Phonotactic preferences in Polish and English: Quantitative perspective

Katarzyna Dziubalska-Kołaczyk
dkasia@ifa.amu.edu.pl
Grzegorz Krynicki
krynicki@ifa.amu.edu.pl

Adam Mickiewicz University
Aim

• present a more comprehensive approach to phonotactics than the one originally proposed in Beats-&-Binding model
• corroborate this approach by statistical evidence from Polish and English
B&B phonotactics

• intersegmental cohesion depends on the complex interplay of adjacent segments, as allowed by language-specific phonotactics

• intersegmental cohesion determines syllable structure, rather than being determined by it (if one insists on the notion of the „syllable”)
B&B phonotactics

the universal preferences specify the optimal shape of a particular cluster in a given position by referring to the **Net Auditory Distance Principle (NAD Principle)**

\[
\text{NAD} = |\text{MOA}| + |\text{POA}| + |\text{LX}|
\]

whereby MOA, POA and LX are the absolute values of differences in the Manner of Articulation, Place of Articulation and Voicing of the neighbouring sounds respectively.

Example:

\[
\text{NAD (C1,C2)} \geq \text{NAD (C2,V)}
\]

meaning:

In word-initial double clusters, the net auditory distance (NAD) between the two consonants should be greater than or equal to the net auditory distance between a vowel and a consonant neighbouring on it.
B&B phonotactics

• the phonotactic preferences specify the universally required relationships between net auditory distances within clusters which guarantee, if respected, preservation of clusters

• clusters, in order to survive, must be sustained by some force counteracting the overwhelming tendency to reduce towards CV's

• this force is a perceptual contrast defined above as NAD
# Table of consonants

<table>
<thead>
<tr>
<th>4</th>
<th>3</th>
<th>2</th>
<th>1</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>obstruent</strong></td>
<td><strong>sonorant</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>stop</td>
<td>fricative</td>
<td>sonorant stop</td>
<td>approximant</td>
<td>V</td>
</tr>
<tr>
<td>affricate</td>
<td></td>
<td></td>
<td>semiV</td>
<td></td>
</tr>
<tr>
<td>p b</td>
<td>φ β</td>
<td>m</td>
<td>w</td>
<td>labial</td>
</tr>
<tr>
<td></td>
<td>f v</td>
<td>n</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>t d t d</td>
<td>θ δ</td>
<td>n</td>
<td>r l</td>
<td>coronal</td>
</tr>
<tr>
<td>t d</td>
<td>s z s z</td>
<td>n</td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>t d</td>
<td>s z s z</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k g c j</td>
<td>c z x y</td>
<td>j</td>
<td></td>
<td>dorsal</td>
</tr>
<tr>
<td>C J</td>
<td></td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>?</td>
<td>h</td>
<td></td>
<td></td>
<td>laryngeal (glottal)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>6</td>
</tr>
</tbody>
</table>
B&B phonotactics

• consider the preference for initial double clusters
  \[ \text{NAD} (C1,C2) \geq \text{NAD} (C2,V) \]

• let us now define two Net Auditory Distances between the sounds \((C1, C2)\) and \((C2, V)\) where
  
  \[ 
  \begin{align*}
  C1 & : (MOA1, POA1, Lx1) \\
  C2 & : (MOA2, POA2, Lx2) \\
  V & : (MOA3, Lx3)
  \end{align*} 
  \]

  in terms of the following metric for \((C1, C2)\) cluster
  
  \[ |MOA1 - MOA2| + |POA1 - POA2| + |Lx1 - Lx2| \]
  
  and
  
  \[ |MOA2 - MOA3| + |Lx2 - Lx3| \]

  for \((C2, V)\) cluster
B&B phonotactics

Example:

in CCV in E. Try

\[ t = (4, 2, 0), \quad r = (1, 2, 1), \quad V = (0, 0, 1) \]

\[ \text{NAD (C1, C2)} = |4-1| + |2-2| + |0-1| = 3+0+1=4 \]

\[ \text{NAD (C2, V)} = |1-0| + |1-1| = 1+0=1 \]

thus, the preference

\[ \text{NAD (C1,C2)} \geq \text{NAD (C2,V)} \]

is observed because \( 4 > 1 \)

- NAD makes finer predictions than the ones based exclusively on sonority:

\[ \text{prV} > \text{trV}, \quad \text{krV} > \text{trV}, \quad \text{trV} > \text{drV}, \text{ etc.} \]
Selected Polish clusters and NAD

Cluster types in Polish acc. to NAD

pr  fr  lv  mś  rd  fk  mb  śk  sk

MOA+POA+Lx  C2V  NAD
Phonotactic Calculator - General Purpose

Enable fine-tuning and developing phonotactic theories by statistical analysis of phonetic dictionaries and phonetically annotated corpora from various languages
Phonotactic Calculator - Requirements

• Various cluster lengths at all word positions
• Formulating new phonotactic hypotheses
• Feedback on predictability of a phonotactic hypothesis
• Choice or customization of
  – available phone sets, features of each phone and scores for each feature
  – available phonetic dictionaries and languages (PolSynt, Festvox, Festival)
  – metrics used for calculating distances between phones (taxicab, euclidean)
  – accepted phonetic alphabets (IPA, SAMPA)
Phonotactic calculator

Calculate Net Auditory Distances between the sounds of a cluster

1. Number of consonants in the cluster
   - 2
   - 3
   - 4
   - 5
   - 6

2. Part of the word where the cluster is located
   - onset
   - medial position
   - offset

3. Assume input strings to be whole words or just consonant clusters?
   - consonant clusters
   - words

5. Phonetic alphabet used above
   - IPA
   - SAMPA
   - X-SAMPA

6. What metric should be used to calculate NAD
   - taxicab $\sum_{i=1}^{c} (MO_{i} + PO_{i})$, where $c$ is a number of consonants in cluster
   - euclidean $\sqrt{MO_{i}^2 + PO_{i}^2}$
   -
Phonotactic calculator

Calculate Net Auditory Distances between the sounds of a cluster

1. Number of consonants in the cluster
   - 2
   - 3
   - 4
   - 5
   - 6

2. Part of the word where the cluster is located
   - Onset
   - Medial position
   - Offset

3. Assume input strings to be whole words or just
Testing hypotheses (preferences) about Net Auditory Distance between sounds in a cluster

1. Test the following hypotheses
   1. consequent: nad(c1,c2) => nad(c2,v)           antecedent: B C C V *
   2. consequent: nad(v,c1) => nad(c1,c2)           antecedent: V C C B
   3. consequent: nad(v1,c1) => nad(c1,c2) AND nad(c1,c2) < nad(c2,v2)   antecedent: V C C V
   4. consequent: nad(c1,c2) < nad(c2,c3) AND nad(c2,c3) => nad(c3,v)   antecedent: B C C C V
   5. consequent: nad(v,c1) <= nad(c1,c2) AND nad(c1,c2) > nad(c2,c3)   antecedent: V C C C B
   6. consequent: nad(v1,c1) <= nad(c1,c2) AND nad(c2,c3) < nad(c3,v2)   antecedent: V C C C V

7. Format the consequent as above
   Format the antecedent as above

...on the clusters that do not contain any morphological boundary
   ○ from Polish
   ○ from English

2. Comparison of tests of selected hypotheses on clusters containing and devoid of morphological boundaries
   ○ Testing 1st hypothesis on 5000 clusters from top frequency** Polish words containing CCV cluster in onset position
   ○ Testing 2nd hypothesis on 2000 clusters from top frequency** English words containing VCC cluster in coda position

3. Show details of NAD calculation
   ○ No
   ○ Yes (Operates only if you selected any of the options in point 2)

Notice:

*  B = word boundary, V = vowel, C = consonant
** Frequency lists were compiled from European Union documentation corpus of approx. 20mln tokens.
Empirical data

- Phonetic dictionaries for English (Festival)
- Phonetically transcribed word lists and frequency lists (PolSynt)
- Annotating these resources for morphological information
  - simplex vs complex words
  - clusters containing and devoid of morphological boundary
Automatic selection of simplexes

• English:
  – 127 040 CMU entries
  – 20.9% of these were recognized by PC Kimmo and classified as simplex
  – 91.2% of these were not compounds. Final list of 10245 entries (8.06% of CMU)

• Polish
  – Phonetically transcribed 120 000 entries of Great PWN dictionary
  – Semi-automatic heuristics (removing words with derivational morphemes and potential compounds) resulted in 13691 words
Manual selection of simplexes

- English: list of 2000 VCC clusters classified manually into
  - 1114 containing morphological boundary
  - 886 not containing any morphological boundaries

- Polish: list of 5000 CCV clusters classified manually into
  - 162 containing morphological boundary
  - 4838 not containing any morphological boundaries
Results of testing 6 phonotactic preferences on semi-automatic simplexes

<table>
<thead>
<tr>
<th>POLISH</th>
<th>Clusters that apply</th>
<th>Clusters that meet the preference</th>
<th>Perc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>nad(c1,c2) ≥ nad(c2,v)</td>
<td>708</td>
<td>346</td>
<td>48.87%</td>
</tr>
<tr>
<td>nad(v,c1) &lt;= nad(c1,c2)</td>
<td>416</td>
<td>134</td>
<td>32.21%</td>
</tr>
<tr>
<td>nad(v1,c1) ≥ nad(c1,c2) &amp; nad(c1,c2) ≤ nad(c2,v2)</td>
<td>3793</td>
<td>1798</td>
<td>47.40%</td>
</tr>
<tr>
<td>nad(c1,c2) &lt; nad(c2,c3) &amp; nad(c2,c3) ≥ nad(c3,v)</td>
<td>105</td>
<td>70</td>
<td>66.67%</td>
</tr>
<tr>
<td>nad(v,c1) ≤ nad(c1,c2) &amp; nad(c1,c2) &gt; nad(c2,c3)</td>
<td>9</td>
<td>6</td>
<td>66.67%</td>
</tr>
<tr>
<td>nad(v1,c1) ≥ nad(c1,c2) &amp; nad(c2,c3) &lt; nad(c3,v2)</td>
<td>555</td>
<td>135</td>
<td>24.32%</td>
</tr>
<tr>
<td>mean</td>
<td></td>
<td></td>
<td>47.69%</td>
</tr>
</tbody>
</table>
Results of testing 6 phonotactic preferences on semi-automatic simplexes

<table>
<thead>
<tr>
<th>ENGLISH</th>
<th>Clusters that apply</th>
<th>Clusters that meet the preference</th>
<th>Perc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>nad(c1,c2) (\geq) nad(c2,v)</td>
<td>1232</td>
<td>1004</td>
<td>81.49%</td>
</tr>
<tr>
<td>nad(v,c1) =&lt; nad(c1,c2)</td>
<td>929</td>
<td>663</td>
<td>71.37%</td>
</tr>
<tr>
<td>nad(v1,c1) (\geq) nad(c1,c2) &amp; nad(c1,c2) (\leq) nad(c2,v2)</td>
<td>1243</td>
<td>549</td>
<td>44.17%</td>
</tr>
<tr>
<td>nad(c1,c2) &lt; nad(c2,c3) &amp; nad(c2,c3) (\geq) nad(c3,v)</td>
<td>91</td>
<td>91</td>
<td>100.00%</td>
</tr>
<tr>
<td>nad(v,c1) (\leq) nad(c1,c2) &amp; nad(c1,c2) &gt; nad(c2,c3)</td>
<td>27</td>
<td>23</td>
<td>85.19%</td>
</tr>
<tr>
<td>nad(v1,c1) (\geq) nad(c1,c2) &amp; nad(c2,c3) &lt; nad(c3,v2)</td>
<td>159</td>
<td>32</td>
<td>20.13%</td>
</tr>
<tr>
<td><strong>mean</strong></td>
<td></td>
<td></td>
<td><strong>67.06%</strong></td>
</tr>
</tbody>
</table>
# Results of testing 6 phonotactic preferences on manual simplexes

## POLISH

<table>
<thead>
<tr>
<th>Hyp. no. 1. $\text{nad}(c_1,c_2) \geq \text{nad}(c_2,v)$</th>
<th>Clusters that apply</th>
<th>Clusters that meet the preference</th>
<th>Perc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphologically complex</td>
<td>5000</td>
<td>2453</td>
<td>49.06%</td>
</tr>
<tr>
<td>Morphologically simple</td>
<td>4838</td>
<td>2412</td>
<td>49.86%</td>
</tr>
</tbody>
</table>

## ENGLISH

<table>
<thead>
<tr>
<th>Hyp. no. 2. $\text{nad}(v_1,c_1) \leq \text{nad}(c_1,c_2)$</th>
<th>Clusters that apply</th>
<th>Clusters that meet the preference</th>
<th>Perc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphologically complex</td>
<td>2000</td>
<td>1063</td>
<td>53.15%</td>
</tr>
<tr>
<td>Morphologically simple</td>
<td>1114</td>
<td>404</td>
<td>36.27%</td>
</tr>
<tr>
<td>Morphologically simple</td>
<td>886</td>
<td>659</td>
<td>74.38%</td>
</tr>
</tbody>
</table>
Conclusions on quantitative analysis

• Phonotactic preferences are met in Polish and English to a moderately high degree (47% and 67% resp.)

• Both in Polish and English, morphologically simple words meet selected preferences (1st and 2nd resp.) to a greater degree than morphologically complex words

• More experiments are necessary to prove statistical significance of differences between morphologically simple and complex words with respect to their compliance with all phonotactic preferences
Morphonotactics
(cf. Dressler & Dziubalska-Kołaczyk 2007)

• morphonotactics is the area of interaction between morphotactics and phonotactics
• phonotactic preferences hold for monomorphemic, ”lexical” words
• the less respected the preferences are, the more marked clusters arise
• morphonotactic clusters (across morpheme boundaries) are much more likely to be marked
Morphonotactics: English examples

• exclusively morphotactically motivated consonant sequences are the word-final clusters /-fs, -vz/ as in laughs, loves, wife’s, wives, which occur only in plurals, third singular present forms and in Saxon genitives

• also /-bz, -gz, -ðz, -Øs, -mz, -md, -nz/ (except in names), as in bobs, Bob’s, eggs, deaths, wreathes, clothes, times, seems, seemed, tons
Morphonotactics: German examples

• exclusive morphological motivation exists for the clusters /-mst/, as in kämm+st ‘you comb’, schlimm+st ‘worst’, ge+sims+t ‘with a moulding or mantlepiece’, /-xst, -fst/, as in lach+st ‘you laugh’, tun+lich+st ‘if possible’, schläf+st ‘you sleep’, zu+tief+st ‘deepest’, with the affricate /-pfst/, as in tropf+st ‘you drip’, stampf+st ‘you stamp’ and in the longer consonant clusters /-rkst/, as in werk+st ‘you work’, ver+korks+t ‘kink’, /-lkst/, as in welk+st ‘you fade’, /-nkst/, as in stink+st ‘you stink’, /-lpst, -mpst/, as in stülp+st ‘you turn up’, selb+st ‘self’, tramp+st ‘you tramp’, plumps+(s)t ‘you plop’
Morphonotactics: Polish examples

• there is no monomorphemic \textit{ws}- [fs-] cluster
• \textit{wsz}- [fʂ-] occurs in the fossilized but frequent prefixoids \textit{wsze}, \textit{wszech}, \textit{wszem} ‘all, everybody’, in archaic \textit{wszędy} ‘everywhere’, in frequent \textit{wszystko} ‘everything’ (all of which are semantically related in an irregular way), and in archaic \textit{wszak} ‘after all’
• \textit{wsi}- [fɕ-] appears in the Russian loan \textit{wsio} ‘everything’ and in the colloquial pronunciation of the abbreviation \textit{WSJO} [fɕo] from the recent term \textit{Wyższa Szkoła Języków Obcych} ‘college of modern languages’
• all the other instances of the three initial clusters are of a morphonotactic nature