# A computational study of gaps in obstruent inventories

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

## 1. Overview

- Goal: comparing theories of inventory structure in accounting for the distribution of obstruent gaps
- Two major types of theories
  - segment markedness, implicational universal (e.g., De Lacy, 2003; Gamkrelidze, 1975; Greenberg, 1966)
  - symmetry or economy of inventory and feature system (e.g., Clements, 2003; Dunbar & Dupoux, 2016)
- Gaps and inventory shapes:
  - distribution of gaps affects the shape of an inventory
- Focus on obstruents:
  - a more homogeneous subset as a start
  - claims on obstruents' interactions in distribution (Gamkrelidze, 1975)

## 2. Task & Data Description

- Task: identifying the held-out sound from a gapped inventory (cf. Cotterell & Eisner, 2017)
  - Gap: absence of an [$\alpha$ voice] sound in a certain PoA when a [$-\alpha$ voice] sound exists in the same PoA
  - Counterpart of gap: an attested sound with a different PoA, paired with gap to form a data point
- Example: actual inventory from Wogeo (Exter 2003):

| b | d | g | |
|---|---|---|---|
| – | t | k | |
| v | – | | Gaps: /p/, /z/ |
| f | s | | Data points: /p/-/t/, /p/-/k/, /z/-/v/ |

- Data source: the PHOIBLE database (Moran et al., 2014)
  - 2155 inventories in total, phonetically detailed feature set
  - subset in this study: [$-$soronant] sounds, 1874 inventories

## 3. Models

(1) Example data points from Wogeo

| b | d | g | | b | d | g | |
|---|---|---|---|---|---|---|---|
| – | t | k | **(1a)** | – | – | k | **(1b)** |
| – | – | | Is /**v**/ or /**z**/ attested? | v | – | | Is /**p**/ or /**t**/ attested? |
| f | s | | correct answer: /**v**/ | f | s | | correct answer: /**t**/ |

- **Markedness Models**
  - **Grounded markedness**
    - defined with constriction site and aerodynamics
    - **voiced** obstruents with a **backer** PoA are marked
    - **voiceless** obstruents with a **fronter** PoA are marked
    - better motivated for stops (e.g., Westbury & Keating, 1986; Ohala, 1983; Ohala & Riordan, 1979)
    - **(1a)**: /**v**/ is fronter and less marked
    - **decision: /v/ is attested (correct!)**
  - **Segment frequency markedness**
    - **typologically more rare** $\rightarrow$ more marked
    - rationale: typological data as approximation of markedness based on articulation, perception, or analytic bias
    - **(1a)**: /**z**/ and alveolars in general are both typologically more frequent than /v/ and labiodentals
    - **decision: /z/ is attested (incorrect!)**

(2) Feature table* for data point **(1b)**, /**p**/ vs /**t**/

| | p | t | k | b | d | g | f | s | v |
|---|---|---|---|---|---|---|---|---|---|
| voice | – | – | – | + | + | + | – | – | + |
| high | – | – | + | – | – | + | – | – | – |
| distributed | – | + | – | – | + | – | – | – | – |
| labiodental | – | – | – | – | – | – | + | – | + |
| strident | – | – | – | – | – | – | – | + | – |

*The feature set is derived from removing features iteratively, starting from low-entropy ones, until having a set required for representing all segments uniquely

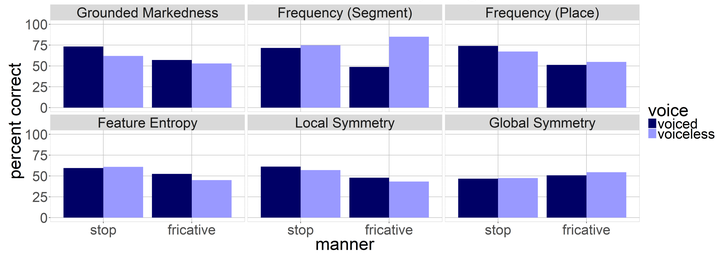
- **Feature-systemic Models**
  - **Local and Global Symmetry** (Dunbar & Dupoux, 2016)
    - **Local symmetry**: better if more pairs of segments differ only in one feature
      - interpretation: filling in the space of possible feature combinations in a more symmetrical way
      - **(1b)**: having /**t**/ gives /**t**/-/d/, /**t**/-/s/
        having /**p**/ gives /**p**/-/k/, /**p**/-/b/, /**p**/-/f/, /**p**/-/s/
      - **decision: /p/ is attested (incorrect!)**
    - **Global symmetry**: better if the use of [+] and [−] features are more balanced
      - interpretation: one-feature natural classes with balanced sizes are preferred
      - **(1b)**: having /**p**/ requires [strident] and increases the amount of [−] values
      - **decision: /t/ is attested (correct!)**
  - **Feature Entropy/Economy** (Mukherjee et al. 2007)
    - entropy measure: $\sum_{f \in F}(-\frac{p_f}{N}\log_2\frac{p_f}{N} - \frac{q_f}{N}\log_2\frac{q_f}{N})$
    - more skewed distribution of [+] and [−], fewer required features $\rightarrow$ lower entropy $\rightarrow$ better economy
    - prefers the segment that gives an inventory better predictability
    - **(1b)**: having /**t**/: 3.43 ($1 + 0.81 \times 3$)
      - 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 ($1 + 0.81 \times 2 + 0.53 \times 2$)
      - 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
    - **decision: /t/ is attested (correct!)**

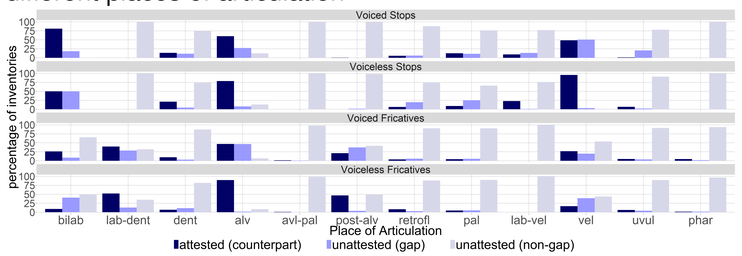
## 4. Results

- Models with better performance: segment frequency (68.4%), grounded markedness (63.2%), PoA frequency (61.0%)
- Grounded markedness: most successful with voiced stops, followed by voiceless stops
- Segment frequency:
  - additional success in voiceless fricatives: /s/ and /f/ more frequent than other gapped voiceless fricatives
  - artificial neural networks, trained to map feature values of segments to decision on gap/counterpart, can improve the performance to 73% while having a similar breakdown
- Feature-systemic models are not good at predicting gaps

(3) Breakdown of model performance by segment types




(4) Percentage of gapped inventories having gaps/counterparts in different places of articulation




## 5. Discussion

- Properties of the segments are crucial in predicting whether they are likely to be gaps
  - grounded markedness, segment frequency
  - good performance without paying attention to the inventory
- There is an expected difference between stops and fricatives in the usefulness of a simple grounded markedness model
  - stops that have more difficulty in maintaining voicing or creating burst are more likely to be gapped
- Feature-systemic models and predicting gaps:
  - what we learn here: not every segment makes an inventory more symmetrical or reduces a system's entropy
  - additional experiment: global symmetry and feature entropy prefer gapless inventories (as defined in this study)