

# A computational study of gaps in obstruent inventories

1. OVERVIEW	2. TASK & DATA DESCRIPTION	4. RESULTS
<ul> <li>Goal: comparing theories of inventory structure in accounting for the distribution of obstruent gaps</li> <li>Two major types of theories <ul> <li>segment markedness, implicational universal (e.g., De Lacy, 2003; Gamkrelidze, 1975; Greenberg, 1966)</li> <li>symmetry or economy of inventory and feature system (e.g., Clements, 2003; Dunbar &amp; Dupoux, 2016)</li> </ul> </li> <li>Gaps and inventory shapes: <ul> <li>distribution of gaps affects the shape of an inventory</li> </ul> </li> <li>Focus on obstruents: <ul> <li>a more homogeneous subset as a start</li> <li>claims on obstruents' interactions in distribution (Gamkrelidze, 1975)</li> </ul> </li> </ul>	<ul> <li>Task: identifying the held-out sound from a gapped inventory (cf. Cotterell &amp; Eisner, 2017) <ul> <li>Gap: absence of an [α voice] sound in a certain PoA when a [-α voice] sound exists in the same PoA</li> <li>Counterpart of gap: an attested sound with a different PoA, paired with gap to form a data point</li> </ul> </li> <li>Example: actual inventory from Wogeo (Exter 2003): <ul> <li>b</li> <li>d</li> <li>g</li> <li>t</li> <li>k</li> <li>v</li> <li>Gaps: /p/, /z/</li> <li>f</li> <li>s</li> <li>Data points: /p/-/t/, /p/-/k/, /z/-/v/</li> </ul> </li> <li>Data source: the PHOIBLE database (Moran et al., 2014) <ul> <li>2155 inventories in total, phonetically detailed feature set</li> <li>subset in this study: [-soronant] sounds, 1874 inventories</li> </ul> </li> </ul>	<ul> <li>Models with better grounded marked</li> <li>Grounded marked lowed by voiceles</li> <li>Segment frequent</li> <li>additional su frequent that</li> <li>artificial neur segments to performance</li> <li>Feature-systemic</li> <li>(3) Breakdown of mod Grounded Markedness</li> </ul>
<b>3. MODELS</b> (1) Example data points from Wogeo         b       d       g         -       t       k       (1a)         -       -       k       (1b)         -       -       Is /v/ or /z/ attested?       v       -       Is /p/ or /t/ atte         f       s       correct answer: /v/       f       s       correct answer	<ul> <li>Feature-systemic Models         <ul> <li>Local and Global Symmetry (Dunbar &amp; Dupoux, 2016)</li> <li>Local symmetry: better if more pairs of segments differ only in one feature</li> <li>/t/</li> </ul> </li> </ul>	Jo       50         25       0         Feature Entropy         100         75         50         25         0         50         25         0         50
<ul> <li>Markedness Models         <ul> <li>Grounded markedness</li> <li>defined with constriction site and aerodynamics</li> <li>voiced obstruents with a backer PoA are marked</li> <li>voiceless obstruents with a fronter PoA are marked</li> <li>better motivated for stops (e.g., Westbury &amp; Keating, 1986; Ohala, 1983; Ohala &amp; Riordan, 1979)</li> <li>(1a): /v/ is fronter and less marked</li> </ul> </li> </ul>	<ul> <li>interpretation: filling in the space of possible feature combinations in a more symmetrical way</li> <li>(1b): having /t/ gives /t/-/d/, /t/-/s/ having /p/ gives /p/-/k/, /p/-/b/, /p/-/f/, /p/-/s/</li> <li>decision: /p/ is attested (incorrect!)</li> <li>* Global symmetry: better if the use of [+] and [-] features are more balanced</li> <li>interpretation: one-feature natural classes with balanced sizes are preferred</li> </ul>	(4) Percentage of gap different places of single of the second se
<ul> <li>Segment frequency markedness</li> <li>typologically more rare → more marked</li> <li>rationale: typological data as approximation of markedness based on articulation, perception, or analytic bias</li> <li>(1a): /z/ and alveolars in general are both typologically more frequent than /v/ and labiodentals</li> <li>decision: /z/ is attested (incorrect!)</li> </ul>	<ul> <li>(1b): having /p/ requires [strident] and increases the amount of [-] values</li> <li>decision: /t/ is attested (correct!)</li> <li>Feature Entropy/Economy (Mukherjee et al. 2007)</li> <li>* entropy measure: ∑<sub>f∈F</sub>(-<sup>p<sub>f</sub></sup>/<sub>N</sub> log<sub>2</sub> <sup>p<sub>f</sub></sup>/<sub>N</sub> - <sup>q<sub>f</sub></sup>/<sub>N</sub> log<sub>2</sub> <sup>q<sub>f</sub></sup>/<sub>N</sub>)</li> <li>* more skewed distribution of [+] and [-], fewer required features → lower entropy → better economy</li> </ul>	<ul> <li><b>5. DISCUSSION</b></li> <li>Properties of the sare likely to be gate</li> <li>grounded mate</li> <li>good perform</li> </ul>
(2) Feature table* for data point (1b), /p/ vs /t/ $ \begin{array}{c ccccccccccccccccccccccccccccccccccc$	* prefers the segment that gives an inventory better pre- dictability * (1b): having /t/: 3.43 (1 + 0.81 × 3) · voice: $(-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8}) = 1$ · high, dist., labiodental: $(-\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8}) = 0.81$ having /p/: 3.68 (1 + 0.81 × 2 + 0.53 × 2) · voice: $(-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8}) = 1$ · high, labiodental: $(-\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8}) = 0.81$ · dist., strident: $(-\frac{1}{8}\log_2\frac{1}{8} - \frac{7}{8}\log_2\frac{7}{8}) = 0.53$	<ul> <li>There is an expective usefulness of - stops that had ating burst a</li> <li>Feature-systemic         <ul> <li>Feature-systemic</li> <li>what we lear more symmet</li> <li>additional expression</li> </ul> </li> </ul>

OVEDVIEW	2 TACK & DATA DESCRIPTION	
		4. NESULIS
Goal: comparing theories of inventory structure in accounting for the distribution of obstruent gaps	<ul> <li>Task: identifying the held-out sound from a gapped inventory (cf. Cotterell &amp; Eisner, 2017)</li> </ul>	<ul> <li>Models with bette grounded marked</li> </ul>
Two major types of theories	- Gap: absence of an [ $\alpha$ voice] sound in a certain PoA when a [ $-\alpha$ voice] sound exists in the same PoA	<ul> <li>Grounded marked lowed by voiceles</li> </ul>
Lacy, 2003; Gamkrelidze, 1975; Greenberg, 1966)	<ul> <li>Counterpart of gap: an attested sound with a different PoA,</li> </ul>	<ul> <li>Segment frequent</li> </ul>
<ul> <li>– symmetry or economy of inventory and feature system</li> <li>(e.g. Clements 2003: Dunbar &amp; Dupoux 2016)</li> </ul>	<ul> <li>Example: actual inventory from Wogeo (Exter 2003):</li> </ul>	- additional su
Gaps and inventory shapes:	b d g	– artificial neu
<ul> <li>distribution of gaps affects the shape of an inventory</li> </ul>	- L K V $-$ Gaps: /p/, /z/ f a Determinister /p/ /t/ /p/ /l// /z/ /v/	segments to performance
<ul> <li>– a more homogeneous subset as a start</li> </ul>	<ul> <li>Data points. /p/-/k/, /2/-/k/</li> <li>Data source: the PHOIBLE database (Moran et al., 2014)</li> </ul>	<ul> <li>Feature-systemic</li> </ul>
<ul> <li>– claims on obstruents' interactions in distribution (Gamkre- lidze, 1975)</li> </ul>	<ul> <li>2155 inventories in total, phonetically detailed feature set</li> <li>subset in this study: [</li></ul>	(3) Breakdown of mod
	- subset in this study. [-solonant] sounds, 1074 inventories	100 75
MODELS		50 25 0 0
Example data points from Wogeo	<ul> <li>Feature-systemic Models</li> </ul>	Teature Entropy
b d g - t k (1a) b d g 	- Local and Global Symmetry (Dunbar & Dupoux, 2016)	<b>0</b> 50 25
$  \frac{ s }{v}$ or $\frac{z}{z}$ attested? $v$ $ \frac{ s }{p}$ or $\frac{t}{z}$ attest	<pre>ted?     * Local symmetry: better it more pairs of segments dif-     fer only in one feature</pre>	0 stop fricative
Markedness Models	<ul> <li>interpretation: filling in the space of possible feature</li> <li>combinations in a more symmetrical way</li> </ul>	(4) Percentage of gap
<ul> <li>Grounded markedness</li> </ul>	· (1b): having /t/ gives /t/-/d/, /t/-/s/	
<ul> <li>defined with constriction site and aerodynamics</li> </ul>	having / <b>p</b> / gives / <b>p</b> /-/k/, / <b>p</b> /-/b/, / <b>p</b> /-/f/, / <b>p</b> /-/s/	so 25 0 100 75
<ul> <li>* voiced obstruents with a backer PoA are marked</li> <li>* voiceless obstruents with a fronter PoA are marked</li> </ul>	· decision: /p/ is attested (incorrect!)	<b>9</b> 100 100
* better motivated for stops (e.g., Westbury & Keating,	* Global symmetry: better if the use of [+] and [-] fea- tures are more balanced	
1986; Ohala, 1983; Ohala & Riordan, 1979) * (1a): /v/ is fronter and less marked	<ul> <li>interpretation: one-feature natural classes with bal- ancod sizes are proferred</li> </ul>	bilab lab-dent dent
* decision: /v/ is attested (correct!)	• (1b): having /p/ requires [strident] and increases	∎atteste
- Segment frequency markedness	the amount of [-] values	
* rationale: typological data as approximation of marked-	- Feature Entropy/Economy (Mukheriee et al. 2007)	
ness based on articulation, perception, or analytic bias	* entropy measure: $\sum_{f \in F} \left(-\frac{p_f}{N} \log_2 \frac{p_f}{N} - \frac{q_f}{N} \log_2 \frac{q_f}{N}\right)$	are likely to be ga
* (1a): /z/ and alveolars in general are both typologically more frequent than /v/ and labiodentals	<ul> <li>* more skewed distribution of [+] and [-], fewer required</li> </ul>	<ul> <li>grounded ma</li> </ul>
* decision: /z/ is attested (incorrect!)	* prefers the segment that gives an inventory better pre-	– good perforn
Feature table* for data point (1b), /p/ vs /t/	dictability	<ul> <li>There is an expect the usefulness of</li> </ul>
P     K     D     Q     T     S     V       voice     -     -     +     +     +     -     +	* (1b): having /t/: 3.43 (1 + 0.81 × 3) · voice: $(-\frac{4}{2} \log_2 \frac{4}{2} - \frac{4}{2} \log_2 \frac{4}{2}) = 1$	- stops that ha
high + - +	• high, dist., labiodental: $\left(-\frac{2}{8}\log_2\frac{2}{8} - \frac{6}{8}\log_2\frac{6}{8}\right) = 0.81$	ating burst a
distributed $ +$ $   -$ labiodental $      -$	having /p/: 3.68 (1 + 0.81 × 2 + 0.53 × 2)	Feature-systemic
strident         -<	• voice: $\left(-\frac{4}{8}\log_2\frac{4}{8} - \frac{4}{8}\log_2\frac{4}{8}\right) = 1$ • high. labiodental: $\left(-\frac{2}{3}\log_2\frac{2}{8} - \frac{6}{3}\log_2\frac{6}{3}\right) = 0.81$	- what we lear more symme
*The feature set is derived from removing features iteratively, starting from low- entropy ones, until having a set required for representing all segments uniquely	• dist., strident: $\left(-\frac{1}{8}\log_2\frac{1}{8} - \frac{7}{8}\log_2\frac{7}{8}\right) = 0.53$	– additional ex
	daajajan, /t/ ja attaatad (aarraat)	trony nrefer (

Sheng-Fu Wang, Department of Linguistics, New York University shengfu.wang@nyu.edu

\* decision: /t/ is attested (correct!)





er performance: segment frequency (68.4%), dness (63.2%), PoA frequency (61.0%)

edness: most successful with voiced stops, folss stops

ncy:

uccess in voiceless fricatives: /s/ and /f/ more an other gapped voiceless fricatives

ural networks, trained to map feature values of decision on gap/counterpart, can improve the e to 73% while having a similar breakdown

## models are not good at predicting gaps



## apped inventories having gaps/counterparts in articulation



segments are crucial in predicting whether they aps

- narkedness, segment frequency
- mance without paying attention to the inventory cted difference between stops and fricatives in a simple grounded markedness model
- ave more difficulty in maintaining voicing or creare more likely to be gapped
- models and predicting gaps:
- rn here: not every segment makes an inventory etrical or reduces a system's entropy
- experiment: global symmetry and feature entropy prefer gapless inventories (as defined in this study)