PHONOLOGICAL (DIS)SIMILARITY
REDUPLICATION, CONFUSABILITY, AND THE LEXICON IN BENGALI

SAMEER UD DOWLA KHAN, REED COLLEGE
GLOSS COLLOQUIUM, 5 FEBRUARY 2016
OUTLINE

- Overview
  - Identity and similarity in phonology
- Echo reduplication
  - Identity avoidance, with a puzzle from English
  - Production data from Bengali
  - Gradient similarity avoidance
- Similarity metric
  - Shared natural classes
  - Weighted shared natural classes
- Other expressions of gradient similarity
  - Lexical statistics
  - Perceptual confusability
- Synthesis of results
Many processes incorporate **categorical identity**

- **Reduplication**
- **Sibilant harmony** in Chumash (Beeler 1970)
  - [kiʃkín] + [us] ⇒ *[kiʃkínus] ⇒ [kiskínus] ‘I saved it for him’
- **Haplology** (identity avoidance)
  - merry + -ly ⇒ merrily
  - silly + -ly ⇒ *sillily ⇒ silly (adv.)

Often broadened to a natural class: **categorical similarity**

- **Vowel harmony** in Turkish (Tosun 1999)
  - [kwz] + [ţar] ⇒ [kwzţar] ‘girls’
  - [jyz] + [ţar] ⇒ *[jyzţar] ⇒ [jyzlær] ‘faces’
But some phenomena in perception and the lexicon are best described as involving **gradient similarity**

- Lexical **cooccurrence** effects in Muna (Coetzee & Pater 2005)
  - [d] is found in fewer roots with [t] than with [n]
- Perceptual **confusability** in English (Cutler et al. 2004)
  - [tʃ] is misidentified as [t] more often than as [dʒ]

- It’s possible that cases of supposed categorical identity/similarity are in fact extreme cases of gradient similarity
  - cf. Vowel harmony in Hungarian (Hayes & Londe 2006)
GOAL FOR THIS TALK

- Present data illustrating the phenomenon of **fixed-segment echo reduplication** in Bengali

- Demonstrate that it is a case of **gradient similarity**

- Explore what kind of **metric** underlies the patterns seen

- Investigate **lexical and perceptual** expressions of similarity as well as a comparison
Echo reduplication

- \([\text{daktə} \text{ʃmFaktə}] \) ‘doctor\text{DISMISSIVE}’
- As opposed to \([\text{daktə} \text{daktə}] \) ‘(real/prototypical) doctor’
- Most common in lgs across southern Asia

Phonological properties

- Total reduplication
- Systematic replacement of some material in reduplicant (RED) with one or more fixed segments

Semantic properties

- Typically denotes \text{generalization}: ‘X, etc.’, ‘superset of X’
- In some lgs, it can also be \text{disparaging}
ECHO REDUPLICATION

- **Representative examples:**

  - **Turkish** $[m_F]$ (Southern 2005)
    - [kutu] ‘box’ $\Rightarrow$ [kutu $m_F$utu] ‘box(es), etc.’
    - [citap] ‘book’ $\Rightarrow$ [citap $m_F$itap] ‘book(s), etc.’
    - [ʃaka] ‘fun’ $\Rightarrow$ [ʃaka $m_F$aka] ‘easily’, ‘calmly’

  - **Eastern A-Hmao** $[ú_F]$ (Mortensen 2006)
    - [ámâ] ‘eye’ $\Rightarrow$ [ámú$_F$ âmâ] ‘eyes, ears, mouth, and nose’
    - [píndzâu] ‘demon’ $\Rightarrow$ [píndzú$_F$ píndzâu] ‘spirits of all kinds’
    - [kílåw] ‘strip of cloth’ $\Rightarrow$ [kílú$_F$ kílåw] ‘strips of cloth, etc.’
 Unlike prototypical reduplication, echo reduplication typically requires the base and RED to be non-identical.

- Unlike “emergence of the unmarked” cases of base-RED nonidentity, e.g. Sanskrit (Steriade 1988)
- Unlike “default fixed segmentism”, e.g. Yoruba (Alderete et al. 1999)

- Presence of the fixed segment should be enough to generate base-RED nonidentity...
- ...unless the fixed segment is identical to the segment it is meant to replace
IDENTITY AVOIDANCE

- [m]-initial words in Turkish [m_F] have no echo form
  - [para] ‘money’ ⇒ [para m_Fara] ‘money, etc.’
  - [masa] ‘table’ ⇒ *[masa m_Fasa] ‘towel, etc.’ ⇒ NO OUTPUT

- [m]-initial words in Abkhaz [m_F] take backup [tʃ_F] (Vaux 1996)
  - [gádžak’] ‘fool’ ⇒ [gádžak’ m_Fádžak’] ‘fool, etc.’
  - [maát] ‘money’ ⇒ *[maát m_Faát] ⇒ [maát tʃ_Faát] ‘money, etc.’

- In Classical Tibetan [a_F], base takes backup [o_F] (Beyer 1992)
  - [ndzog] ⇒ [ndza_F g ndzog] ‘jumbled up’
  - [glen] ⇒ [gla_F n glen] ‘very stupid’
  - [ŋan] ⇒ *[ŋa_F n ŋan] ⇒ [ŋan ŋo_F n] ‘miserable’
IDENTITY AVOIDANCE

- Through various means, lgs work to **avoid categorical identity** between base and RED in echo forms
- Survey of echo forms in >100 lgs of India found **identity avoidance** in every case (Trivedi 1990)

- Previous work on echo forms generally describe a straightforward case of **categorical identity avoidance**
- No one has yet confirmed that this avoidance pattern does not extend to natural classes, or that it is not gradient
A PUZZLE FROM ENGLISH

What about English [ʃm_F]?
- [daktœ] ‘doctor’ ⇒ [daktœ ʃm_Faktœ] ‘doctor\textsubscript{DISMISSIVE}’
- [skul] ‘school’ ⇒ [skul ʃm_Ful] ‘school\textsubscript{DISMISSIVE}’

Online survey, 190 respondents (Nevins & Vaux 2003)

Identity avoidance: 95–97% of speakers rejected echo forms with [ʃm_F] for the 3 [ʃm]-initial words
- [ʃmuz] ‘schmooze’ ⇒ *[ʃmuz ʃm_Fuz] ‘schmooze\textsubscript{DISMISSIVE}’

Interestingly, 30% of speakers also rejected echo forms with [ʃm_F] for the one [ʃn]-initial word... why??
- [ʃnaz] ‘schnozz’ ⇒ *[ʃnaz ʃm_Faz] ‘schnozz\textsubscript{DISMISSIVE}’
A PUZZLE FROM ENGLISH

Possible explanations:

The “two dialects” possibility
- 65% of subjects obey **identity avoidance**
- 30% obey **categorical similarity avoidance**, where \( [ʃn] \) and \( [ʃm] \) are of the same category: “sounds similar to \( [ʃm_F] \)”

The “matter of degree” possibility
- 95% obey **gradient similarity avoidance**, of whom:
  - 65% considered \( [ʃn] \) and \( [ʃm_F] \) are sufficiently dissimilar
  - 30% considered \( [ʃn] \) and \( [ʃm_F] \) are excessively similar
A PUZZLE FROM ENGLISH

- Another possible explanation: “this isn’t English”
  - **Humorous** and possibly **peripheral** to the language
  - **Less common** in English than in other lgs
  - `[ʃm]` is **highly marked**, restricted to **borrowings** from Yiddish
  - Construction is possibly borrowed from Yiddish (Southern 2005)
To understand if echo reduplication can employ gradient similarity avoidance, we need a lg in which:

- Echo reduplication is a fully productive, linguistic feature
- The fixed segment is a relatively unmarked sound
- The fixed segment has many similar sounds

**Bengali\(^1\) is an ideal test case**

- Default fixed segment \([t_F]^2\): crosslinguistically unmarked
- \([t]\) has high token freq. (definite marker & classifier \([-\text{ta}]\))
- Attested backup fixed segments \([m_F \ f_F \ p_F \ u_F]\) (Ray et al. 1966)
- Inventory has many \([t]\)-like sounds: \([t^h \ d \ d^h \ t \ t^h \ d \ t^c \ ...]\) (Khan 2010)

---

\(^1\) Specifically, urban colloquial Bangladeshi varieties

\(^2\) \([t^h \ d \ d^h]\) can be retroflex in Bengali, but are typically alveolar in these varieties (Khan 2010)
Does echo reduplication in Bengali involve...
- Categorical identity avoidance,
- Categorical similarity avoidance, or
- Gradient similarity avoidance?

If it is the latter, how can similarity be objectively measured on a gradient scale?

As a comparison, we can investigate other parts of Bengali phonology that expected to employ this gradient similarity:
- Lexical cooccurrence restrictions
- Perceptual confusability
EXPERIMENT I: PRODUCTION

- **Basic design:** native speakers produce echo RED for base stimuli with carefully-selected initial C.

- **Expectations:**
  - [kasi] ‘cough’ ⇒ [kasi t_fasi]
  - [tika] ‘vaccine’ ⇒ *[tika t_Fika] ⇒ [tika m_Fika] (identity violation)

- **Question: how will sounds similar to \([t_F]\) behave?**
  - \([tʰajʃ:a] \text{ ‘having stuffed’ } ⇒ [tʰajʃ:a t_fajʃ:a] \text{ (no violation)}?\)
    OR
  - \([tʰajʃ:a] ⇒ *[tʰajʃ:a t_fajʃ:a] ⇒ [tʰajʃ:a f_fajʃ:a] \text{ (similarity violation)}?\)
EXPERIMENT I: STIMULI

- **60 stimulus** words
  - Disyllabic stems
  - Content words: N, A, V (perfective participles)
- **2 registers** of urban colloquial Bangladeshi Bengali
  - High register: closer to written Kolkata Standard
  - Low register: closer to eastern regional varieties
- **Produced** by adult female speaker
  - Proficient in both registers
  - 2 reps per variety = 240 recordings
  - Recorded in sound-treated booth on Telex M-540 mic
60 test words fell under 3 conditions:
- **Identity**: [t]-initial words
- **Similarity**: words with [t]-like initials
  - Coronal obstruents [tʰ d ʈ ʈʰ tɬ sɭtʃʰ ʃ]
- **Control**: words with non-[t]-like initials
  - Coronal sonorants [n l ɭ]
  - Non-coronals [k h p f bʱ m]
## EXPERIMENT 1: STIMULI

Consonants of Bangladeshi Standard Bengali (Khan 2010)

<table>
<thead>
<tr>
<th></th>
<th>Labial</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Post-Alv</th>
<th>Velar/Glot</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stop</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>p b b̃</td>
<td>t̰ t̰ʰ</td>
<td>t̰ʰ d̰ d̰̃</td>
<td>t̰ʰ d̰ d̰̃</td>
<td>k k̃ h g̃ g̃̃</td>
</tr>
<tr>
<td><strong>Affricate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>tζ tζ̃</td>
<td>tζ̃ dζ dζ̃</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fricative</strong></td>
<td>f</td>
<td>s</td>
<td></td>
<td></td>
<td>h</td>
</tr>
<tr>
<td><strong>Liquid</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>l r</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nasal</strong></td>
<td>m</td>
<td></td>
<td>n</td>
<td></td>
<td>(ŋ)</td>
</tr>
</tbody>
</table>
# EXPERIMENT I: STIMULI

- **Consonants of Bangladeshi Standard Bengali** (Khan 2010)

## Identity | Similarity | Control

<table>
<thead>
<tr>
<th></th>
<th>Labial</th>
<th>Dental</th>
<th>Alveolar</th>
<th>Post-Alv</th>
<th>Velar/Glot</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Stop</strong></td>
<td>p b bʱ</td>
<td>t tʰ d dʱ</td>
<td>t tʰ d dʱ</td>
<td>k kʰ g gʱ</td>
<td></td>
</tr>
<tr>
<td><strong>Affricate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Fricative</strong></td>
<td>f</td>
<td>s</td>
<td></td>
<td></td>
<td>h</td>
</tr>
<tr>
<td><strong>Liquid</strong></td>
<td></td>
<td></td>
<td>l r</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Nasal</strong></td>
<td>m</td>
<td></td>
<td>n</td>
<td></td>
<td>(ŋ)</td>
</tr>
</tbody>
</table>
EXPERIMENT I: SETUP

- **30 speakers** of Bengali
  - Varied dialect background
  - Residents of CA
  - Paid $10

- **Heard stimulus**
  - Participant selected preferred register
  - Order randomized for each speaker

- **Asked to produce** echo reduplicated form
  - [kaʃi] ‘cough’ ⇒ [kaʃi tFaʃi] ‘cough, etc.’ given as example

- Responses were transcribed
Identity words will never use \([t_F]\)

Control words will always use \([t_F]\)

Similarity words are what are being tested:
- Hypothesis 1: \(\text{similarity} = \text{control}\) (categorical identity)
- Hypothesis 2: \(\text{similarity} = \text{identity}\) (categorical similarity)
- Hypothesis 3: \(\text{similarity}\) is on a continuum

\[
\begin{align*}
\text{Identity} & \quad \text{Similarity} & \quad \text{Control} \\
* [t...t_F] & \not\equiv [t^h...t_F] & = [b^f...t_F] \\
\end{align*}
\]

\([t^h\text{ajs:a}] \Rightarrow [t^h\text{ajs:a} \ t_F\text{ajs:a}]\)
EXPERIMENT I: HYPOTHESES

- **Identity** words will never use $[t_F]$.
- **Control** words will always use $[t_F]$.
- **Similarity** words are what are being tested:
  - Hypothesis 1: $\text{similarity} = \text{control}$ (categorical identity).
  - Hypothesis 2: $\text{similarity} = \text{identity}$ (categorical similarity).
  - Hypothesis 3: similarity is on a continuum.

\[
\begin{align*}
\text{Identity} & : *[t...t_F] = *[t^h...t_F] \\ 
\text{Similarity} & : *[t^h...t_F] \neq [b^f...t_F] \\ 
\text{Control} & :
\end{align*}
\]

$[t^h\text{ajʃːa}] \Rightarrow *[t^h\text{ajʃːa} t_F\text{ajʃːa}] \Rightarrow [t^h\text{ajʃːa} \text{m}_F\text{ajʃːa}]$
EXPERIMENT I: HYPOTHESES

- **Identity** words will never use $[t_F]$
- **Control** words will always use $[t_F]$
- **Similarity** words are what are being tested:
  - Hypothesis 1: $\text{similarity} = \text{control}$ (categorical identity)
  - Hypothesis 2: $\text{similarity} = \text{identity}$ (categorical similarity)
  - Hypothesis 3: $\text{similarity}$ is on a continuum

\[
\begin{align*}
\text{Identity} & \quad \neq \quad \text{Similarity} & \quad \neq \quad \text{Control} \\
*[t...t_F] & \quad ?[t^h...t_F] & \quad [b^f...t_F] \\
[t^{h}\text{ajʃː}a] & \Rightarrow [t^{h}\text{ajʃː}a \ t^{f}\text{ajʃː}a] & \sim [t^{h}\text{ajʃː}a \ f^{f}\text{ajʃː}a]
\end{align*}
\]
Hypothesis 3 was borne out
Similarity words lie on a continuum
- Disprefer [t_F] but not outright ungrammatical
- Some consonants are more [t]-like in behavior than others

Seems like Cs that take [t_F] less often are also phonetically closer to [t]

Least likely to be replaced by [t_F]  Most likely to be replaced by [t_F]
EXPERIMENT I: RESULTS

% $[t_F]$ use in RED

Base-initial consonant
EXPERIMENT I: RESULTS

Base-initial consonant

% $[t_F]$ use in RED

$\text{t} \quad ^{th} \quad \text{d} \quad \text{t} \quad \text{s} \quad ^{th} \quad \text{k} \quad \text{t} \text{c} \quad \text{d} \quad \text{j} \quad \text{n} \quad \text{l} \quad \text{b}^{^h} \quad \text{f} \quad \text{p} \quad \text{m}$
Echo reduplication in Bengali appears to incorporate a notion of **gradient similarity avoidance**
- No straightforward clustering of consonants
- Heavy overlap across clusters
- Like the “matter of degree” hypothesis from English puzzle
NEW QUESTIONS

- We should confirm our suspicion that our reduplication data can be modeled on an **objective scale of similarity**
- Is there a **metric** that Bengali speakers are using to calculate the similarity of an initial C and [t]?
- Metric has to be **gradient**, possibly **language-specific**
Best-known option is **shared natural classes (SNC) metric** (Frisch et al. 1995/2004)

Similarity of two Cs is based on the **number of natural classes they share** in the inventory

Universal claim with language-specific application

**Hypothesis:** the more natural classes shared between a C and [t], the less likely it will take [t\_F] in its echo RED
In the SNC metric, similarity of $C_1$ and $[t]$ is quantified as:

$$\text{sim}(C_1, t) = \frac{\text{# natural classes shared by } (C_1, t)}{\text{# shared natural classes} + \text{# non-shared natural classes}}$$

Compared SNC-similarity (line) to Exp 1 results (bars)
$r^2 = .584, p < 0.01^*$
The SNC metric does an okay job overall ($r^2 = .584$).

However, the area where it crucially fails to predict the data is the *similarity set* (coronal obstruents).

The metric treats [t] as inherently more similar to [t̪] and [tɕ] than to [tʰ]... is there a way to adjust that?
Original SNC metric derives directly from the phoneme inventory and feature set

But what if we maintain the basic model but incorporate feature weights?

Let’s try a little thought experiment

Weighting [dist] over [spread gl]: the [t - ɾ] distinction can be “heavier” than the [t - ţʰ] distinction

If this improves our metric, we can then pursue the question of whether these weights are justified
In an SNC-like model with feature weights, similarity of $C_1$ and $[t]$ is quantified as follows: (Wilson, p.c.)

$$sim(C_1, t) = \exp\left(-\sum_{i=1}^{\#\text{features}} w_i (1 - \delta_i(C_1, t))\right)$$

- $w_i = \text{weight of the feature } f_i$
- $\delta_i(C_1, t) = 1$ (feature value shared) or 0 (not shared)

Where weights are drawn from the variation in the reduplication results, as follows:
WEIGHTED SNC: METRIC

- Probability of \([t_F]\) use in the RED of a base with initial \(C_1\)

\[ P = \left( \frac{m!}{(n!(m-n)!) \left(1 - \text{sim}(C_1, t)\right)^n \left(\text{sim}(C_1, t)\right)^{m-n}} \right) \]

\(P\) = probability that \(C_1\)-initial base will be reduplicated with \([t_F]\) \(n\) times out of a total of \(m\) trials

\(m\) = number of reduplications for \(C_1\)-initial word

\(n\) = number of reduplications with \([t_F]\) for \(C_1\)-initial word

- Compared weighted similarity (line) to Exp 1 results (bars)
WEIGHTED SNC: CORRELATION

\( r^2 = .855, p < 0.01^* \)
With **4 adjusted feature weights**, the SNC metric can closely model the reduplicative data \( (r^2 = .855) \)

- [voice]: .554
- [distributed]: .400
- [strident]: .249
- [spread glottis]: .198
- All other features have a weight of 0.100
Okay, but have we compromised the model?

Is it no longer a similarity metric, but just a model of the reduplicative data?

Let’s see if our reduplicative data resemble other areas where gradient, lg-specific similarity is arguably relevant:

- Lexical **cooccurrence** (McCarthy 1994)
- Perceptual **confusability** (Shepard 1972)
Similarity of two Cs is often negatively correlated with their cooccurrence within roots (Greenberg 1950)
- English: two LAB or two DOR are underattested in [sCVC]: skip, speak, skim, smack..., *smap, *scog, *spobe, *speam (Fudge 1969)
- Arabic: velars & uvulars rarely cooccur within roots (Frisch et al. 2004)

Hypothesis: the less often a C cooccurs with [t] in a root, the less often it will take [t_F] in its echo RED
- If we see a strong correlation with the reduplicative data, this could be independent support for our weighted model
Similarity of initial $C_1$ and medial [t] is the inverse of their observed / expected lexical cooccurrence: (Frisch et al. 2004)

$$sim(C_1, t) = \frac{\# [C_1 VCV]}{\# [CVCV]} \times \frac{\# [CVtV]}{\# [CVCV]}$$

Examined the cooccurrence of all initial Cs with medial [t] in CVCV roots in a corpus of Bengali (Mallik et al. 1998)

Compared cooccurrence rate (line) to Exp 1 results (bars)
COOCCURRENCE: CORRELATION

$r^2 = .004, p = 0.81$
The lexical cooccurrence model of similarity **fails to predict** the observed \([t_F]\)-avoidance patterns \((r^2 = .004)\)

Possible explanations:
- **Lexical cooccurrence in Bengali involves similarity, but echo reduplication does not** (unlikely, see results)
- **Lexical cooccurrence in Bengali does not involve similarity, while echo reduplication does** (possible)
- **Low \(n\)?** Corpus had 865 CVCV roots; 64 with medial \([t]\)
  - cf. Arabic corpus of 2674 roots (Frisch et al. 2004)
The other area to look for the effects of gradient similarity is in **perceptual confusability**
- Hindi: [ʈ] is misidentified as [ʈ] more than as [ɖ] (Ahmed & Agrawal 1968)

**Hypothesis**: Cs more likely to be (mis)perceived as [ʈ] are also less likely to take [ʈᵢ] in echo RED
- If we see a strong correlation with the reduplicative data, this could be independent support for our weighted model
EXPERIMENT II: SETUP

- **Multiple Forced Choice** (MFC) listening experiment
  - Participants identify the consonant they hear
  - Run in Praat *(Boersma & Weenink 2013)*
  - Sony MDR-V200 headphones connected to laptop
  - Experiments took place in quiet room in participants’ homes

- **25 speakers** of Bengali (13F, 12M)
  - Reported no hearing difficulties
  - Varied dialect background
  - Residents of or visitors to CA
  - Paid $20
EXPERIMENT II: STIMULI

- **54 syllables**
  - Onsets: 27 legal [Ca] syllables (all Cs but [ŋ ŋ])
  - Codas: 27 legal [aC] syllables (all Cs but [dʱ h])

- Produced by adult female speaker
  - Best of several reps was normalized for amplitude

- **Blocked by 3 masking conditions**
  - Multi-talker **babble**
  - Pink **noise**
  - **Quiet** (no added sound)

- **54 syllables x 3 conditions x 3 reps = 486 trials**
EXPERIMENT II: TASK

আপনি যে অক্ষর শেখেছেন, নেটা মাউস দিয়ে ক্রিক করুন।
ক্রিক করার পরে, স্পেস বারটা চাপলে আগানে।

ক খ গ ঘ ঙ
চ ছ জ ঝ
ট ঠ ড ঢ
ত থ দ ধ ন
প ফ ব ভ ম

র ল
শ স হ
ড

স্পেস বারটা চাপুন
EXPERIMENT II: HYPOTHESES

- The C most confused with [t] should be [tʰ]
  - Generalized: aspiration should be the most confusable feature

- Next most confused with [t] should be [d]
  - Generalized: voicing should be the 2nd most confusable feature

- After that should be [ɹ]
  - Generalized: [distributed] and other minor place distinctions should be the 3rd most confusable

- After that should be [s]
  - Generalized: [strident] and other manner-related distinctions should be less confusable than the preceding
- **Onset accuracy:** 92% in quiet, 70% in noise, 60% in babble

### Feature accuracy in onsets

<table>
<thead>
<tr>
<th>Feature</th>
<th>Quiet (%)</th>
<th>Noise (%)</th>
<th>Babble (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>asp</td>
<td>92</td>
<td>70</td>
<td>60</td>
</tr>
<tr>
<td>voi</td>
<td>80</td>
<td>60</td>
<td>50</td>
</tr>
<tr>
<td>MinPl</td>
<td>90</td>
<td>75</td>
<td>65</td>
</tr>
<tr>
<td>cont</td>
<td>95</td>
<td>80</td>
<td>70</td>
</tr>
<tr>
<td>MajPl</td>
<td>92</td>
<td>72</td>
<td>62</td>
</tr>
<tr>
<td>son</td>
<td>85</td>
<td>68</td>
<td>55</td>
</tr>
</tbody>
</table>
**EXPERIMENT II: RESULTS**

- **Coda accuracy**: 66% in quiet, 39% in noise, 34% in babble
Similarity of $C_1$ and $[t]$ as drawn from confusion rate is quantified as follows: (Shepard 1972)

$$sim(C_1, t) = \frac{\# (C_1:t) + \# (t:C_1)}{\# (C_1:C_1) + \# (t:t)}$$

Compared Exp 2 perceptions to Exp 1 productions
- Removed “quiet” condition results (at ceiling)
- Looked at onsets and codas separately
ONSET CONFUSIONS: CORRELATION

$r^2 = 0.572, p < 0.01^*$
CODA CONFUSIONS: CORRELATION

$\% [t_F] \text{ use in RED}$

$r^2 = .795, p < 0.01^*$

Base-initial consonant
Consonant confusions with \[t\] in coda position are well correlated with the reduplicative results \((r^2 = .795)\).

But! Echo reduplication involves judging the similarity of onsets; why does the reduplicative data more closely resemble coda confusion?
- Onset confusions with \[t\] were overall rare.
- Acoustic cues are perceptually less salient in codas (Wright 2004), so this is where similarity (not just identity) is likely more often relevant.
SYNTHESIS OF RESULTS

- Okay, we need a recap.

- What did we do again?
  - Task 1: examine **fixed segment choice** in echo reduplication
  - Task 2: establish that fixed segment choice is **predicted by SNC**
  - Task 3: improve the SNC in a thought experiment with **weights**
  - Task 4: find no correlation with **lexical statistics**
  - Task 5: find significant correlation with **coda confusions**
The current study demonstrates that fixed segment choice in Bengali echo reduplication is **highly variable**.

I argue that the choice of fixed segment involves a systematic avoidance of **similarity**, because:
- The patterns are (partially) predicted by the **SNC metric**.
- The patterns correlate with **confusion rates** (in codas).
The patterns clearly show that this similarity is **gradient**.

Echo reduplication is one of many phenomena previously treated as categorical but more recently seen as gradient:
- e.g. vowel harmony in Hungarian (Hayes & Londe 2006)
The current study proposes a modified version of the SNC metric of similarity

*I propose feature weighting* for lg-specific application in diverse phonological phenomena

The study also provides an interesting case in which the SNC metric can measure similarity in phonological phenomena *other than* lexical cooccurrence effects
REMAINING QUESTIONS

- Is Bengali echo reduplication a special case, or should we look for gradient similarity in many more lgs?

- Why are the lexical cooccurrence effects of Bengali so different from the reduplicative results?

- How does this change as speakers deal with multiple phoneme inventories, e.g. bilinguals?
This study was supported in part by Reed College’s Stillman Drake Fund.

Many thanks to my participants and stimulus producers, to Colin Wilson (JHU), Kie Ross Zuraw (UCLA), Marc Garellek (UCSD), and Megha Sundara (UCLA), and to the audience here at the University of Oregon!

অসংখ্য ধন্যবাদ!
[ওঁoŋkʰo oʊoːnːobaŋ]
Ahmed, Rais; Agrawal, S. S. 1968. Significant features in the perception of (Hindi) consonants. JASA 45(3).

Alderete, John; Beckman, Jill; Benua, Laura; Gnanadesikan, Amalia; McCarthy, John; Urbanczyk, Suzanne. 1999. Reduplication with fixed segmentism. Linguistic Inquiry 30(3), 327-364.


Cutler, Ann; Weber, Andrea; Smits, Roel; Cooper, Nicole. 2004. Patterns of English phoneme confusions by native and non-native listeners. JASA 116(6), 3668-3678.


Wright, Richard. 2004. A review of perceptual cues and cue robustness. In Hayes, Bruce; Kirchner, Robert; Steriade, Donca (eds.) *Phonetically Based Phonology*. Cambridge.