Are universal markedness hierarchies learnable from the native lexicon? The case of gemination in Hungarian

Lilla Magyar

(MIT)

WCCFL 33 (28 April, 2015)
1. Introduction
   - Gemination in loanwords
   - Gemination in the native phonology
   - Questions and hypotheses
2. Testing Hypothesis 1
   - Wug test
3. Testing Hypothesis 2
   - Corpus study
4. Testing Hypothesis 3
   - The model
   - Simulation 1
   - Simulation 2
   - Simulation 3
   - A note on failed simulations
   - Simulation 4
5. Conclusions
Loan gemination cross-linguistically

Gemination is a cross-linguistically widespread phenomenon:

- Japanese: [kat:o] ‘cut’ (Kubozono (2008))
- Italian: [fan:] ‘fan’ (Passino (2004))
- Kannada: [kap:u] ‘cup’ (Sridhar (1990))
- Telugu: [ro:qːu] ‘road’ (Krishnamurti & Gwynn (1985))
Loan gemination cross-linguistically

General characteristics of loan gemination:

- a singleton consonant in the source word is geminated in the loanword, even if the doubling does not have an orthographic reflex in the source word
- source languages do not allow phonetic geminates
- in Japanese, Kannada, Telugu and Finnish, the final consonant of the source word is geminated, and a vowel is inserted after the final consonant
- in Italian and Hungarian, the final consonant of monosyllabic words is geminated
a singleton consonant in words borrowed from English, German (and occasionally, from French) often undergo gemination

loan gemination without orthographic reflex in the source word is most common in word-final position in monosyllables

gemination depends on consonant class as well

voiceless obstruents are geminated more often than other consonants

nasals are geminated more often than liquids

voiced fricatives are never geminated when there is no orthographic geminate in the source word
Gemination in the native Hungarian phonology

- 25 consonants
- All consonants have a long version
- Seemingly no restriction on gemination
- Both singletons and geminates can occur in intervocalic and word-final position
- Some consonants may be geminated only across morpheme boundaries
- The frequency distribution of singletons and geminates in the native phonology is not clear
### Gemination in the native Hungarian phonology

<table>
<thead>
<tr>
<th>Short</th>
<th>Long</th>
</tr>
</thead>
<tbody>
<tr>
<td>b [b]</td>
<td>bb [b:] jobb ‘better’</td>
</tr>
<tr>
<td>c [ts]</td>
<td>cc [t:s] koccan ‘crush’</td>
</tr>
<tr>
<td>cs [tʃ]</td>
<td>ccs [tʃ:] loccsan ‘splash’</td>
</tr>
<tr>
<td>d [d]</td>
<td>dd [d:] tedd ‘do-2nd.p.sing.-imperative’</td>
</tr>
<tr>
<td>f [f]</td>
<td>ff [f:] puffad ‘bloat up’</td>
</tr>
<tr>
<td>g [g]</td>
<td>gg [g:] agg ‘very old’</td>
</tr>
<tr>
<td>k [k]</td>
<td>kk [k:] rokkant ‘handicapped’</td>
</tr>
<tr>
<td>l [l]</td>
<td>ll [l:] hall ‘hear’</td>
</tr>
<tr>
<td>m [m]</td>
<td>mm [m:] kammog ‘trudge’</td>
</tr>
<tr>
<td>n [n]</td>
<td>n [n:] fenn ‘up (above)’</td>
</tr>
<tr>
<td>p [p]</td>
<td>pp [p:] koppan ‘make a knocking sound’</td>
</tr>
<tr>
<td>r [r]</td>
<td>rr [r:] arra ‘that way’</td>
</tr>
<tr>
<td>s [ʃ]</td>
<td>ss [ʃ:] lassan ‘slowly’</td>
</tr>
<tr>
<td>sz [ʒ]</td>
<td>ssz [ʒ:] osszon ‘divide-3rd.p.sing.-imperative’</td>
</tr>
<tr>
<td>t [t]</td>
<td>tt [t:] vett ‘buy-3rd.p.sing.-past’</td>
</tr>
<tr>
<td>v [v]</td>
<td>vv [v:] savval ‘acid-instrumental’</td>
</tr>
<tr>
<td>z [z]</td>
<td>zz [z:] hozz ‘bring-2nd-sing-imperative’</td>
</tr>
</tbody>
</table>
Gemination cross-linguistically

Implicational hierarchies and universal rankings (Podesva (2002), Steriade (2004)):

- geminate $\rightarrow$ geminate voiceless obstruent
- geminate fricative $\rightarrow$ geminate stop
- geminate sonorant $\rightarrow$ geminate sonorant of lower sonority
- *YY (glides) $\gg\gg$ *LL (liquids) $\gg\gg$ *NN (nasals)
- it lines up with patterns of gemination in Hungarian loanwords
since gemination in loanwords seems to reflect patterns of universal geminate markedness, it looks like native speakers of Hungarian have some awareness of such hierarchies

if all consonants can be geminated in the native phonology, where does this knowledge come from?

can it come from the native phonology (if we take a closer look at it)?

if so, is it learnable based on phonotactic generalisations?
Hypotheses

The following hypotheses will be tested:

**Hypothesis 1**: Even native speakers of Hungarian (a language which allows all kinds of geminates) have some awareness of universal geminate markedness.

**Hypothesis 2**: This knowledge comes from the native lexicon: the frequency distribution of geminates in the native phonology reflects patterns of universal markedness.

**Hypothesis 3**: These patterns can be learned from the native Hungarian lexicon based on phonotactic generalisations.
Testing material

- the test contains 236 words (118 word pairs)
- all words are nonce monosyllables ending in a short vowel + short consonant or geminate sequence
- word pairs are minimal pairs: a word ending in a particular short vowel + short consonant and another word ending in the same vowel + the geminated form of the same consonant (e.g. mok [mok] - mokk [mokː])
- filler items include nonce word pairs - one with a short and one with a long vowel
Participants and task

- 115 native speakers of Hungarian participated in the experiment
- the test was administered online and participants remained anonymous
- participants were presented with a list of target and filler items (all in word pairs)
- items were presented in written form (representing pronunciation)
- they were asked to decide which of the two nonce words they considered more well-formed as a native Hungarian word or a Hungarianised loanword
The summary of results averaged for consonant class (in descending order with respect to geminate proportion):

<table>
<thead>
<tr>
<th>Consonant class</th>
<th>Singleton %</th>
<th>Geminate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiceless affricates</td>
<td>44%</td>
<td>56%</td>
</tr>
<tr>
<td>Voiceless stops</td>
<td>47%</td>
<td>53%</td>
</tr>
<tr>
<td>Voiceless fricatives</td>
<td>47%</td>
<td>53%</td>
</tr>
<tr>
<td>Liquids</td>
<td>48%</td>
<td>52%</td>
</tr>
<tr>
<td>Nasals</td>
<td>61%</td>
<td>39%</td>
</tr>
<tr>
<td>Voiced stops</td>
<td>65%</td>
<td>35%</td>
</tr>
<tr>
<td>Voiced fricatives</td>
<td>85%</td>
<td>15%</td>
</tr>
</tbody>
</table>
Monosyllables containing a short vowel followed by a singleton or a geminate consonant extracted from the Hungarian Webcorpus (Halácsy et al. (2004)):

<table>
<thead>
<tr>
<th>Consonant class</th>
<th>Singleton %</th>
<th>Geminate %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Voiceless affricates</td>
<td>35%</td>
<td>65%</td>
</tr>
<tr>
<td>Voiceless stops</td>
<td>51%</td>
<td>49%</td>
</tr>
<tr>
<td>Voiceless fricatives</td>
<td>54%</td>
<td>46%</td>
</tr>
<tr>
<td>Voiced stops</td>
<td>65%</td>
<td>35%</td>
</tr>
<tr>
<td>Nasals</td>
<td>78%</td>
<td>22%</td>
</tr>
<tr>
<td>Liquids</td>
<td>72%</td>
<td>28%</td>
</tr>
<tr>
<td>Voiced fricatives</td>
<td>92%</td>
<td>8%</td>
</tr>
</tbody>
</table>
Comparison to wug test results

Strong positive correlation with wug test results ($r=0.85$):
Comparison to wug test results

- both reflect universal tendencies in some ways
- voiceless obstruent geminates are more common than other geminates
- nasal geminates are somewhat more common in the corpus than in the wug test
- liquid geminates are unusually highly rated in the wug test
- external factors in speakers’ judgements (nonce words resemble existing words, word endings)
- exiguous amount of monosyllables with relevant parameters in the corpus
The model

- implemented in the UCLA phonotactic learner (Hayes & Wilson (2008))
- based on the principle of Maximum Entropy
- trained on a set of data (extracted from the corpus)
- comes up with weighted constraints based on phonotactic generalisations
- assigns scores (sums of violations) to forms listed as the testing data
Settings

- full monosyllables from the Hungarian Webcorpus (type frequencies) used as training data
- tested on rhymes (short vowel and short consonant or geminate sequences)
- scores assigned by the learner are compared to wug test judgements (how much a word ending in a certain short vowel + short consonant sequence is (dis)preferred over a word ending in the same vowel + same geminate consonant sequence - e.g. [op] vs. [op:])**
Scores assigned by the learner

How the learner assigns scores:

- the learner assigns a score to each form
- the score is the sum of violations of all constraints by each form
- the higher the score, the lower the probability of a given form
Score conversion

Global probability:

- probability of a form given all the other forms

\[ P(\text{ibb}) = \frac{\exp(-h(\text{ibb}))}{\exp(-h(\text{ibb})) + \exp(-h(\text{opp})) + \exp(-h(\text{et})) + \ldots} \]

\[ P(\text{ibb}) + P(\text{ib}) + P(\text{opp}) + P(\text{et}) + \ldots = 1 \]

- for a better comparison with wug test results, scores will be converted to local probabilities
Score conversion

Local probability:

- probability of a form given another form (which it is compared to)
- for example, the probability of a word ending in [ib:] compared to [ib]

\[
P(ibb) = \frac{\exp(-h(ib))}{\exp(-h(ibb)) + \exp(-h(ib))}
\]

\[
P(ibb) + P(ib) = 1
\]
Results

Very weak positive correlation ($r=0.25$)
the model is trained on all the rhymes found in the corpus

it is tested on final singletons and geminates (e.g. [b] vs. [b:], [t] vs. [t:]

predicted local probabilities of words ending in certain singletons and geminates are tested against native speaker preferences (how much a singleton consonant was preferred over the geminated form of the same consonant - e.g. all words ending in [b] vs. all words ending in [b:])
Results

Very weak positive correlation \( (r=0.14) \)
Settings

- the learner is trained on final consonants extracted from the corpus (preceding material is removed)
- tested on word-final consonants (data given as [b], [b:], [p], [p:] etc.)
- tested against native speakers’ preferences (just like in the previous simulation)
Results

Weak positive correlation (r=0.31)
Reasons for failure

Problems with the model:

1. if there are a lot of features, some restrictions have to be included in the settings, otherwise learning will not terminate in reasonable time.
2. it needs to come up with more constraints to capture the relevant generalisations.
3. however, this may not be the case, since the model was able to replicate corpus frequencies (almost perfectly).
Reasons for failure

Other issues:

- since there is only a small amount of monosyllables in the native Hungarian lexicon and there are many (accidental) gaps, people are not able to replicate fine-grained details (e.g. different vowel + consonant combinations, differences in place of articulation) in their judgements on nonce words
- instead, they make broader generalisations, that is, they have intuitions about which consonant classes are more common as geminates
- therefore a new, more restricted learning model is needed
the learner was tested on final consonants represented as consonant classes
other parts of words than the final consonant were cut and word-final consonants (both singletons and geminates) were organised into a list
the data were collapsed across consonant classes: for example, [p], [t] and [k] were transcribed as T or T:, a general segment representing singleton and geminate voiceless stops
the learner was tested on consonants collapsed into consonant classes in word final position
predictions by the learner and native speaker judgements on (dis)preferences for singletons over geminates by each consonant were compared
**Constraints**

* [+del.rel,-long]: *affricates (1.44)  
* [-son,+voice,+cont,+long]: *long voiced fricatives (1.375)  
* [+nasal,+long]: *long nasals (1.14)  
* [+voice,+cont,+long]: *long voiced sonorants and sibilants (1.1)  
* [-son,+voice,+cont,-long]: *short voiced obstruents (0.979)  
* [+del.rel,+long]: *long affricates (0.831)  
* [-son,+voice,+long]: *long voiced obstruents (0.764)  
* [-son,+cont,+long]: *long sibilants (0.695)  
* [-son,+cont,-long]: *short sibilants (0.54)  
* [+voice,+cont,-long]: *short voiced sibilants (0.164)
Results

Strong positive correlation ($r=0.86$)
A model using handcrafted markedness constraints

- in Simulation 4, the learner created constraints which are like geminate markedness constraints
- let us see how a model with handcrafted markedness constraints would predict the probability of singletons and geminates by each consonant class
- implemented using the Maxent Grammar Tool (Hayes & Wilson (2008))
Settings

- OT tableaus and constraints were used as input data
- each tableau contains a singleton and a geminated version of a segment representing a consonant class, violation marks for constraints and observed probabilities for each possible output candidate
Constraints

*ZZ: *voiced fricatives geminates
*SS: *voiceless fricatives geminates
*TT: *voiceless stop geminates
*DD: *voiced stop geminates
*T:S: *voiceless affricate geminates
*NN: *nasal geminates
*LL: *liquid geminates
**Constraints**

*Z*: *singleton voiced fricatives
*S*: *singleton voiceless fricatives
*T*: *singleton voiceless stops
*D*: *singleton voiced stops
*TS*: *singleton voiceless affricates
*N*: *singleton nasals
*L*: *singleton liquids
Results

Strong positive correlation \((r=0.86)\) (same results as Simulation 4):
Conclusions

Hypothesis 1

- native speakers’ geminate vs. singleton preferences in the wug test are similar to what we find in loanwords, the native Hungarian lexicon, and also, across languages
- there are some differences between distribution of singletons and geminates in the corpus and in nonce word judgements, which might be due to other factors
- many nonce words looked like famous brand names or ends of common polysyllabic words - as reported by participants
- since there are relatively few monosyllabic words with the relevant parameters (ending in short vowel + short / long consonant), external factors like these can influence subjects
Hypothesis 2

- the distribution of geminates and singletons in the corpus lines up with universal geminate markedness hierarchies
- native speakers’ preferences in nonce words and loanwords are drawn from native patterns in the lexicon
Hypothesis 3

- speakers do not generalise fine-grained distinctions like vowel + singleton / geminate co-occurrence restrictions or differences in place of articulation from the lexicon
- this might be the case because there is not enough data to generalise from (low type frequency of different short vowel + short consonant / geminate combinations)
- the more simplified and specific the learning model is, the more likely it is to match native speakers’ judgements on nonce monosyllables
- it indicates that there may be bias for geminate markedness
Future work

- incorporating external factors (which came up in the wug test) into the analysis
- look at geminates in different contexts (not just word-final position)
- look at token frequency (in addition to type frequency distributions)
- build a model which is able to ignore other detail and concentrate on geminate markedness (if needed and possible)
Thank you for comments and discussion: Adam Albright, Donca Steriade, Michael Kenstowicz, Edward Flemming, Miklós Törkenczy, Katalin Mády, Csaba Oravecz, and audiences at the 22th Manchester Phonology Meeting and Phonology Circle at MIT.