Onset Cluster Typologies

Stephen Jay and Steve Parker
Dallas International University

12th Annual DFW Metroplex Linguistics Conference
November 6, 2020

Abstract
Most formal models of phonology invoke sonority to explain the set of consonant clusters allowed in specific languages. While such proposals account for many languages, they ultimately prove too weak when confronted with more exhaustive data sets. Part of the problem is that theories of onset cluster typology have not been tested against inventories from hundreds of languages. When larger samples are considered, formal models fail to accommodate the attested combinations of syllable-initial consonant clusters ($_{\sigma}[C_1C_2V]$) in many languages.

Previous approaches to the phonotactics of consonant clusters include the following:

- MSD: Minimum Sonority Distance (Steriade 1982, Selkirk 1984, Levin 1985)
- SD: Sonority Dispersion (Clements 1990)
- SR: Sonority Rise (Flemming 2008)
- SA: Sonority Angle (Fullwood 2014)

Each of these models calculates the gradient harmony of consonant clusters in terms of a different mathematical formula based on the distance between the sonority indices of C_1, C_2, and sometimes the following vowel. However, none of them are able to handle the full range of empirical facts. For example, MSD favors obstruent+glide (OG) clusters, so it cannot produce a language where C_2 must be a liquid without additional mechanisms. Conversely, SD evaluates obstruent+liquid onsets (OL) as unmarked, so it cannot generate languages where C_2 is always a glide. All of these approaches are therefore partially right, yet partially wrong.

In this paper we highlight a new model called Minimum Distance to Offset (MDO). Like MSD approaches it sets a minimum threshold for sonority distance between C_2 and C_1 in each language. However, it further arranges onset cluster types into several continua with a constant C_2 in each scale. For example, the glide offset continuum consists of a fixed ranking encoding the universal preference OG $\succ$ NG $\succ$ LG $\succ$ GG.

Each of these competing models is compared against a database listing the inventory of permissible onset clusters in over 300 languages worldwide. The results indicate that previous approaches massively undergenerate the attested typological combinations and are therefore too restrictive. In contrast to this, the MDO proposal successfully accounts for the full array of language types. While the flexibility inherent in the MDO model is necessary, we argue that it is not overly powerful. We note many hypothetical language types it cannot produce.

We conclude by showing how the preference scales of the MDO approach can be implemented using markedness constraint families in a stringency relationship (de Lacy 2004, 2006). For instance, the following constraints target increasingly more inclusive combinations of onset clusters, with C_2 fixed as a glide: *{GG}, *{GG,LG}, *{GG,LG,NG}, *{GG,LG,NG,OG}. This subset of the model allows for languages in which C_2 can only be a liquid, nasal, obstruent, or combination thereof.

1. Overview and highlights

- Sonority-based models of consonant clusters account for many languages.
- Drawing from a database of over 300 languages, we summarize possible inventory combinations of onset clusters cross-linguistically.[1]
- When evaluated against this fuller range of typological facts, current models undergenerate the attested combinations in many languages.
- We propose a partially new approach – Minimum Distance to Offset – which produces the great majority of all known language types, without needing additional formal devices.
- For example, the glide offset continuum posits the following scale of relative harmony: OG ≻ NG ≻ LG ≻ GG.
- This model can be implemented in Optimality Theory using constraint families in a stringency relationship. To illustrate, the asymmetrical entailments in this scale are produced by the following freely-permutable constraints: *{GG}, *{GG,LG}, *{GG,LG,NG}, *{GG,LG,NG,OG}.

2. The universal sonority scale

(1) Universal sonority scale, modal version (Clements 1990, Kenstowicz 1994, Smolensky 1995):

natural class:	vowels	>	glides	>	liquids	>	nasals	>	obstruents
abbreviation:	V		G		L		N		O
sonority index (SI):	5		4		3		2		1

Table 1: Exhaustive list of 16 possible combinations of consonant clusters, grouped by sonority class (the numbers in parentheses indicate the sonority differential (SD) between C_2 and C_1) (Parker 2012:108)

		first consonant			
		obstruent	nasal	liquid	glide
second consonant	obstruent	OO (0)	NO (–1)	LO (–2)	GO (–3)
	nasal	ON (1)	NN (0)	LN (–1)	GN (–2)
	liquid	OL (2)	NL (1)	LL (0)	GL (–1)
	glide	OG (3)	NG (2)	LG (1)	GG (0)

(2) (a) sonority rises, 6 clusters (enclosed in dark bold borders): ON, OL, OG, NL, NG, LG [Sonority Differential > 0]. Called *core* or *rising clusters*.
(b) sonority equal, 4 clusters (diagonally from top left to bottom right): OO, NN, LL, GG [Sonority Differential = 0]. Called *plateau clusters* or *plateaus*.
(c) sonority falls, 6 clusters (in shaded cells): NO, LO, LN, GO, GN, GL [Sonority Differential < 0]. Called *reversed clusters* or *reversals*.

[1]We gratefully acknowledge several interns and students who helped compile this database: Katherine Bare, Timothy Palmer, Moriah Rose, Lydia Stebbins, and Matthew Woods.

3. Limitations of our study

In this paper we purposely do not address several issues:

- coda clusters. Our impression is that syllable-final combinations do not follow the predictions of sonority-based models as consistently as onsets do.
- tri-consonantal onsets with a monotonic sonority rise, such as /klwV/. These are quite rare.
- word-initial sequences involving reversed sonority, such as /lpV/ or /spV/ (the notorious problem of *[sC]itis*). Despite frequent claims that these violate the Sonority Sequencing Principle (SSP), in most cases there is no empirical evidence that these are tautosyllabic. Rather, many kinematic studies show that the initial segment is not part of the same prosodic constituent as the [CV] string (Gafos *et al*. 2014, 2020; Hermes *et al*. 2013; Shaw and Gafos 2015; Shaw *et al*. 2011). Consequently, these are best analyzed as an extra-syllabic appendix which does not violate the SSP *stricto sensu* (Vaux and Wolfe 2009).

4. The problem

Previous approaches to the phonotactics of consonant clusters:

- MSD: Minimum Sonority Distance (Steriade 1982, Selkirk 1984, Levin 1985)
- SD: Sonority Dispersion (Clements 1990)
- SR: Sonority Rise (Flemming 2008)
- SA: Sonority Angle (Fullwood 2014)

Each of these models calculates the gradient markedness of consonant clusters in terms of a different mathematical formula based on the distance between the sonority indices of C_1, C_2, and sometimes the following vowel.

For example, MSD, SR, and SA all favor obstruent + glide (OG) onset clusters. However, there are many languages which only have OL onset clusters.

On the other hand, SD evaluates obstruent + liquid onsets (OL) as unmarked, which correctly predicts OL-only languages. However, there are many languages which have OG-only clusters.

Furthermore, the typological predictions of such approaches have never been submitted to a systematic confirmation based on a statistically robust sample of languages.

Ultimately, none of these approaches are able to handle the full range of empirical facts, so they are too weak. All of them are partially right, yet partially wrong.

5. Proposed model

Our approach has the flexibility to handle both liquid offset languages (CL only) and glide offset languages (CG only):

Overview of Minimum Distance to Offset (MDO) model
(Parker 2016, 2017)

	3		2		1	(sonority distance)
glide offset continuum	OG	≻	NG	≻	LG	
liquid offset continuum			OL	≻	NL	
nasal offset continuum					ON	

Some generalizations and claims about this model:

- The set of permissible two segment onset clusters in any given language can be formed by combining a continuous group of one or more members of decreasing sonority distance from any of these offset continua.

Nevertheless,

- The relative markedness of a cluster increases as the sonority distance decreases (left to right). This is expected.
- Furthermore, the markedness of a cluster also increases as the relative sonority of its offset decreases (top to bottom). This may be a novel finding.

The upshot of this approach is that it predicts both OG-only and OL-only languages, while maintaining that OG is universally more harmonic than OL, all else being equal.

6. Summary of previous approaches

(a) Minimum Sonority Distance (MSD)
(Steriade 1982, Selkirk 1984, Levin 1985)

Assuming the Sonority Indices in (1), and given an onset composed of two segments, C_1 and C_2, if $a = SI(C_1)$ and $b = SI(C_2)$, and if $a \leq b$, then the language-specific MSD = x such that $b - a \geq x$, where $x \in \{0, 1, 2, 3\}$ (Parker 2012:110).

MSD=3	MSD=2	MSD=1	MSD=0
OG	OG,OL,NG	OG,OL,NG,LG,NL,ON	OG,OL,NG,LG,NL,ON,GG,LL,NN,OO

This evaluates OG as universally unmarked because the sonority distance between C_1 and C_2 is the largest possible. But this model cannot produce a liquid offset language (OL, or OL + NL).

(b) Sonority Dispersion Principle (Clements 1990)

$$D = \sum_{i=1}^{m} \frac{1}{d_i^2}$$

where d = distance between the sonority indices of each pair of segments
m = number of pairs of segments (including nonadjacent ones), where
$m = n\,(n-1)\,/\,2$, and where n = number of segments

D=0.56	*D*=1.17	*D*=1.36	*D*=2.25	*D*=undefined
OLV	OGV, ONV	NGV, NLV	LGV	GG, LL, NN, OO

This evaluates OL as universally unmarked because the sonority distances between C_1, C_2, and the vowel are evenly and maximally dispersed. But it cannot produce glide offset languages (CG only).

(c) Sonority Rise (Flemming 2008)

A perceptually-based model for predicting the propensity of a cluster to be repaired by vowel epenthesis. It is not presented as a theory for generating inventories of onset clusters universally. Nevertheless, the scale it produces can easily be adapted for such a purpose.

$$SR = 1 - \frac{C_2 - C_1}{V - C_1}$$

SR=0.25	*SR*=0.33	*SR*=0.50	*SR*=0.67	*SR*=0.75	*SR*=1.00
OG	NG	OL, LG	NL	ON	GG, LL, NN, OO

(d) Sonority Angle (Fullwood 2014)

Similar to the Sonority Rise approach in (c) above. It differs from SR in favoring OL over LG (a good result). SA also distinguishes between the four plateau clusters, unlike SR.

$$SA = \arctan(V - C1) - \arctan(C2 - C1)$$

SA=0.08	*SA*=0.14	*SA*=0.22	*SA*=0.32	*SA*=0.46	*SA*=0.54	*SA*=0.79	*SA=1.11*	*SA*=1.25	*SA*=1.33
OG	NG	OL	LG	NL	ON	GG	LL	NN	OO

Sonority Rise and Sonority Angle converge on evaluating OG as better than OL, as in the MSD model. However, they both have the problem of preferring NG over OL.

7. Results

A database compiled at GIAL and DIU lists over 1,200 languages known to contain onset clusters.

Of these, around 330 have been analyzed to tabulate their inventory of syllable-initial consonant clusters in terms of sonority. This sample of languages has not been tested for genetic or areal bias. However, at this point we are not making any claims about the relative frequency of particular language types. Rather, we are simply establishing which types exist, or seem not to exist. So this is not a major confound.

Among languages which permit onset clusters, the 6 core or rising sonority types (OG, OL, ON, NG, NL, LG) can hypothetically be arranged into a maximum of 63 possible combinations consisting of one to six types.

Of these 63 hypothetical combinations, 22 language types clearly exist in our sample, while 41 do not. The predictions of the five competing models for these 22 attested inventories are as follows:

Table 2: Number of language types compatible with each model, in terms of permissible combinations of core onset clusters (out of 63 total possibilities)

	Minimum Distance to Offset	Minimum Sonority Distance	Sonority Dispersion	Sonority Rise	Sonority Angle
attested types (22)	22	3	8	7	6
non-attested types (41)	1	0	0	0	0

In Table 2, the MDO approach successfully accounts for all 22 onset combinations which occur among our sample of 329 languages. The other four approaches fail to capture many of these systems.

When we add the four plateau clusters (to the six rising sonority types) and recalculate the factorial typology, the total number of possible combinations (1-10 types) increases to 1023. Of these, 66 language types are documented in our sample. The remaining 957 combinations are either missing completely or problematic (due to incomplete data or questionable analyses).

Our approach is designed to easily capture these attested languages as well:

Minimum Distance to Offset (MDO) model, fuller version
(Parker 2016, 2017)

	3		2		1		0	(sonority distance)
glide offset continuum	OG	≻	NG	≻	LG	≻	GG	
liquid offset continuum			OL	≻	NL	≻	LL	
nasal offset continuum					ON	≻	NN	
obstruent offset continuum							OO	

Table 3: Number of language types compatible with each model, in terms of permissible combinations of core plus plateau onset clusters (out of 1023 total possibilities)

	Minimum Distance to Offset	Minimum Sonority Distance	Sonority Dispersion	Sonority Rise	Sonority Angle
attested types (66)	51	14	8	12	6
non-attested types (957)	62	11	3	9	4

In terms of the fuller set of 10 cluster types (core + plateau) analyzed in Table 3, the MDO model again accounts for a much larger proportion of attested language types than its competitors. These other four approaches are clearly too weak. On the other hand, the MDO approach also predicts a greater number of unattested combinations. This is not nearly as serious a problem, however, since many of these gaps could be accidental. Given more exhaustive data from the world's languages, most of these missing cases might eventually be filled in.

Among the 15 attested language types which the MDO model fails to predict, most of these involve complications such as clusters limited to word-initial position. We suspect that with further scrutiny, many of these potential counterexamples will in fact disappear as valid cases.

As noted in §6, a major problem for three of the four previous models (MSD, SR, and SA) is that they cannot generate CL-only languages since they evaluate OG as universally unmarked. This predicts that it should be "selected" first in any language allowing clusters. Consequently, they require additional machinery (unrelated to sonority) to account for inventories which are limited to just OL, or OL + NL. Some examples of these two types of systems are listed in (4e) below.

On the other hand, the Sonority Dispersion Principle posits OL as universally preferred, so it cannot handle CG-only languages. Many examples of these are illustrated in (4a-d).

8. OT implementation

The stringency approach popularized by de Lacy (2004, 2006) for sonority-based stress attraction provides an elegant way to capture the four offset continua of the MDO model (§5). In his proposal, constraints can be freely ranked (permuted). We posit the following families of markedness constraints which target increasingly more inclusive combinations of onset clusters, grouped according to a fixed offset consonant:

(3) (a) *{GG}, *{GG,LG}, *{GG,LG,NG}, *{GG,LG,NG,OG} [glide offset continuum]
(b) *{LL}, *{LL,NL}, *{LL,NL,OL} [liquid offset continuum]
(c) *{NN}, *{NN,ON} [nasal offset continuum]
(d) *{OO} [obstruent offset continuum]

For example, if the constraint *{GG,LG,NG,OG} in (3a) is highly ranked, it rules out glide offset clusters altogether. This produces languages in which C_2 can only be a liquid, nasal, obstruent, or combination thereof.

By ranking the relevant faithfulness constraints at different points among these constraints, various types of glide offset inventories are produced:

(4) (a) **FAITH** » *{GG}, *{GG,LG}, *{GG,LG,NG}, *{GG,LG,NG,OG}
result: OG, NG, LG, and GG (example languages: Shilluk, Kamba)

(b) *{GG} » **FAITH** » *{GG,LG}, *{GG,LG,NG}, *{GG,LG,NG,OG}
result: OG, NG, and LG, but not *GG (example languages: Ga'dang, Western Parbate Kham)

(c) *{GG}, *{GG,LG} » **FAITH** » *{GG,LG,NG}, *{GG,LG,NG,OG}
result: OG and NG, but not *LG or *GG (example languages: Angataaha, Kenyang)

(d) *{GG}, *{GG,LG}, *{GG,LG,NG} » **FAITH** »*{GG,LG,NG,OG}
result: OG, but not *NG, *LG, or *GG (example languages: Bambassi, Dadibi, Zoque)

(e) *{GG}, *{GG,LG}, *{GG,LG,NG}, *{GG,LG,NG,OG} » **FAITH**
result: no glide offset clusters whatsoever (example languages: Eastern Kayah Li, Emberá-Catío, Kalasha, Parauk Wa, Eastern Katu, Isirawa, Ngarinyin, Yamomámi)

Furthermore, this model can easily be extended to include or exclude reversed sonority clusters as well. Thus, the complete expansion of the obstruent offset continuum in (3d) is as follows:

(5) Obstruent offset continuum, exhaustive version (including reversals; cf. 3d):

*{GO}, *{GO,LO}, *{GO,LO,NO}, *{GO,LO,NO,OO}

9. Conclusion

Previous accounts of onset cluster phonotactics are clearly inadequate as universal statements of which types of languages do and do not exist, so they must be rejected. In their place we need a new model. We have proposed the Minimum Distance to Offset approach, which successfully covers much more empirical ground. It may not match the typological variety of attested onset inventories perfectly, but it is clearly an improvement. So we are moving in the right direction. But is it too powerful?

In the interest of showing what *cannot* be done with this approach, we list several languages types which it systematically cannot produce. These consist of any one or more clusters starting from the marked end of an offset continuum, without the less marked members (such as a language with GG only). Or any combination of clusters on a continuum that skips over intervening members, e.g., OG plus LG without *NG.

Table 4: Language types predicted not to exist, in terms of permissible onset clusters

1.	GG only
2.	LG (only)
3.	NG (etc.)
4.	GG, LG
5.	GG, NG
6.	GG, OG
7.	LG, NG
8.	LG, OG
9.	GG, LG, NG
10.	GG, LG, OG
11.	GG, NG, OG
12.	LL
13.	NL
14.	LL, NL
15.	LL, OL
16.	NN
17.	any combination of {1-16}

In our database there are no compelling cases of languages having any of these inventories. There are a few marginal "hits," but all of these involve complications such as non-canonical clusters types which occur only in loanwords, ideophones, etc. So their status is questionable.

References

Clements, George. 1990. "The Role of the Sonority Cycle in Core Syllabification." In *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, edited by John Kingston and Mary Beckman, 283–333. Cambridge: Cambridge University Press.

de Lacy, Paul. 2004. "Markedness Conflation in Optimality Theory." *Phonology* 21, no. 2 (August): 145–99.

de Lacy, Paul. 2006. *Markedness: Reduction and Preservation in Phonology*. Cambridge: Cambridge University Press.

Flemming, Edward. 2008. "Asymmetries between Assimilation and Epenthesis." Unpublished manuscript, MIT. PDF.

Fullwood, Michelle Alison. 2014. "The Perceptual Dimensions of Sonority-Driven Epenthesis." In *Supplemental Proceedings of the 2013 Annual Meeting on Phonology*, edited by John Kingston, Claire Moore-Cantwell, Joe Pater, and Robert Staubs. Washington, DC: Linguistic Society of America.

Gafos, A. I., Charlow, S., Shaw, J. A., and Hoole, P. 2014. "Stochastic time analysis of syllable-referential intervals and simplex onsets." *Journal of Phonetics* 44 (Dynamics of Articulation and Prosodic Structure), 152–66.

Gafos, Adamantios I., Jens Roeser, Stavroula Sotiropoulou, Philip Hoole, and Chakir Zeroual. 2020. "Structure in mind, structure in vocal tract." *Natural Language & Linguistic Theory* 38: 43–75.

Hermes, A., Mücke, D., and Grice, M. 2013. "Gestural coordination of Italian word-initial clusters: the case of 'impure s.'" *Phonology* 30: 1–25.

Kenstowicz, Michael. 1994. *Phonology in Generative Grammar*. Malden, MA: Blackwell.

Levin, Juliette. 1985. "A Metrical Theory of Syllabicity." PhD diss., MIT.

Parker, Steve. 2012. "Sonority distance vs. sonority dispersion–a typological survey." In *The Sonority Controversy*, edited by Steve Parker, 101–65. Berlin: De Gruyter Mouton.

Parker, Steve. 2016. "Reconsidering Sonority Dispersion and Liquid vs. Glide Offsets: What Do the Typological Facts Indicate?" CUNY conference on sonority, January 14-15.

Parker, Steve. 2017. "Reconsidering Sonority Dispersion and Liquid vs. Glide Offsets: What Do the Typological Facts Indicate?" *Winak: Revista de Estudios Interculturales* 26: 11–42.

Selkirk, Elizabeth. 1984. "On the Major Class Features and Syllable Theory." In *Language Sound Structure: Studies in Phonology Presented to Morris Halle by His Teacher and Students*, edited by Mark Aronoff and Richard T. Oehrle, 107–36. Cambridge, MA: The MIT Press.

Shaw, J. A., and Gafos, A. I. 2015. "Stochastic time models of syllable structure." *PLoS ONE* 10(5), e0124714.

Shaw, J. A., Gafos, A. I., Hoole, P., and Zeroual, C. 2011. "Dynamic invariance in the phonetic expression of syllable structure: a case study of Moroccan Arabic consonant clusters." *Phonology* 28(3), 455–490.

Smolensky, Paul. 1995. "On the internal structure of the constraint component Con of UG." Handout of a talk presented at UCLA, April 7. Rutgers Optimality Archive 86.

Steriade, Donca. 1982. "Greek Prosodies and the Nature of Syllabification." PhD diss., MIT.

Vaux, Bert, and Andrew Wolfe. 2009. "The appendix." In Eric Raimy and Charles E. Cairns (eds.), *Contemporary Views on Architecture and Representations in Phonology*, 101-43. (Current Studies in Linguistics). Cambridge: MIT Press.