ICPhS 2023
August 7–11, 2023
Prague Congress Center, Czech Republic

20th International Congress of Phonetic Sciences (ICPhS)
August 7–11, 2023
Prague Congress Center, Czech Republic

Weighted phonetic distances in the evaluation of phonotactics

Paula Orzechowska, Katarzyna Dziubalska-Kołaczyk

paulao@amu.edu.pl, dkasia@amu.edu.pl

Introduction

One of the core themes of phonological research has been the relationship between the complexity of **consonant clusters** (e.g. in terms of sonority: van de Vijver & Baer-Henney 2012; CVC type: Jusczyk et al. 1994) and their usage. Our previous work on German (2019) has shown that the frequency of word-initial clusters is best predicted by the **manner of articulation distances** between two consonants (CC), and a consonant and a vowel (CV).

Goals. Although linear regression has been commonly used to model the relationship between structure and usage, it is not always best-fitted to frequency data (e.g. high-freq /ʃt/). Thus, here we compare the predictions of **linear regression** (which is limited to linear dependencies) and **XGBoost**, a method used to model non-normal distributions.

Net Auditory Distance

NAD (Dziubalska-Kołaczyk 2014) predicts cluster preferability based on three types of distances between pairs of segments and well-formedness conditions.

sonorant / obstruent distinction (SO)

place of articulation (POA)

S	A	F	N	L		G	V	
5.0	4.5	4.0	3.0	2.5	2	1.0	0	
p b			m				1.0	bilabial
	pf	f v					1.5	lab-dent
t d	ʦ	s z	n	l			2.0	alveolar
		ʃ ʒ					2.5	post-alv.
		ç ʝ				j	3.0	palatal
k g		x	ŋ				3.3	velar
		ʁ			ʀ		3.6	uvular
							4.0	(radical)
		h					5.0	(glottal)

manner of articulation (MOA)
Stop, Affricate, Fricative, Nasal, Liquid, Glide, Vowel

NAD(C1C2) ≥ NAD(C2V), where:
- NAD(C1C2) = |(MOA1-MOA2)| + |(POA1-POA2)|+ |SO|
- NAD(C2V) = |(MOA1-MOA2)| + |SO|

Example of preferred /bj/ (NAD product = 6)
- NAD (C1C2) = |5-1| + |1-3| + |1| = 4 + 2+ 1 = 7
- NAD(C2V) = |1-0| + 0 = 1

Gradient-boosted decision trees

M	Independent variables	MSE
1	MOA_C2V (0.46)+MOA_C1C2 (0.28)+ POA_C1C2 (0.27)+SO_C1C2 (0)+SO_C2V (0)	260
2	MOA_C2V (0.35)+POA_C1C2 (0.34) +MOA_C1C2 (0.19)+SO_C1C2 (0.12)	261
3	MOA_C2V (0.45)+POA_C1C2 (0.29)+MOA_C1C2 (0.26)+SO_C2V (0)	261
4	MOA_C2V (0.43)+MOA_C1C2 (0.29) +POA_C1C2 (0.28)	257
5	MOA_C1C2 (0.56)+POA_C1C2 (0.44)	247
6	MOA_C2V (0.57)+MOA_C1C2 (0.43)	252
7	MOA_C2V (0.51)+POA_C1C2 (0.49)	258
8	**MOA_C1C2 (1)**	**239**
9	POA_C1C2 (1)	252
10	**MOA_C2V (1)**	**242**
11	MOA_C1C2 (0.54)+POA_C1C2 (0.44) +SO_C1C2 (0.03)	248
12	**MOA_C1C2 (0.82)+SO_C1C2 (0.18)**	**240**
13	POA_C1C2 (1)+SO_C1C2 (0)	259
14	SO_C1C2 (1)	259
15	**MOA_C2V (0.77)+SO_C2V (0.23)**	**234**
16	SO_C2V (1)	259
17	NAD_C2V (0.56)+NAD_C1C2 (0.44)	278
18	NAD_C1C2 (1)	291
19	**NAD_C2V (1)**	**242**
20	NAD_product (1)	290

- The best model is characterized by the lowest mean squared error (MSE).
- The values in brackets provide the estimated importance of variables.

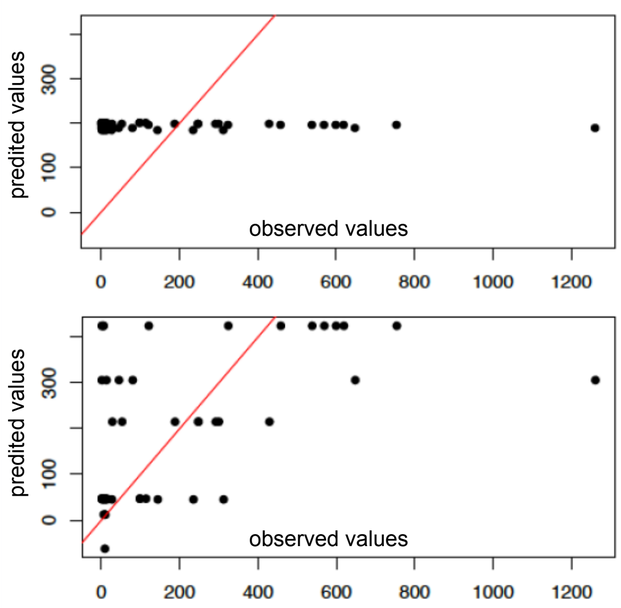
Data

Cl	Freq	Cl	Freq	Cl	Freq
ʃt	1261	bl	248	ps	27
pʀ	754	gl	247	sv, sm	14
ʃp	648	tsv	235	sp	13
gʀ	619	pl	188	ks, bj, gn	10
fʀ	599	kv	144	sts	9
kʀ	569	ʃʀ	121	sn, tj	8
tʀ	538	ʃn	114	fj, sʀ	6
bʀ	458	ʃm	99	tv	5
kl	429	kn	98	vʀ	4
dʀ	324	sk	80	sf, gm	2
ʃv	312	pfl	53	pn, ʃk, km, pfʀ	1
fl	300	st	45		
ʃl	292	sl	28		

- *Leipziger Wortschatz-Portal*
- Cl = 46 CC cluster types
- Freq = type frequency of all words starting with a cluster

Results

linear regression



XGBoost

- Scatterplots present the primacy of the XG Boost method on the example of model (15) that includes MOA(C2V) and SO(C2V).
- The predicted values of XGBoost are much closer to the observed values.

M	Linear regression	XGBoost
8	251.69	238.94
10	264.93	241.84
12	260.11	239.66
15	273.22	234.31
19	266.14	241.84

- Additionally, MSE values for the remaining best models are lower using the XGBoost method.

Conclusions

1. XGBoost seems to be an adequate method for probing hypotheses on the structure – usage hypotheses.

2. The findings offer a starting point for introducing weights to NAD, i.e. increasig the **weight of MOA and SO distances** (and possibly eliminating POA distances).

- Model (15) includes MOA(C2V) and SO(C2V). Other best models partially overlap with (15): (8, 12) include MOA(C1C2), and (10, 19) include MOA(C1C2) and/or SO(C1C2).
- Models with POA(C1C2) have a greater MSE.

3. The results testify to the relevance of **fine-grained categories** in the study of phonotactics, supporting previous work on the relevance of manner (sonority) distances (Parker 2012; Selkirk 1984), also in German (Wiese & Orzechowska in press).

- Neither NAD product nor cumulative NAD values account for the frequency data.

4. The finding that C-V distances constitute the best correlates of type frequency reflects the universal salience of such a sequence (Maddieson 1999; Ohala 1990).

References

Dziubalska-Kołaczyk, K. 2014. Explaining phonotactics using NAD. Language Sciences 46 (A): 6–17.
Jusczyk, P.W; Luce, P.A. & Charles-Luce, J. 1994. Infants' sensitivity to phonotactic patterns in the native language. J. of Memory and Language 33(5), 630-645.
Leipziger Wortschatz-Portal. Access online: <wortschatz.informatik.uni-leipzig.de/de>.
Maddieson, I. 1999. In search of universals. In: Ohala, J.J., Hasegawa, Y., Ohala, M., Granville, D. & Bailey, A. C. (eds.), Proceedings of the 14th ICPhS, 2521-2528.
Ohala, J.J. 1990. The phonetics and phonology of aspects of assimilation. In: Kingston, J. & Beckman, M. (eds.), Papers in Laboratory Phonology I. Cambridge: CUP, 258- 275.
Orzechowska, P., Dziubalska-Kołaczyk., K. 2022. Gradient phonotactics and frequency: A study of German initial clusters. Italian J. of Linguistics 34(1), 103–138.
Parker, S. 2012. Sonority distance versus sonority dispersion—A typological survey. In: Parker, S. (ed.), The Sonority Controversy. Berlin: Walter de Gruyter, 101–166.
Selkirk, E.O. 1984. On the major class features and syllable theory. In: Aronoff, M. & Oehrle, R.T. (eds.), Language sound structure. Cambridge, MA: The MIT Press, 107–136.
Stites, J., Demuth, K. & Kirk, C. 2004. Markedness vs frequency effects in coda acquisition. In: Brugos, A., Micciulla, L. & Smith, C.E. (eds.), Proceedings of the 28th Annual Boston University Conference on Language Development, 565-576.
Wiese R., Orzechowska, P. (in press). Structure and usage do not explain each other: An analysis of German word initial clusters, Linguistics.

Wydział Anglistyki Uniwersytet im. Adama Mickiewicza w Poznaniu

UAM