Gradient similarity in antigemination: evidence from allomorph selection

Aljoša Milenković
aljosamilenkovic@g.harvard.edu

OCP 22, Universiteit van Amsterdam
February 05, 2025

Antigemination (AG)

avoidance of adjacent identical consonants

1. PROCEDURAL: restrictions on application of phonological processes (McCarthy 1986; Borowsky 1987; Yip 1988)
2. STATIC/PHONOTACTIC: domain-internal cooccurrence restrictions (Pierrehumbert 1993; Frisch et al. 2004)

This talk: Type 1, procedural AG

Near identity avoidance

- AG not limited to fully identical consonants (1)
- *sufficiently similar* nonidentical consonants also avoided (2)

English:

(1) ɪ-epenthesis to break up /d-d/:
/flʌd-d/ → [flʌdɪd] *flooded*

(2) also in /t-d/ (not a geminate):
/pɛt-d/ → [pɛtɪd] *petted*

Baković 2005: AG and assimilation

- features "ignored" for determination of identity **trigger assimilation** (Baković, 2005, 2006, 2017; Pajak and Baković, 2010)
- LOOKAHEAD EFFECT: AG targets sequences that could become a geminate by assimilation (Adler and Zymet, 2021)
- Formal OT (Prince and Smolensky, 2004) account: partial identity avoidance as a joint effect of NoGem and Agree constraints (Baković, 2005, 2006)

Near identity AG: NoGem–Agree interaction

Works well in strict-ranking OT:

(3)

/pɛt-d/			NoGem	Agree$_{\text{voi}}$	DepV	Ident$_{\text{voi}}$
a.		pɛtd		*!		
b.		pɛtt	*!			*
c.	☞	pɛtɪd			*	

MaxEnt HG: original vs. derived geminates

Baković (2005)'s model in MaxEnt HG (Goldwater and Johnson, 2003; Hayes and Wilson, 2008)

(4)

		NoGem	Agree$_{voi}$	DepV	Ident$_{voi}$	$\mathscr{H}$	$e^{\mathscr{H}}$	p
		2	2	1	1			
/dd/	a. dd	−1				−2	.14	.27
	b. dɪd			−1		−1	.37	.73
/td/	a. dd	−1			−1	−3	.05	.12
	b. tɪd			−1		−1	.37	.88

Prediction: greater cumulative penalty of derived compared to original geminates → stronger avoidance

This talk

- Is this prediction a desirable one? B/C/S data suggest otherwise
- Phonologically-conditioned allomorph selection; not regular phonology
- Mobile *a*-morphemes: free variation between C# and CV# allomprphs

(5) Mobile *a*-morphemes

a.	s tɔːrtɔːm	∼	sa tɔːrtɔːm	‘with a cake’
b.	k tɔːrɲu	∼	ka tɔːrɲu	‘toward the tower’
c.	dɔbr-ɔːg	∼	dɔbr-ɔːga	‘good-GEN.SG.M/N’

- [sa] strongly preferred over [s] before words starting with [s], [ʃ], [z], or [ʒ] (Stevanović, 1991; Barić et al., 1997)

BCS obstruent inventory

voiceless	p	t	k	f	s	ʃ	x	ʦ	ʧ	ʨ
voiced	b	d	g		z	ʒ			ʤ	ʥ

[+anterior] fricatives/affricates

[−anterior] fricatives/affricates

Avoidance of the [s] allomorph before [s], [ʃ], [z], and [ʒ]-initial words in line with Baković (2005)'s theory, given that B/C/S display **voicing and anteriority assimilation in sandhi**

Voicing assimilation

Both word-internally and in sandhi:

(6) $$\begin{bmatrix} -\ \text{son} \\ \alpha\ \text{voi} \end{bmatrix} \rightarrow [\beta\ \text{voi}]\ /\ \_\begin{bmatrix} -\ \text{son} \\ \beta\ \text{voi} \end{bmatrix}$$

/iz-/ ‘out of, from’:

(7)	a.	iz-raːditi	‘work out’	⟨izraditi⟩
	b.	iz rata	‘from the war’	⟨iz rata⟩
(8)	a.	is-kupiti	‘gather’	⟨iskupiti⟩
	b.	is kutɕɛː	‘from the house’	⟨iz kuće⟩

Anteriority assimilation

Both word-internally and in sandhi:

(9) $$\begin{bmatrix}\text{CORONAL}\\ +\ \text{cont}\\ +\ \text{ant}\end{bmatrix} \rightarrow [-\ \text{ant}]\ /\ \_\begin{bmatrix}\text{CORONAL}\\ -\ \text{ant}\end{bmatrix}$$

/iz-/ ‘out of, from’:

(10)	a.	iʃ-ʧupati	‘pull out’	⟨iščupati⟩
	b.	iʃ ʧɛga	‘from what’	⟨iz čega⟩
(11)	a.	iʒ-ʤikʎati	‘grow’	⟨iždžikljati⟩
	b.	iʒ ʤɛpa	‘out of the pocket’	⟨iz džepa⟩

Corpus survey

Extracted bigrams with s/sa and the following word from the {bs,hr,sr}WaC corpora (Ljubešić and Klubička, 2014):

- Bosnian: bsWac
- Croatian: hrWaC
- Serbian: srWaC

	N bigrams	N unique lemmas
bsWaC	1,749,389	77,536
hrWac	8,420,018	216,275
srWaC	3,301,108	117,278

- All corpora lemmatized and morphosyntactically tagged

Corpus survey

Corpus search excluded (via regex):

1. **acronyms** (discrepancy between spelling and pronunciation):

   (12) ⟨s SAD-om⟩ [s ɛ̱s a dɛɔm] ‘with the USA’

2. **spelling errors** (diacritic omission)

   (13) ⟨s cijim⟩ [ʃ ʧijim] ‘with whose’

3. **lexicalized expressions** (invariable realization, not governed by the phonological grammar):

   (14) [sa mnɔːm] *[s mnɔːm] ‘with me’

Segmental effects on [s]/[sa] realization

[s] more disfavored before **voiced obstruents** than elsewhere:

(15) AGREE$_{\text{voi}}$
Assess a violation for every pair of obstruents that disagree in voice.

[s] more disfavored before **posterior coronals** than elsewhere:

(16) AGREE$_{\text{ant}}$
Assess a violation for every pair of coronal obstruents that disagree in anteriority.

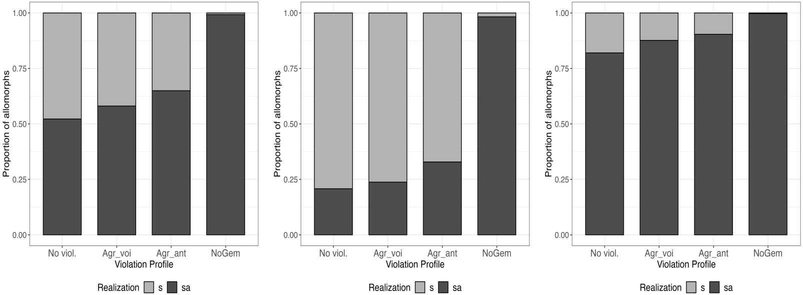
AG: [s] avoided before {s, z, ʃ, ʒ}:

(17) NOGEMINATE
Assess a violation for adjacent identical consonants.

Analysis: individual effects of constrains

violation profile	environment	[s] violates
baseline	__#{i, ɛ, a, ɔ, u}, __#{ʋ, m, r, l, n, ʎ, ɲ, j}, __#{p, t, k, f, x, ʦ},	no violation
voice mismatch	__#{b, d, g}	Agree$_{\text{voi}}$
anteriority mismatch	__#{ʧ, ʨ}	Agree$_{\text{ant}}$
geminate	__#s	NoGem

Results: independent constraint contributions



(a) Bosnian

(b) Croatian

(c) Serbian

Figure 1: Proportion of s/sa (y-axis) by violation profile (x-axis).

Logistic regression analysis

```
cbind(sa, s) ∼ agree_voi * agree_ant * no_gem
```

	Croatian				Serbian			
	β	SE	z	p	β	SE	z	p
(Intercept)	-1.34	.00	-1355	.000	1.52	.00	907	.000
AGREE$_{voi}$:1	.17	.00	61	.000	.44	.01	77	.000
AGREE$_{ant}$:1	.62	.01	83	.000	.72	.02	40	.000
NOGEM:1	5.37	.01	698	.000	4.25	.03	150	.000
AGREE$_{voi}$:1* AGREE$_{ant}$:1	-.2	.03	-6	.000	-.61	.05	-13	.000
AGREE$_{voi}$:1* NOGEM:1	-1.1	.01	-80	.000	-.95	.06	-17	.000
AGREE$_{ant}$:1* NOGEM:1	-1.8	.02	-83	.000	-2.03	.07	-29	.000
AGREE$_{voi}$:1* AGREE$_{ant}$:1* NOGEM:1	.66	.04	15	.000	-.87	.11	-8	.000

See Appendix 1 for more details & data.

Interpretation of regression results

Main effects:

- positive main effect of AGREE$_{\text{voi}}$ → likelihood of [sa] increases in AGREE$_{\text{voi}}$-violating environments
- positive main effect of AGREE$_{\text{ant}}$ → likelihood of [sa] increases in AGREE$_{\text{ant}}$-violating environments
- strong positive main effect of NOGEM → AG: likelihood of [sa] increases substantially before [s] (full identity pair)

Interaction effects → partial identity pairs:

- negative interaction effect of AGREE$_{\text{voi}}$ and NOGEM → likelihood of [sa] before [z] and [ʒ] drops relative to [s#s]
- negative interaction effect of AGREE$_{\text{ant}}$ and NOGEM → likelihood of [sa] before [ʃ] and [ʒ] drops relative to [s#s]

Degree of feature overlap ⇔ strength of avoidance

- SUBLINEARITY (at the level of data): *negative* interaction effects between both AGREE constraints and NOGEM

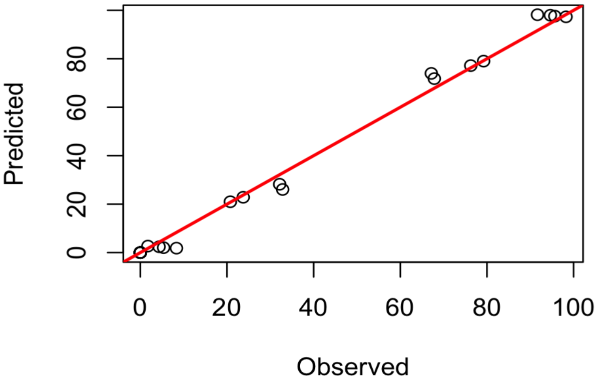
(18) Avoidance scale (strongest to weakest)
s#s > s#z > s#ʃ > s#ʒ

- additional verification: logistic regression with stepwise difference coded comparisons (only Croatian reported).

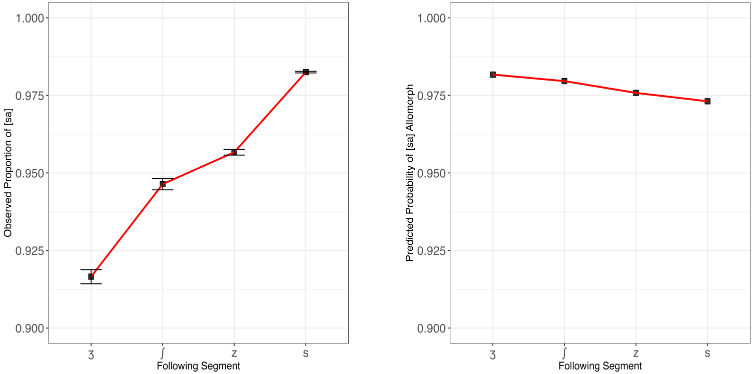
`realization ∼ context_recoded`

	β	SE	z	p
(Intercept) [s#s] (**"0"**)	3.1	.01	453.4	.000***
[s#z] (**"1"**) vs. [s#s] (**"0"**)	-.93	.01	-69.2	.000***
[s#ʃ] (**"2"**) vs. [s#z] (**"1"**)	-.22	.02	-10.4	.000***
[s#ʒ] (**"3"**) vs. [s#ʃ] (**"2"**)	-.47	.02	-19.95	.000***

MaxEnt model (Croatian): overall model fit



Zooming in: s∼sa#{s, z, ʃ, ʒ}



(a) Observed rate of [sa]

(b) Predicted probability of [sa]

Flattening (and slight reversal) in the MaxEnt model (!)

Conclusion

- Implemented in MaxEnt HG, Baković (2005)'s constraint model predicts more robust avoidance of derived geminates compared to original geminates, or no difference
- Impossible pattern: original geminates avoided more robustly than derived ones; **attested in BCS**
- **generalization**: degree of avoidance gradiently proportional to feature overlap:
  - full overlap → strongest avoidance
  - more feature mismatches → less robust avoidance

Conclusion

- MaxEnt model employing constraints à la Baković (2005):
  1. overestimates the strength of avoidance in partial identity pairs
  2. slightly underestimates the strength of avoidance in the s#s pair

❖ not problematic if phonologically-conditioned allomorph selection is external to phonology proper

Acknowledgments

- Special thanks to Michael Becker and Kevin Ryan for continuous support
- This project has been partly funded from the Harvard linguistics research fund

▲ Data & analysis script

▲ MaxEnt model files

References I

Adler, J. and Zymet, J. (2021). Irreducible parallelism in phonology. *NLLT*, 39(2):367–403.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge university press.

Baković, E. (2005). Antigemination, assimilation and the determination of identity. *Phonology*, 22(3):279–315.

Baković, E. (2006). Partial identity avoidance as cooperative interaction. *WECOL 2004*.

Baković, E. (2017). Apparent 'sufficiently similar'degemination in catalan is due to coalescence. *Proceedings of the Linguistic Society of America*, pages 1–9.

Barić, E., Lončarić, M., Malić, D., Pavešić, S., Peti, M., Zečevic, V., and Znika, M. (1997). *Hrvatska gramatika*. Školska knjiga.

Borowsky, T. (1987). Antigemination in english phonology. *Linguistic Inquiry*, 18(4):671–678.

References II

Breiss, C. (2020). Constraint cumulativity in phonotactics: Evidence from artificial grammar learning studies. *Phonology*, 37(4):551–576.

Flemming, E. (2021). Comparing maxent and noisy harmonic grammar. *Glossa*, 6.

Frisch, S. A., Pierrehumbert, J. B., and Broe, M. B. (2004). Similarity avoidance and the ocp. *Natural language & linguistic theory*, 22(1):179–228.

Goldwater, S. and Johnson, M. (2003). Learning OT constraint rankings using a maximum entropy model. In J. Spenader, A. E. and Östen Dahl, editors, *Proceedings of the Stockholm workshop on Variation within Optimality Theory*, pages 111–120.

Hayes, B. (2022). Deriving the wug-shaped curve: A criterion for assessing formal theories of linguistic variation. *Annual Review of Linguistics*, 8(1):473–494.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *LI*, 39(3):379–440.

References III

Ljubešić, N. and Klubička, F. (2014). {bs, hr, sr} wac-web corpora of bosnian, croatian and serbian. In *Proceedings of the 9th web as corpus workshop (WaC-9)*, pages 29–35.

Mayer, C., Tan, A., and Zuraw, K. R. (2024). Introducing maxent. ot: an r package for maximum entropy constraint grammars. *Phonological Data and Analysis*, 6(4):1–44.

McCarthy, J. (1986). OCP effects: Gemination and antigemination. *LI*, 17(2):207–263.

Pajak, B. and Baković, E. (2010). Assimilation, antigemination, and contingent optionality: the phonology of monoconsonantal proclitics in polish. *Natural Language & Linguistic Theory*, 28:643–680.

Pierrehumbert, J. (1993). Dissimilarity in the arabic verbal roots. In *Proceedings of the Northeast Linguistics Society*, volume 23, pages 367–381.

Prince, A. and Smolensky, P. (1993/2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Blackwell. Technical Report, Rutgers University and University of Colorado at Boulder, 1993. Revised version Blackwell, 2004.

References IV

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Smith, B. and Pater, J. (2020). French schwa and gradient cumulativity. *Glossa: a journal of general linguistics*, 5(1).

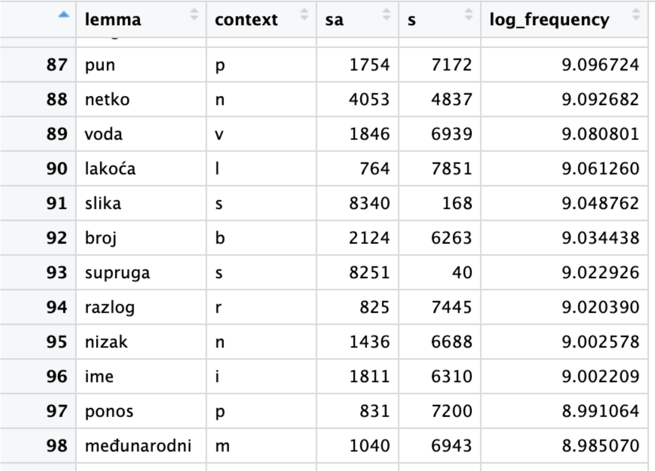
Stevanović, M. (1991). *Savremeni srpskohrvatski jezik:(gramatički sistemi i književnojezička norma).. Uvod, fonetika, morfologija*, volume 1. Naučna knjiga.

Yip, M. (1988). The obligatory contour principle and phonological rules: A loss of identity. *Linguistic inquiry*, 19(1):65–100.

Zuraw, K. and Hayes, B. (2017). Intersecting constraint families: an argument for Harmonic Grammar. *Lg*, 93(3):497–548.

Regression model: data set

Number of instances of each allomorph by the following lemma:



	lemma	context	sa	s	log_frequency
87	pun	p	1754	7172	9.096724
88	netko	n	4053	4837	9.092682
89	voda	v	1846	6939	9.080801
90	lakoća	l	764	7851	9.061260
91	slika	s	8340	168	9.048762
92	broj	b	2124	6263	9.034438
93	supruga	s	8251	40	9.022926
94	razlog	r	825	7445	9.020390
95	nizak	n	1436	6688	9.002578
96	ime	i	1811	6310	9.002209
97	ponos	p	831	7200	8.991064
98	međunarodni	m	1040	6943	8.985070

Regression model: variables

Logistic regression implemented in R (R Core Team, 2021)

Dependent variable: counts of each allomorph's realization aggregated by lemma: `cbind(sa,s)` (see Baayen, 2008, 197 for the method)

Fixed predictors:

1. `Agree_voi`: coded "1" for voicing mismatch between [s] and the following sound, "0" elsewhere;
2. `Agree_ant`: "1" for anteriority mismatch between [s] and the following sound, "0" elsewhere
3. `NoGem`: "1" if the following sound is [s], [z], [ʃ], or [ʒ], "0" elsewhere

and interactions between the fixed predictors

Data

Available in a GitHub repository

Background

- MaxEnt HG shown to be superior to other constraint-based frameworks that accommodate variation (Zuraw and Hayes, 2017; Breiss, 2020; Smith and Pater, 2020; Flemming, 2021; Hayes, 2022)
- Only Croatian data
- MaxEnt models for the Bosnian and Serbian data will be provided in the foreseeable future
- Implemented in R, using the `maxent.ot` package (Mayer et al., 2024)

Constraints #1

(19) *Mobile *a*
Assess a violation for every occurrence of mobile *a*.

Rationale: strong synchronic dispreference for mobile *a*: mobile *a*-allomorphs strongly dispreferred in all three languages, virtually unavailable in modern language in prepositions that are larger than a single consonant (e.g., *nad*∼*nada* ‘above’ is almost invariably realized as [nad])

(20) Have-μ
Assess a violation for every word (accentful or clitic) that contains no moras.

Function: to penalize vowel-less realizations [s]/[k]

Constraints #2

(21) NoGeminate
Assess a violation for adjacent identical consonants.

Function: penalizes true geminates: [s#s], but not e.g. [s#ʃ]

(22) Agree$_{\text{voi}}$
Assess a violation for every pair of obstruents that disagree in voice.

Function: triggers voicing assimilation

(23) Ident$_{\text{voi}}$
Assess a violation for every output segment which has a different voicing specification from its input correspondent.

Function: opposes anteriority assimilation

Constraints #3

(24) Agree$_{\text{ant}}$
Assess a violation for every pair of coronal obstruents that disagree in anteriority.

Function: triggers anteriority assimilation

(25) Ident$_{\text{ant}}$
Assess a violation for every output segment which has a different anteriority specification from its input correspondent.

Function: opposes anteriority assimilation

Learned weights

constraint	learned weight
*Mobile a	4.07
Haveμ	2.75
NoGem	4.92
Agree$_{\text{voi}}$	21.83
Ident$_{\text{voi}}$	.11
Agree$_{\text{ant}}$	16.05
Agree$_{\text{voi}}$	.28

Data

Simulation files & analysis script available in a GitHub repository