **accumulate sufficient evidence** for licitness
- **Further comparisons**
  - Human judgments on English & **other languages**
  - Comparison to H&W and **other models**
- How can we learn **syllable contact constraints** in this framework?
- How does SWG fare on languages with **smaller vowel spaces?**
  - Prediction: **more onsets/codas will pass TSP** and be licit because ***N* will be smaller**