Weighted phonetic distances in the evaluation of phonotactics

Paula Orzechowska, Katarzyna Dziubalska-Kołaczyk

Adam Mickiewicz University in Poznań
paulao@amu.edu.pl, dkasia@amu.edu.pl

ABSTRACT

Net Auditory Distance is a phonotactic principle that computes the preferability of consonant clusters based on a universally preferred distribution of phonetic distances between adjacent segments. The distances pertain to place, manner and laryngeal features, all of which are expected to contribute to cluster structure. In this paper, we make an attempt at extending the principle by determining the most optimal combination of weighted distances required to predict the frequency of word-onset consonant groups in German. Statistical modelling using gradient boosted decision trees has revealed that, among others, the manner (sonority) distance of the consonant-vowel transition is the most important predictor of cluster frequency. The findings make it possible to assign weights to distances used in the computation of phonotactic preferability, ultimately leading to a more refined classification of German word-onsets.

Keywords: consonant clusters, phonetic features, weights, frequency, German

1. INTRODUCTION

One of the main themes in phonological research is the relationship between structure and usage. Efforts have been invested in the study of combinations of consonants in the languages of the world, revealing that their distribution patterns can be related to markedness and frequency [1-3]. This relationship has been also investigated for such clusters in Standard German, with markedness defined in terms of phonetic / phonological properties [4-6]. For instance, [6] asked which phonetic distances characterizing neighbouring segments best predict the frequency of word-initial clusters. The distances were derived from the Net Auditory Distance (NAD) principle (Sect. 2) that measures the size of contrast between adjacent consonants (C) and vowels (V) in terms of the place of articulation, manner of articulation and laryngeal features. The results of the study suggest that type frequency increases with an increase in the manner of articulation distance between two consonants (see [7] for similar results).

In this paper, we expand on the previous study by analysing the same dataset using different statistical methods. Although linear regression has been used to model the relationship between structure and usage, it is not always best-fitted to the nature of frequency data. Outliers lower the efficiency of linear models; [6] point to the lack of normality in residuals caused numerous low-frequency clusters and high-frequency /ʃt/. Thus, this analysis compares the predictions of linear regression and XGBoost, a technique used to model non-normal distributions.

In the following sections, we introduce the phonotactic model (2), propose a method of weighing distances based on XGBoost values (3), and discuss implications for a new weighted version of NAD (4).

2. PHONOTACTIC MARKEDNESS

2.1. Net Auditory Distance

NAD [8] is a measure of auditory distances between pairs of segments in a cluster, proposed in the framework of Natural Phonology [9,10]. Phonotactic preferability is computed based on well-formedness conditions incorporated from higher linguistic levels.

NAD is grounded in perceptual contrast, following the psychological principle of figure and ground [11]. It assumes that segments forming a universally preferred cluster should be sufficiently different from each other in order to be clearly perceived. Perceptual clarity is best achieved by the combination of a quieter C and a louder V [12,13] for two reasons. First, auditory cues in CV transition are more robust and richer in place cues than in VC [14,15]. Additionally, CV interface facilitates more precise articulation compared to VC [16].

The relevance of phonetic contrast is phonologically captured by the principle of clarity of perception [9], suggesting that larger phonetic distances between adjacent segments are expected to facilitate perception. This is reflected in NAD, where markedness conditions are based on a well-defined arrangement of net distances between pairs of consonants and vowels neighbouring on them in word-initial, medial and final (C)CCs.

2.2. Computation of individual distances

Calculations are computed over 3 types of distances: the manner (MOA) and place (POA) of articulation as well as the sonorant/obstruent (SO) contrast, along the values in Table 1 and well-formedness conditions. The condition for a preferred initial CC requires that the distance between two consonants be larger than or equal to the distance between a consonant and the following vowel.

NAD(C1C2) ≥ NAD(C2V), where:

NAD(C1C2) = |(MOA1-MOA2)| + |(POA1-POA2)| + |SO|, and NAD(C2V) = |(MOA1-MOA2)| + |SO|.

In NAD, MOA distances correspond to sonority distances [17], which rise by 1 from a vowel towards the least sonorous plosives. POA distances, in turn, relate to the segments' anatomical location in the vocal tract [18]. The SO distance captures the sonorant/obstruent distinction. Distance=0 specifies S+S and O+O sequences, while distance=1 holds for sequences that belong to different classes.

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)

Table 1: NAD distances for German (capitals refer to Stop, Fricative, Affricate, Nasal, Liquid, Glide).

For example, the NAD computation for /bj/ involves: C1C2=|5-1|+|1-3|+|1|=7, and C2V=|1-0|+0=1. The cluster meets the well-formedness condition (7≥1), and is thus preferred. The total distance value for a cluster, referred to as 'NAD product', is an index expressing the degree of cluster preferability. NAD product is calculated by means of subtraction: NAD(C1C2)-NAD(C1V), and amounts to 6 for /bj/. The larger the NAD product value, the better the cluster. (For a detailed exposition of other conditions and more detailed NAD categories see [8].)

The numerical values assigned to each consonant reflect the fact that various languages may require phonological distinctions that are more or less fine-grained. Thus, in line with the sonority scale in [19], the most recent version of NAD contains 12 MOA classes for consonants and 5 classes of vowels based on height and peripherality (see [20] for an overview of the revised NAD version).

3. ANALYSIS

3.1. Data

The data analysed in this study were drawn from [4], who extracted a cluster list from phonological descriptions, coursebooks, dictionaries [21-23]. Some clusters are found in rare words, proper nouns and loans, e.g. /bj skv/ (*Björn*, *Squash*). For reasons of space, the proposed analysis is based only on CC types, as presented in Table 2.

Clusters (n=46)
bj bl bʀ dʀ fj fl fʀ gl gm gn gʀ kl km kn ks kʀ kv pfl pfʀ pl pn ps pʀ sf sk sl sm sn sp sʀ st sʦ sv ʃk ʃl ʃm ʃn ʃp ʃʀ ʃt ʃv tj tʀ tv ʦv vʀ

Table 2: Initial CC clusters in German.

Table 3 presents type frequency of each cluster, i.e. the cumulative frequency of all words starting with this cluster. The data were drawn from the corpus of newspaper texts *Leipziger Wortschatz-Portal* [24], which contains 1.65 million word types. In rows with multiple clusters, the frequency value refers to an individual CC.

Cl	Freq	Cl	Freq	Cl	Freq
ʃt	1261	bl	248	ps	27
pʀ	754	gl	247	sv, sm	14
ʃp	648	ʦv	235	sp	13
gʀ	619	pl	188	ks, bj, gn	10
fʀ	599	kv	144	sʦ	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		

Table 3: Type frequency of initial CC clusters.

3.2. Procedure

We ran a series of analyses using linear regression and XGBoost models, and the R environment [25]. The goal was to identify those independent variables that best predict type frequency (i.e. variables generating models with the lowest mean squared error). We compared linear regression with models using XGBoost [26], i.e. gradient-boosted decision trees. Both techniques make it possible to find relations between variables but linear regression is limited to linear dependencies, while XGBoost is also sensitive to non-linear relations. Informally speaking, the latter method divides data into homogenous groups depending on the values of the dependent variable. The contribution of a variable to a model is expressed with a numerical value, indicating the variable's importance. The comparison of models was based on the mean squared error obtained through 5-fold cross-validation process (C-V MSE). A complete list of tested models M (numbered 1-20) is given in Table 4. The MSE values are rounded to the nearest tens.

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

Table 4: Independent variables constituting tested models. The estimated importance of each variable in XGBoost is given in brackets. The best models (with the lowest C-V MSE values) are marked in bold type.

Type frequency was coded as a dependent variable. The tested models involved different combinations of independent predictors derived from the well-formedness condition in Sect. 2.2, i.e. distances for pairs of segments (MOA_C1C2, POA_C1C2, SO_C1C2, MOA_C2V, SO_C2V), summated distances for pairs of segments (NAD_C1C2, NAD_C2V) and NAD product.

3.3. Results

In this section, we present XGBoost analyses for CC word onsets based on 5 best XGBoost models, and compare them with linear regression models.

The best model (15, C-V MSE=234.31) features two variables; MOA and SO distances for the C2V transition. The model constitutes a better fit to the frequency data than linear regression (C-V MSE=273.22). The primacy of the non-linear model is visualized below.

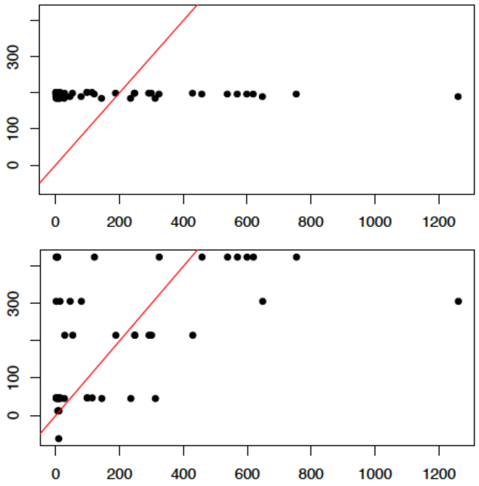

Figure 1: Scatterplots presenting observed vs. expected values for linear regression (top) and XGBoost (bottom) using MOA_C2V and SO_C2V variables (model 15). In both plots, the diagonal line represents the line of perfect fit.

Figure 1 juxtaposes the distribution of the predicted (y axis) and observed (x axis) values in linear regression and XGBoost. As can be seen, the predicted values of XGBoost are much closer to the observed values. All expected values of linear regression lie within a very narrow range (ca. 180-210), which is not in line with the observed values (most of which range from 1 to ca. 600). The XGBoost model is much more flexible, displaying a greater variety of expected values (mostly in the range from 1 to ca. 400). The strength of XGBoost lies in the lack of assumption on linearity, and its ability to capture non-linear dependencies. Model 1 5 captures the contribution of MOA_C2V and

SO_C2V, whereby the importance of the manner distance (77%) outweighs the SO distance (23%).

Other models with low C-V MSE values are listed under (8, 10, 12, 19). The summary of the prediction error obtained through cross-validation for the models is given in Table 5. In each case, the decision tree method provides better results. For clarify of presentation, we also include model 15.

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

Table 5: Comparison of the CV-MSE measures for 5 best linear regression and XGBoost models.

4. TOWARDS WEIGHTED AVERAGE NAD

The findings of the present study offer a starting point for introducing weights to the NAD principle. The analyses have revealed not only a novel statistical method that can be successfully used in determining such weights but also a selection of distances that could be possibly eliminated from future NAD calculations due to being of little use in predicting type frequency.

The model that best predicts the frequency of CCs includes two predictors: MOA_C2V and SO_C2V, among which the manner distance plays a prime role (ca. 80%). Other best models in Table 5 partially overlap with model 15: (8, 12) are also based on MOA_C1C2, and the remaining models (10, 19) include MOA_C1C2 and/or SO_C1C2.

It must be emphasized that models including a wider range of variables tend to display a greater C-V MSE compared to simpler models (e.g. 2, 3). This holds true particularly for models that contain POA_C1C2 and/or SO_C1C2, SO_C2V. In fact, the importance of the latter two variables tends to be low or none in models with three or more independent variables (1-3) or models using POA (9, 13). This suggests that information provided by SO duplicates information arising from POA, but can supplement the information from MOA.

Finally, cumulative distances were shown to be poor predictors of cluster frequency. NAD_C1C2 and NAD product are characterized by the highest C-V MSE values.

5. CONCLUSIONS

The findings of the study support previous theoretical and empirical work on phonotactics. First of all, they suggest that manner (or sonority) distances [27-29] constitute a relevant phonotactic primitive and motivate the core structure of word-initial clusters in German. A similar observation was made in an independent study by [7]. Using the same data and a different set of predictors, [7] demonstrated that manner of articulation distances and voicing constitute the backbone structure of German clusters. That is, average frequency values of clusters were shown to be higher for clusters displaying larger manner distances and contrast in voicing. These observations go in line with [30] arguing that an adequate description of phonotactic possibilities of a language requires operating on weighted phonological features.

Moreover, the results, along with the previous contributions, testify to the relevance of fine-grained categories in the study of phonotactics. Neither NAD product nor cumulative NAD values were shown to account for the frequency data. This observation might also explain why the sonority slope of clusters, usually expressed in binary terms as well-formed/ill-formed, preferred/dispreferred, turns out to be an insufficient criterion in accounting for complex sequences of segments. In turn, subtle sub-segmental cues and gradient classification of structures are expected to be more informative, as argued in [30].

Previous work [7] using the same dataset revealed the importance of CC distances. Here, we have showed that MOA and SO distances for the consonant-vowel transition constitute the best correlates of type frequency. The CV transition, representing an alternation of a quieter C and louder V, is the most salient sequence cross-linguistically [13, 14]. This observation is captured by the distance matrix in [20]: in CCV, the sequence of consonants should display larger contrast compared to CV.

Finally, let us note that type frequency should be viewed as reflecting the phonotactic potential of a language. The relevance of the manner of articulation might also suggest a direction through which the phonotactic inventory of German has developed historically (see also [31] for similar results), and a way in which it might expand. The question on how every-day usage shapes cluster inventories (token frequencies) will be addressed in another study.

6. REFERENCES

[1]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.

[2] 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.

[3] Van de Vijver, R., Baer-Henney, D. 2012. Sonority intuitions are provided by the lexicon. In: Parker, S. (ed.), *The Sonority Controversy*. Berlin: de Gruyter, 195-215.

[4] Orzechowska, P., Wiese, R. 2015. Preferences and variation in word-initial phonotactics: A multi-dimensional evaluation of German and Polish. *Folia Linguistica* 49, 439-486.

[5] Ulbrich, C.; Alday, P., Knaus, J., Orzechowska, P., Wiese, R. 2016. The role of phonotactic principles in language processing. *Language, Cognition and Neuroscience* 31(5), 662-682.

[6] Orzechowska, P., Dziubalska-Kołaczyk., K. 2022. Gradient phonotactics and frequency: A study of German initial clusters. *Italian J. of Linguistics* 34(1), 103–138.

[7] Wiese R., Orzechowska, P. (in press). Structure and usage do not explain each other: An analysis of German word-initial clusters, *Linguistics*.

[8] Dziubalska-Kołaczyk, K. 2019. On the structure, survival and change of consonant clusters. *Folia Linguisitca Historica* 40(1), 107-127.

[9] Donegan, P. J. & Stampe, D. 1979. The study of natural phonology. In: Dinnsen, D.A. (ed.), *Current Approaches to Phonological Theory*. Bloomington: Indiana University Press, 126-173.

[10] Dressler, W.U. 1985. *Morphonology: The dynamics of derivation*. Ann Arbor: Karoma Publishers.

[11] Dressler, W. U. 2009. Natural phonology as part of natural linguistics. *Poznań Studies in Contemporary Linguistics* 45(1), 32-42.

[12] Greenberg, J. H. 1978. Some generalizations concerning initial and final consonant clusters. In: Greenberg, J. H. (ed.), *Universals of Human Language*. Stanford: Stanford University Press, 243-279.

[13] Maddieson, I. 2013. Syllable structure. In: Dryer, M., Haspelmath, M. (eds.), *The World Atlas of Language Structure Online*. Munich: Max Plank Digital Library.

[14] 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.

[15] 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.

[16] Ohala, J.J. & Kawasaki, H. 1984. Prosodic phonology and phonetics. *Phonology Yearbook* 11, 113-129.

[17] Goldsmith, J.A. 1990. *Autosegmental and Metrical Phonology*. Blackwell.

[18] Ladefoged, P. 2006. *A Course in Phonetics* (5th ed.). Boston: Heinle & Heinle.

[19] Parker, S. 2008. Sound level protrusions as physical correlates of sonority. *J. of Phonetics* 36, 55-90.

[20] Dziubalska-Kołaczyk, K., Pietrala, D. 2020. The NAD Phonotactic Calculator – an online tool to calculate cluster preferability across languages. In: Sauer, H., Chruszczewski, P.P. (eds.). *Mostly Medieval. In Memory of Jacek Fisiak.* San Diego: Academic Publishing, 445-458.

[21] Meinhold, G., Stock, E. 1980. *Phonologie der Deutschen Gegenwartssprache*. Leipzig: VEB Bibliographisches Institut Leipzig.

[22] Duden Aussprachewörterbuch 1990 / 2009 (6th ed.). *Duden: Wörterbuch der deutschen Standardaussprache*. Mannheim / Wien / Zürich: Dudenverlag.

[23] Wiese, R. 2000. *The Phonology of German* (2nd ed). Oxford: Clarendon Press.

[24] *Leipziger Wortschatz-Portal*. Access online: <wortschatz.informatik.uni-leipzig.de/de>.

[25] R Core Team, R: *A language and environment for statistical computing*, R foundation for statistical computing, Vienna, Austria, 2020.

[26] Chen, T., Guestrin C. 2016. XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining,* 785-794.

[27] Parker, S. 2012. Sonority distance versus sonority dispersion—A typological survey. In: Parker, S. (ed.), *The Sonority Controversy.* Berlin: Walter de Gruyter, 101–166.

[28] 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.

[29] Steriade, D. 1982. *Greek prosodies and the nature of syllabification*. PhD diss., Cambridge, MA: The MIT Press.

[30]Orzechowska, P. 2019. *Complexity in Polish phonotactics: On features, weights, rankings and preferences*. Springer.

[31] Baumann, A., Wissing, D. 2018. Stabilizing determinants in the transmission of phonotactic systems: Diachrony and acquisition of coda clusters in Dutch and Afrikaans. *Stellenbosch Papers in Linguistics Plus* 55, 77-107.