What echo reduplication reveals about phonological similarity

Sameer ud Dowla Khan, Brown University
Thursday 20 April 2012
Reed College
Outline

- **Background: echo reduplication**
  - Properties of echo reduplication crosslinguistically

- **Experiment: identity and similarity**
  - Similarity avoidance in Bengali echo reduplication

- **Analysis: measurement of similarity**
  - Measurement of consonant similarity in Bengali

- **General discussion**
  - Summary and further questions
Background

Properties of echo reduplication crosslinguistically
Echo reduplication

- **Subtractive reduplication**
  - Bengali
    - [goli] ‘alley’
    - [oli goli] ‘alleys, etc.’

- **Fixed-segment (S_F) reduplication**
  - Bengali
    - [kaʃi] ‘cough’
    - [kaʃi t_Faʃi] ‘cough, etc.’
  - English
    - [kɔf] ‘cough’
    - [kɔf ʃm_Fɔf] ‘coughˌDISMISSIVE’
Echo reduplication

- **Subtractive reduplication**
  
  Bengali
  
  [goli] ‘alley’
  [oli goli] ‘alleys, etc.’

- **Fixed-segment (\(S_F\)) reduplication**
  
  Bengali
  
  [kaʃi] ‘cough’
  [kaʃi t\(_F\)aʃi] ‘cough, etc.’

  English
  
  [kɔf] ‘cough’
  [kɔf \(f_{mf}f\)] ‘cough\(_\text{DISMISSIVE}\)’
Fixed-segment reduplication

- In FSR, **fixed material** \( S_F \) associated with a particular construction is found in the R instead of a copy of B material.

- The fixed material can be:
  - A consonant (most common)
  - A vowel
  - A CV sequence
  - A stem
Fixed-segment reduplication

- **Consonantal $S_F$**
  
  *Kashmiri (Koul):* $S_F = [v_F]^1$
  
  [nalki] ‘faucet’
  
  [nalki $v_F$alki] ‘faucet, etc.’

- **Vocalic $S_F$**
  
  *A-Hmao (Mortensen 2005):* $S_F = [í_F]$
  
  [and[pʰɔu] ‘mouth’
  
  [ánd[pʰí $F$ánd[pʰɔu] ‘cheeks, nose, etc.’

---

$^1$ IPA: Koul & Wali (2006)
Fixed-segment reduplication

- **[CV] $S_F**
  - *Tamil (Keane 2001):* $S_F = [ki_F]^1$
  - $[\text{ve}l\text{ai}]$ ‘white’
  - $[\text{ve}l\text{ai }ki_F\text{ai}]$ ‘white, etc.’

- **Stem $S_F$**
  - *Russian (Podobryaev 2012):* $S_F = [xuj_F] < \text{‘penis’}$
  - $[\text{mál}t\text{čik}]$ ‘boy’
  - $[\text{mál}t\text{čik }xuj_F\text{á}l\text{t}\text{čik}]$ ‘boy_{DISMISSIVE}’

1 IPA: Keane (2004)
Languages reported to have echo reduplication
Echo reduplication

- Typically conveys **generalization**
  - ‘X, etc.’
  - ‘X and associated things’
  - ‘X in general’
  - ‘superset of which X is a member’

- In some lgs, it conveys a **dismissive** tone
  - Russian: [málʲtɕik xuʃálʲtɕik] ‘boy$_{DISMISSIVE}$’
  - English: [daktəʃmæktə] ‘doctor$_{DISMISSIVE}$’

Obligatory BR-nonidentity

- Most salient **phonological property** of echo reduplication is **obligatory BR-nonidentity**
  - ≥1 phonological difference between B and R
- **Presence of $S_F$** in R usually enough to **generate BR-nonidentity**
  
  *Kashmiri (Koul)*
  
  $[nalki] \Rightarrow [nalki \ v_Falki]$ ‘faucet, etc.’

- **But what if it isn’t?**
  
  *Kashmiri (Koul)*
  
  $[va:zi] \Rightarrow ??[va:zi \ v_Fa:zi]??$ ‘cook, etc.’
Obligatory BR-nonidentity

- Lgs avoid such cases of potential BR-identity by either:
  - Using a designated backup $S_F$
  - Choosing from among the other $S_F$ options
  - Modifying the B instead of in R
  - Deeming the phrase ineffable
Obligatory BR-nonidentity

- Many lgs have a backup $S_F$, kept on reserve for cases of BR-identity

  Abkhaz (Vaux 1996): $S_F = [m_F] (\Rightarrow [tʃ_F])$
  
  /gáʤak’/ $\Rightarrow [gáʤak’ m_Fáʤak’]$ ‘fool, etc.’

  /tʃək’/ $\Rightarrow [tʃək’ m_Fək’]$ ‘horse, etc.’

  /maát/ $\Rightarrow *[maát m_Faát-] \Rightarrow [maát tʃ_Faát-]$ ‘money, etc.’
Obligatory BR-nonidentity

- Other lgs have **multiple $S_F$ options**, always choosing one that avoids BR-identity

  *Farsi (Ghaniabadi et al. 2006):* $S_F = [m_F] \sim [p_F]$

  /tærøzu/ $\Rightarrow [tærøzu m_Færøzu] \sim [tærøzu p_Færøzu]$ ‘scale, etc.’

  /zærif/ $\Rightarrow [zærif m_Færif] \sim [zærif p_Færif]$ ‘slender, etc.’

  /mive/ $\Rightarrow *[mive m_Five] \sim [mive p_Five]$ ‘fruit, etc.’

  /pir/ $\Rightarrow [pir m_Fir] \sim *[pir p_Fir]$ ‘old, etc.’
Obligatory BR-nonidentity

- Some lgs even go so far as to **modify B** when R with $S_F$ would be identical to it.

  *Classical Tibetan (Beyer 1992):* $S_F = [a_F] \ (\Rightarrow \ B \ [o_F])$

  /ndzog/ $\Rightarrow \ [ndz_{a_F}g \ ndzog]$ ‘jumbled up’

  /glen/ $\Rightarrow \ [gl_{a_F}n \ glen]$ ‘very stupid’

  /ŋan/ $\Rightarrow \ *[ŋ_{a_F}n \ ŋan] \Rightarrow \ [ŋan \ ŋo_F.n]$ ‘miserable’
Lastly, some lgs simply deem cases of echo BR-identity to be **ineffable**

*Turkish (Swift 1963):* \( S_F = [m_F] \)

/\hau\u011fu/ \( \Rightarrow [\hau\u011fu \, m_F \hau\u011fu] \) ‘towel, etc.’

/citap/ \( \Rightarrow [\text{citap} \, m_F \text{itap}] \) ‘book, etc.’

/masa/ \( \Rightarrow *[\text{masa} \, m_F \text{asa}] \) ‘table, etc.’ \( \Rightarrow \) NO OUTPUT

\(^1\) IPA: Zimmer & Orgun (1999)
Obligatory BR-nonidentity

- Crosslinguistically, **BR-identity in echo reduplication is ungrammatical**
- Trivedi’s (1990) survey of FSR in ~100 Indian lgs found **obligatory BR-nonidentity in every lg**
- Seems clear... but I still have one question: How sensitive is BR-nonidentity?
Survey

- For example, let’s consider English

  $S_F = [ʃm_F]$

  `/daktə/ ‘doctor’ $\Rightarrow [daktə \text{ʃ}m_F aktə]$

  ‘doctor_{DISMISSIVE}’

  `/skul/ ‘school’ $\Rightarrow ?$

  `/smuð/ ‘smooth’ $\Rightarrow ?$

  `/ʃmuz/ ‘schmooze’ $\Rightarrow ?$

  `/ʃmælts/ ‘schmaltz’ $\Rightarrow ?$

  `/ʃnæz/ ‘schnozz’ $\Rightarrow ?$`
In Nevins & Vaux (2003), 95% of speakers avoided [ʃm_F] in R of [ʃmuz] ‘schmooze’
- *[ʃmuz ʃm_F uz] due to BR-nonidentity
- Fits with cross-linguistic pattern

Interestingly, 30% of speakers also avoided [ʃm_F] in R of [ʃnaz] ‘schnozz’
- *[ʃnaz ʃm_F az]...but why?
- BR-nonidentity generalized to BR-dissimilarity
Curiosity from literature

- What does this mean?
  - Explanation 1:
    - 30% of speakers have a **BR-dissimilarity constraint**
    - 65% have a **BR-nonidentity constraint**
  - Explanation 2:
    - 95% have a **BR-dissimilarity constraint**, of which:
      - 30% felt \([\text{n}]\) and \([\text{m}_F]\) are **too similar**
      - 65% felt \([\text{n}]\) and \([\text{m}_F]\) are **sufficiently dissimilar**
Is this English-specific?

- Or maybe we’re assuming too much from this one data point...
- Is English echo reduplication a weird case?
  - Not as common as in other languages
  - [ʃm]-reduplication is somewhat humorous
  - [ʃm] and [ʃn] are highly marked in English
- Maybe this is just a weird fact of English...
Motivation for an experiment

- To find out if echo reduplication involves BR-nonidentity or BR-dissimilarity...
- We need to study a lg in which:
  - Echo reduplication is a **fully productive, linguistic feature**
  - $S_F$ isn’t such a marked sound
Experiment

What echo reduplication reveals about phonological similarity
Experiment: question

- **Question:** how sensitive is BR-assessment?
  - **Only sensitive** to exact BR-identity
    - Any BR-difference should suffice
  - **Also sensitive** to relative BR-similarity
    - Some BR-differences aren’t dissimilar enough
Experiment: language

- Test case: **Bengali** echo reduplication
  - Default $S_F: [t_F]$
  - Backup $S_F: [m_F] [f_F] [p_F] [u_F]...$

- Why Bengali?
  - Echo reduplication is a very **common feature**
  - Default $[t_F]$ is a relatively **unmarked sound**
  - Many **contrastive but phonetically similar** phonemes: $[t^h] [d] [t] [t^h] [tɕ]...$
Experiment: basic idea

- So we know that a word like [bʰidʃə] ‘having gotten wet’ $\Rightarrow$ [bʰidʃə t_Fidʃə]...
- ...and that a word like [tika] ‘vaccine’ $\Rightarrow$ *[tika t_Fika] $\Rightarrow$ [tika m_Fika]...
- ...but what about a word like [tʰajʃə] ‘having stuffed’?
  - Will it act like [bʰidʃə]? [tʰajʃə t_Fajʃə]
  - Or like [tika]? *[tʰajʃə t_Fajʃə] $\Rightarrow$ [tʰajʃə m_Fajʃə]
Experiment: subjects and procedure

- **Production experiment** with native speaker adults (n=30)
- Heard audio recording of a word
  - Order was randomized for each speaker
- Asked to **produce echo reduplicated form**
  - Did speaker use default $[t_F]$?
  - Or did he/she use a backup $S_F$?
Experiment: stimuli

- 60 test words fell under **three conditions**:
  - **Identity**: [t]-initial words
  - **Similarity**: words with [t]-like initials
    - Coronal obstruents: [tʰ] [d] [t] [tʰ] [tʃ] [s]~[tʃʰ] [ʃ]
  - **Control**: words with **non-[t]-like initials**
    - Coronal sonorants: [n] [l] [ɾ]
    - Non-coronals: [k] [h] [p] [f] [bʰ] [m]
# Experiment: stimuli

## Bengali consonant inventory (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>p b b̄</td>
<td>t t̊ d d̄</td>
<td>t t̊ d d̄</td>
<td></td>
<td>k k̊ g ḡ</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 l̊</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: stimuli

### Bengali consonant inventory (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>p b b̥</td>
<td>t t̥ d d̥</td>
<td>t t̥ d d̥</td>
<td></td>
<td>k k̥ g g̥</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>j</td>
<td>h</td>
</tr>
<tr>
<td><strong>Liquid</strong></td>
<td></td>
<td></td>
<td>ɻ 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: hypothesis 1

- Hypothesis 1: BR-assessment is **only identity-sensitive**
  - **Identity** words will **never** use $[t_F]$
  - **Similarity** words will behave like **Control** words
  - **Control** words will **always** use $[t_F]$

- **Identity $\neq$ Similarity = Control**
  
  $*[t...t_F] \neq [t^h...t_F] = [b^i...t_F]$
Experiment: hypothesis 2

○ Hypothesis 2: BR-assessment is **sensitive to phonetic similarity** across phonemes
  - **Identity** words will **never** use $[t_F]$
  - **Similarity** words will behave like **Identity** words
  - **Control** words will **always** use $[t_F]$

○ **Identity** = **Similarity** ≠ **Control**

  $*[t...t_F] \neq *[t^h...t_F] = [b^a...t_F]$
Experiment: hypothesis 3

- Hypothesis 3: BR-assessment is **strongest in cases of identity**, but also **sensitive to phonetic similarity** across phonemes
  - **Identity** words will **never** use \[t_F\]
  - **Similarity** words will **sometimes** use \[t_F\]
  - **Control** words will **always** use \[t_F\]

- **Identity ≠ Similarity ≠ Control**
  - \*[t...t_F] ≠ ?[t^h...t_F] = [b^f...t_F]
Results

- The results were surprising:
- While Hypothesis 3 came closest, none of the hypotheses was borne out!
  - Identity ≠ Similarity = Control
  - Identity = Similarity ≠ Control
  - Identity ≠ Similarity ≠ Control
- Lots of variation across speakers and words
- But one thing was clear:
Results: general pattern

- BR-assessment is **sensitive to similarity**
- But, this sensitivity is **gradient**
- The **more similar** a consonant is to [t], the **less likely** it is to be replaced by [t₉]

[平原] [平原ʰ] [平原] [平原] ... [平原ʰ] [平原] [平原] [平原] [平原]

Least likely to use [t₉]  Most likely to use [t₉]
Results: grouped by initial consonant

% $[t_F]$ use

B-initial consonant
Results: significant differences

% [t_F] use

B-initial consonant

- t
- tʰ
- d
- ɾ
- s
- sʰ
- k
- ɾ
- tʃ
- d
- ʃ
- n
- l
- bʱ
- f
- p
- m
Results: design issues?

- Was there a problem with the setup?
- Should the similarity condition and control condition be redefined?
Results: clustering

B-initial consonant
Results: gradient similarity

- No, there is **no clustering of consonants** into two or three categories
- Furthermore, **heavy overlap** across the clusters that are found
- Clearly, **similarity is gradient**
New question

- It’s clear some measurement of similarity is needed in Bengali echo reduplication
- So then how do speakers calculate the similarity of a pair of sounds?
Analysis

Measurement of consonant similarity in Bengali
Models of similarity

- Phonological similarity has been measured in different ways in the literature
- Most metrics incorporate:
  - Shared natural classes
  - Correlation with lexical cooccurrence
- Can either of these model Bengali speakers’ notions of similarity?
Shared natural classes: introduction

- One method\(^1\) of calculating similarity is by comparing the \textbf{number of natural classes} of which two sounds are members
  - \([t]\) and \([n]\) share \([+\text{cor}]\) but not \([-\text{voi}]\) or \([-\text{son}]\)
  - \([t]\) and \([p]\) share \([-\text{son}]\) and \([-\text{voi}]\) but not \([+\text{cor}]\)
  - \([t]\) and \([v]\) share \([-\text{son}]\) but not \([-\text{voi}]\) or \([+\text{cor}]\)
- The \textbf{more similar} two consonants are, the \textbf{more natural classes} they will share

\(^1\) Frisch, Pierrehumbert, & Broe (2004)
Shared natural classes: introduction

- This measure takes **lg-specific details** into account, due to different inventories
  - Aspiration is contrastive in Bengali, not English
  - [b] and [d] share [-asp] in Bengali, not English
- Can we apply the SNC metric to the echo reduplication results in Bengali?
In a model of similarity based on **shared natural classes**...

...the similarity of a consonant $C_1$ to [t] can be calculated as follows:

$$sim(C_1, t) = \frac{\# \text{ shared natural classes}}{\# \text{ shared natural classes} + \# \text{ non-shared natural classes}}$$
SNC predictions (line) vs. observed (bars)

$r^2 = .584, p < 0.01^*$
SNC metric: discussion

- SNC metric does fairly well ($r^2 = .584$)
- However, where it doesn’t do well is the most crucial area: coronal obstruents
  - How can we adjust this model to reflect that [t] is more similar to [tʰ] than to [t]?
  - Is there a way to designate certain features as being more important than others?
What if we incorporated different weights for different features, reflecting their importance in similarity measurement?

- Weighting [distributed] over [aspiration] will make ([t], [t]) more different than ([t], [t^h])

What would such a metric look like?
Similarity equation

- In a model of similarity based on **shared weighted features**...
- ...the similarity of a consonant $C_1$ to $[t]$ can be calculated as follows

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

$w_i = \text{weight of the feature } f_i$

$\delta_i(C_1, t) = 1$ (feature value shared) or 0 (not shared)
Where do these weights come from?

- So how do we determine $w_i$?
- Maybe *lexical statistics*?
Many studies\(^1\) claim that the **lexicon of a \lg** reflects notions of similarity.

The **more similar** two consonants are, the **less often they will cooccur** within roots.

- Words like \([\text{f}\text{ʌ}\text{ʤ}]\) and \([\text{p}\text{ɛ}\text{ɡ}]\) are common.
- Words like \(*[\text{ʃ}\text{ʌ}\text{ʤ}]\) and \(*[\text{p}\text{ɛ}\text{b}]\) are underattested.

Can we apply this to the Bengali data?

\(^1\) McCarthy (1994) and many others.
Lexical cooccurrence: metric

- We can turn this around:
- Sound pairs that are underattested within roots must be perceived as more similar
- Thus, **lexical statistics** can be converted into a **similarity score**
Lexical cooccurrence: metric

- In a model of similarity based on **lexical cooccurrence statistics**...
- ...the similarity of a consonant $C_1$ to [t] can be calculated as follows

$$sim(C_1, t) = \frac{\text{obs}[C_1 VCV]}{\text{all roots}} \times \frac{\text{obs}[CVtV]}{\text{all roots}}$$
Lexical cooccurrence: data

- Used a **Bengali corpus** (Mallik et al. 1998) to examine roots where [t] **cooccurs with a consonant (C)**
- Plugged in the numbers to get a **similarity score** for each C paired with [t]
- Compared those similarity scores to the \([t_F]\)-use patterns from my experiment
Lexical cooccurrence: results

$r^2 = .004, p = 0.81$

% $[t_F]$ use

B-initial consonant
Lexical cooccurrence: discussion

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

- It appears that the **Bengali lexicon does not reflect the notions of similarity** at work in the productive grammar

- Thus, we cannot use lexical statistics to adjust our natural classes model
Can weights be used at all?

- So how else can we determine what the weights should be?
- Can weights help us at all?
- Let’s see if we can use the variation in the data itself to determine the weights...
- ...and then worry about where the weights are coming from at some other time
Probability equation

- Probability of $[t_F]$-use in the echo R of a $C_1$-initial B can be calculated as follows

$$P = \frac{(m!)}{(n!(m-n)!)} (1-sim(C_1, t))^n (sim(C_1, t))^{m-n}$$

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
Feature weighting (line) vs. observed (bars)

$r^2 = .855, p < 0.01^*$
Feature weighting: discussion

- A model of similarity that takes **feature weights** into account can closely model the data ($r^2 = .855$)
- Of course, in our case, we used the data to determine the weights

*I’ll talk about some ideas of where this could independently come from in a minute...*
General discussion

Summary and further questions
Crosslinguistically, B and R in echo reduplication must be sufficiently different

In most lgs/studies, this is taken to be a **categorical nonidentity constraint**
- “B and R must be non-identical”
- Assumes that sound pairs can be categorized as either “identical” or “non-identical”
- Even one BR feature mismatch should suffice
Summary

- Data from English show that the constraint is actually **sensitive to phonetic similarity**, not just identity
  - “B and R must be **dissimilar**”
  - Still assumes categorical grouping of sound pairs: “identical”, “similar”, and “dissimilar”
Summary

- Experimental data from Bengali confirm that **BR-assessment is sensitive to phonetic similarity**, not just identity.
- The data also show that in fact, **similarity is gradient**
  - Cannot group sound pairs categorically as “identical”, “similar”, and “dissimilar”
Summary

- Speakers compute the **similarity score of two sounds** using a metric that takes different **feature weights** into account.
- We can derive the feature weights from the pattern itself...
- ...but where do speakers actually get these weights from independently?
Summary

- Weights do not come from the lexicon
  - Similarity in echo reduplication is *not correlated with lexical patterns*
Further questions

- Alternatively, feature weights could come directly from the phoneme inventory
  - The features that are weighted heavily are:
    - [voice]: 0.554
    - [distributed] (=dental vs. alveolar): 0.400
    - [strident]: 0.249
    - [spread glottis] (=aspiration): 0.198
  - All others are weighted 0.1
Further questions

- These are also the features that help make the Bengali phoneme inventory so coronal-heavy.
- In fact, of all features, these make the most phonemic contrasts on their own in the lg.
- Thus, there might be independent evidence of their “weight”.
- Need to do more work to confirm this.
Further questions

- How much of this is **language-specific**?
  - Coronal-heavy inventory?
- How much is **universal**?
  - Feature inventory?
- Future studies on **similarity-sensitive phenomena in other lgs** will take on these questions to build this model of similarity
Thank you!

Special thanks to Kie Ross Zuraw, Colin Wilson, Bruce Hayes, Farida Amin Khan, the 30 participants in my experiment, and everyone in attendance here.