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*)

NAD = |MOA| + |POA| + |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:

NAD (C1,C2) ≥ 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

4	3	2	1	0	
obstruent		**sonorant**			
stop	**fricative**	**sonorant stop**	**approximant**	**V**	
affricate			**semiV**		
p b	ɸ β f v	m ɱ	w	**labial**	**1**
t̪ d̪ t d ʈ ɖ	θ ð s̪ z̪ s z ʂ ʐ ʃ ʒ	n̪ n	r l	**coronal**	**2**
k g c ɟ	ç ʑ x ɣ	ɲ ŋ	j	**dorsal**	**3**
				radical	**4**
ʔ	h			**laryngeal (glottal)**	**5**

B&B phonotactics

- consider the preference for initial double clusters

  NAD (C1,C2) ≥ NAD (C2,V)

- let us now define two Net Auditory Distances between the sounds (C1, C2) and (C2, V) where

  C1 (MOA1, POA1, Lx1)
  C2 (MOA2, POA2, Lx2)
  V (MOA3, Lx3)

  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), r = (1, 2, 1), V = (0, 0, 1)

NAD (C1, C2) = |4-1| + |2-2| + |0-1| = 3+0+1=**4**

NAD (C2, V) = |1-0| + |1-1| = 1+0=**1**

thus, the preference

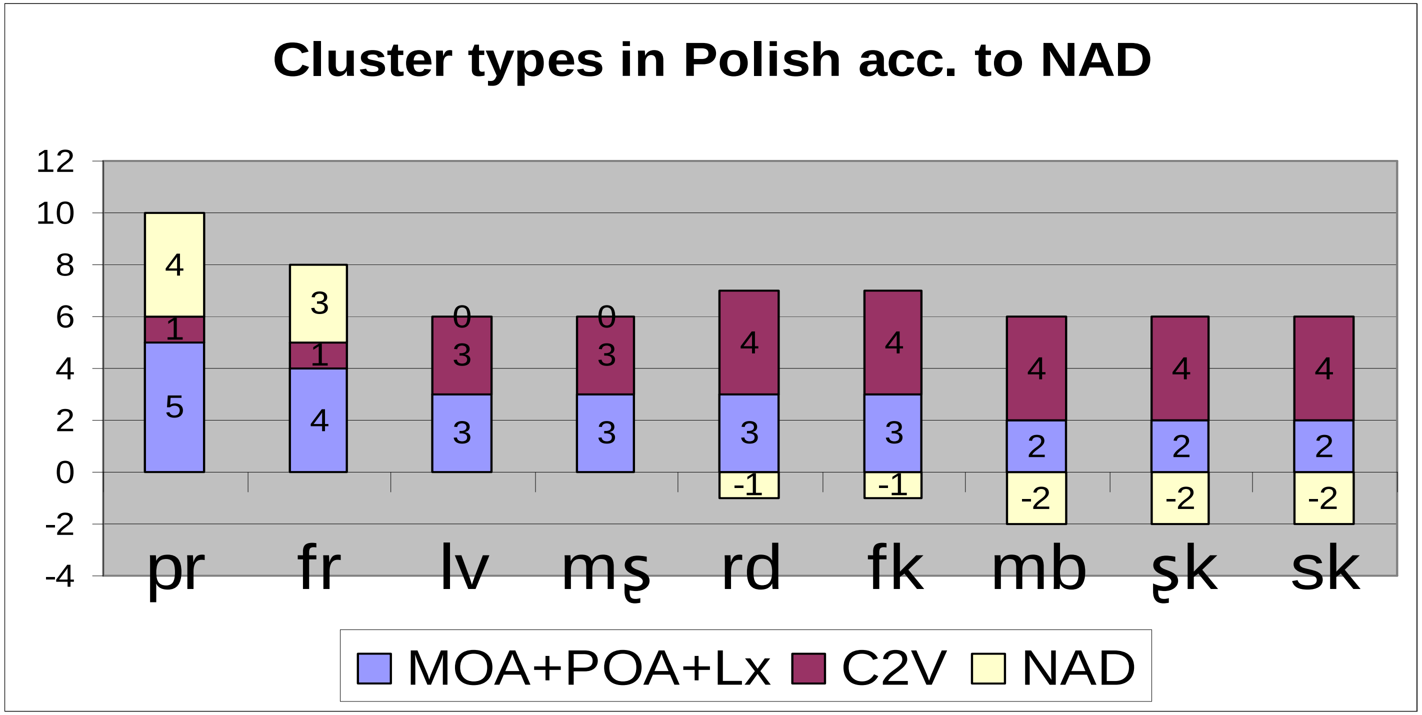
NAD (C1,C2) ≥ NAD (C2,V)

is observed because **4** > **1**

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

prV > trV, krV > trV, trV > drV, etc.

Selected Polish clusters and NAD



Phonotactic Calculator - General Purpose

Enable fine-tunining 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

clusters | hypotheses | search | about

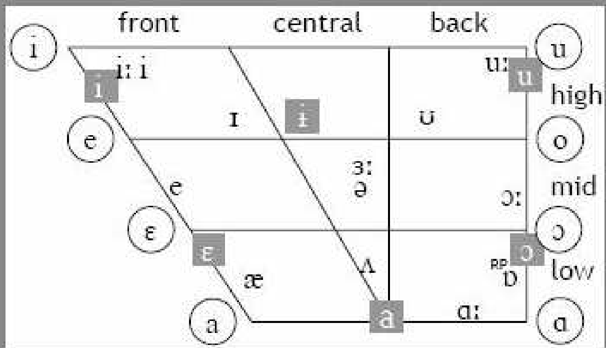
Calculate Net Auditory Distances between the sounds of a cluster

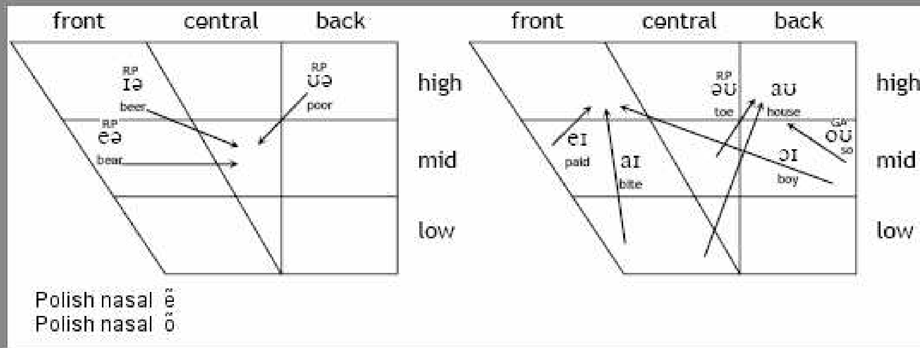
1. **Number of consonants in the cluster**
   ☐ 2 ☐ 3 ☐ 4 ☐ 5 ☐ 6

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

	Bilabi	LabDen	LabVel	Dental	Alveol	PalAlveo	AlvPalat	Retrof	Palata	Velar	Glotta
Plosive	p b			t̪ d̪	t d					k g	ʔ
Nasal	m			n̪	n				ɲ	ŋ	
Trill					r						
Tap					ɾ						
Fricativ		f v		θ ð s̪ z̪	s z	ɕ ʑ	ʃ ʒ			x	h
Affrica				t͡s d͡z		t͡ɕ d͡ʑ	t͡ʃ d͡ʒ				
Approx			w		ɹ			ɻ	j		
LatApp					l						





5. **Phonetic alphabet used above**
   ◉ IPA ○ SAMPA ○ X-SAMPA

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

Phonotactic calculator

clusters | hypotheses | search | about

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**

Phonotactic calculator

clusters | hypotheses | search | about

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)

testuj

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

POLISH	Clusters that apply	Clusters that meet the preference	Perc.
nad(c1,c2) ≥ nad(c2,v)	708	346	48,87%
nad(v,c1) =< nad(c1,c2)	416	134	32,21%
nad(v1,c1) ≥ nad(c1,c2) & nad(c1,c2) ≤ nad(c2,v2)	3793	1798	47,40%
nad(c1,c2) < nad(c2,c3) & nad(c2,c3) ≥ nad(c3,v)	105	70	66,67%
nad(v,c1) ≤ nad(c1,c2) & nad(c1,c2) > nad(c2,c3)	9	6	66,67%
nad(v1,c1) ≥ nad(c1,c2) & nad(c2,c3) < nad(c3,v2)	555	135	24,32%
		mean	47,69%

Results of testing 6 phonotactic preferences on semi-automatic simplexes

ENGLISH	**Clusters that apply**	**Clusters that meet the preference**	**Perc.**
nad(c1,c2) ≥ nad(c2,v)	1232	1004	81,49%
nad(v,c1) =< nad(c1,c2)	929	663	71,37%
nad(v1,c1) ≥ nad(c1,c2) & nad(c1,c2) ≤ nad(c2,v2)	1243	549	44,17%
nad(c1,c2) < nad(c2,c3) & nad(c2,c3) ≥ nad(c3,v)	91	91	100,00%
nad(v,c1) ≤ nad(c1,c2) & nad(c1,c2) > nad(c2,c3)	27	23	85,19%
nad(v1,c1) ≥ nad(c1,c2) & nad(c2,c3) < nad(c3,v2)	159	32	20,13%
		mean	67,06%

Results of testing 6 phonotactic preferences on manual simplexes

POLISH	**Clusters that apply**	**Clusters that meet the preference**	**Perc.**
Hyp. no. 1. nad(c1,c2) >= nad(c2,v)	5000	2453	□49.06%
Morphologically complex	162	□41	25.31%
Morphologically simple	4838	2412	49.86%

ENGLISH	**Clusters that apply**	**Clusters that meet the preference**	**Perc.**
Hyp. no. 2. nad(v1,c1) =< nad(c1,c2)	2000	1063	53.15%
Morphologically complex	1114	404	□36.27%
Morphologically simple	886	659	74.38%

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, -θs, -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 *ws*- [fs-] cluster
- *wsz*- [fʂ-] occurs in the fossilized but frequent prefixoids *wsze*, *wszech*, *wszem* ‘all, everybody’, in archaic *wszędy* ‘everywhere’, in frequent *wszystko* ‘everything’ (all of which are semantically related in an irregular way), and in archaic *wszak* ‘after all’
- *wsi*- [fɕ-] appears in the Russian loan *wsio* ‘everything’ and in the colloquial pronunciation of the abbreviation *WSJO* [fɕo] from the recent term *Wyższa Szkoła Jezyków Obcych* ‘college of modern languages’
- all the other instances of the three initial clusters are of a morphonotactic nature