Phonological Data and Analysis Volume 2, Article 5: 1–35 (2020)
https://doi.org/10.3765/pda.v2art5.25

Received 7 March 2019; revised 11 February 2020; accepted 6 June 2020.

Aggressive reduplication and dissimilation in Sundanese

Juliet Stanton*

New York University – stanton@nyu.edu

Most cases of long-distance consonant dissimilation can be characterized as local (occurring across a vowel) or unbounded (occurring at all distances). The only known exception is rhotic dissimilation in Sundanese (Cohn 1992; Bennett 2015a,b), which applies in certain non-local contexts only. Following a suggestion by Zuraw (2002:433), I show that the pattern can be analyzed in a co-occurrence-based framework (Suzuki 1998) by invoking two unbounded co-occurrence constraints, *[r]…[r] and *[l]…[l], whose effects in local contexts are obscured by a drive for identity between adjacent syllables. Statistical trends in the lexicon are consistent with this analysis. I compare the predictions of this analysis to those of Bennett's (2015a,b) and suggest that the present proposal is preferable.

Keywords: dissimilation; aggressive reduplication; phonotactics; lexical statistics

1 Introduction

Most cases of long-distance consonant dissimilation can be characterized as *local* or *unbounded*. In the local cases, alternations occur only across a single vowel (or, alternatively, between adjacent syllables; the difference between these characterizations is not important here). An example of local dissimilation comes from Yimas (Foley 1991), where the inchoative suffix /aɾa/ dissimilates to [ata] given an [ɾ]-final root (1b) but not otherwise (1c).

(1) Local rhotic dissimilation in Yimas (Foley 1991)

a. *Default form of inchoative suffix is* [aɾa]
/pak-aɾa/ → [pak-aɾa] 'break open'

b. *Local dissimilation:* /…ɾ-aɾa/ → […ɾ-ata]
/apɾ-aɾa/ → [apɾ-ata] 'open, spread'

c. *No non-local dissimilation:* /…ɾ…-aɾa/ → […ɾ…-aɾa]
/aɾaŋ-aɾa/ → [aɾaŋ-aɾa] 'tear into pieces'

In the unbounded cases, dissimilation occurs at all distances. An example of unbounded dissimilation comes from Georgian (Fallon 1993), where the adjective-forming suffix /uri/ dissimilates to [uli] given an [r]-containing stem, regardless of that [r]'s distance from the suffix (2b–c).

* For helpful feedback I am grateful to Gillian Gallagher, Maria Gouskova, Megan Crowhurst, and two anonymous reviewers. For questions I thank audiences at MIT, NYU, and the GLOW 48 Phonology Workshop, as well as the Fall 2018 Computational Phonology class at NYU. Special thanks to Abby Cohn for feedback and discussion, plus help with Sundanese resources; and to Eri Kurniawan, Anhari Raushanfikri, and Ivanakbar Purwamaska for help with glosses.

(2) Unbounded rhotic dissimilation in Georgian (Fallon 1993)

a. *Default form of adjective-forming suffix is* [uri]
/svan+uri/ → [svan-uri] 'Svan'

b. *Local dissimilation:* /…r-uri/ → […r-uli]
/asur+uri/ → [asur-uli] 'Assyrian'

c. *Long-distance dissimilation:* /…r…-uri/ → […r…-uli]
/ast'ronomia+uri/ → [ast'ronomia-uli] 'astronomical'

The third logical possibility I will refer to as *non-local-only* dissimilation, where co-occurrence is permitted in a local configuration but not elsewhere. Cases that fit this description are uncommon. The only known case comes from Sundanese, where (simplifying for now) the plural infix /ar/ dissimilates to [al] given the presence of a non-local [r] (3c), but maps to [ar] if an [r] is local (3b).

(3) Non-local-only rhotic dissimilation in Sundanese (Cohn 1992; Bennett 2015a,b)

a. *Default form of plural is* [ar]
/ar+kusut/ → [k-ar-usut] 'messy (pl.)'

b. *No local dissimilation:* /r-ar…/ → [r-ar…]
/ar+rahɨt/ → [r-ar-ahɨt] 'wounded (pl.)'

c. *Long-distance dissimilation:* /…-ar-…r…/ → […-al-…r…]
/ar+ŋumbara/ → [ŋ-al-umbara] 'go abroad (pl.)'

My interest lies in how the Sundanese data bear on predictions of two competing theories of dissimilation. The theories are Suzuki's (1998) Generalized OCP (or GOCP), which treats dissimilation as the result of anti-similarity constraints, and Bennett's (2015) Surface Correspondence Theory of Dissimilation (or SCTD), which treats dissimilation as a way of avoiding similarity-based surface correspondence. Both theories can generate non-local-only dissimilation, but they do so in different ways. Under the GOCP, non-local-only dissimilation is only possible given the interaction of a preference for unbounded dissimilation with an overriding dispreference for the result of local dissimilation. The SCTD, by contrast, provides an explicit provision for non-local-only dissimilation: cases like (3) can be generated directly, without appealing to any dispreference for the result of local dissimilation.

The remainder of this section introduces the GOCP and SCTD and explicates their predictions regarding the character of non-local dissimilation. Following this I show that the Sundanese case, previewed in (3), is consistent with the more restrictive predictions of the GOCP.

1.1 Non-local dissimilation in the GOCP

Suzuki's (1998) GOCP proposes that dissimilation is motivated by constraints of the form **X*…*Y*, where *X* and *Y* are entities whose co-occurrence is dispreferred (for earlier constraint-based analyses of dissimilation see Holton 1995; Alderete 1997; Myers 1997; *a.o.*). Each **X*…*Y* constraint stands for a family of constraints, where "…" denotes intervening material of differing lengths.[1] To explore the theory's predictions regarding non-local dissimilation, we will consider two constraints from the *[-lateral]…[-lateral] family: one that penalizes co-occurring rhotics separated by only a mora (4), and one that penalizes each pair of rhotics occurring within the word (5). (Throughout this paper I assume that laterals are [+lateral], rhotics are [-lateral], and that no other segments are specified for [±lateral].)

[1] Suzuki's proposed hierarchy is **XY* ≫ **X*-C_0-*Y* ≫ **X*-*μ*-*Y* ≫ **X*-*μμ*-*Y* ≫ **X*-*σσ*-*Y* ≫ … ≫ **X*-∞-*Y*. For expositional simplicity I assume only two instantiations of each co-occurrence constraint, **XμY* and **X*…*Y*, where **X*…*Y* penalizes co-occurrence at all distances. For present purposes this simplified variant makes equivalent predictions.

(4) *[-lateral]μ[-lateral] (*[r]μ[r]):
Assign one * for each pair of [-lateral] segments separated by a mora.

(5) *[-lateral]…[-lateral] (*[r]…[r]):
Assign one * for each pair of [-lateral] segments within the word.

A factorial typology of (4–5), together with IDENT-[±lateral] ("assign one violation for each input [αlateral] segment whose output correspondent is [-αlateral]"), predicts two kinds of dissimliation: local (as in Yimas) and unbounded (as in Georgian). Cases of non-local-only dissimilation are not predicted, as neither *[r]μ[r] nor *[r]…[r] penalizes only non-local co-occurrence (6).

(6) Factorial typology of (4), (5), and IDENT-[±lateral]

Dissimilation type	*Ranking(s)*	*/rμr/ →*	*/r...r/ →*
None	IDENT-[±lateral] ≫ *[r]μ[r] ≫ *[r]…[r] IDENT-[±lateral] ≫ *[r]…[r] ≫ *[r]μ[r]	[rμr]	[r…r]
Local	*[r]μ[r] ≫ IDENT-[±lateral] ≫ *[r]…[r]	[rμl]	[r…r]
Unbounded	*[r]μ[r] ≫ *[r]…[r] ≫ IDENT-[±lateral] *[r]…[r] ≫ *[r]μ[r] ≫ IDENT-[±lateral] *[r]…[r] ≫ IDENT-[±lateral] ≫ *[r]μ[r]	[rμl]	[r…l]

Non-local-only dissimilation is however predicted to occur when *[r]…[r] dominates IDENT[±lateral], and is dominated in turn by a constraint that disprefers the consequences of local dissimilation. An example of such a constraint is one that prefers local assimilation for laterality among liquids. I will assume that this constraint can be implemented as *[αlateral]μ[-αlateral] (7), though nothing rests on this formulation.

(7) *[αlateral]μ[-αlateral] (*[αlat]μ[-αlat]):
Assign one * for each pair of [αlateral] and [-αlateral] segments separated by a mora.

When both *[αlat]μ[-αlat] and *[r]…[r] are high-ranked, the resulting system can exhibit non-local-only dissimilation. I illustrate here with two hypothetical prefixed forms, /ra-rata/ and /ra-tara/, which map to [ra-rata] (8) and [ra-tala] (9) given the ranking *[αlat]μ[-αlat] ≫ *[r]μ[r], *[r]…[r] ≫ IDENT-[±lateral] (one of several rankings that can derive these mappings).

(8) Non-local dissimilation in the GOCP: local [r]s do not dissimilate

/ra-rata/	*[αlat]μ[-αlat]	*[r]μ[r]	*[r]…[r]	IDENT-[±lateral]
☞ a. [ra-rata]		*	*	
b. [ra-lata]	*!			*

(9) Non-local dissimilation in the GOCP: non-local [r]s do dissimilate

/ra-tara/	*[αlat]μ[-αlat]	*[r]μ[r]	*[r]…[r]	IDENT-[±lateral]
a. [ra-tara]			*!	
☞ b. [ra-tala]				*

In this hypothetical system we would expect to find evidence of local assimilation outside of this dissimilatory context, because *[αlat]μ[-αlat] dominates IDENT-[±lateral] by transitivity. For example, /ra-lata/ would be predicted to surface unfaithfully as [ra-rata] (10). (This pattern bears a resemblance to what we find in Sundanese, but I use hypothetical forms here to keep the analysis simple.)

(10) More local assimilation

/ra-lata/	*[αlat]μ[-αlat]	*[r]μ[r]	*[r]…[r]	IDENT-[±lateral]
a. [ra-lata]	*!			
☞ b. [ra-rata]		*	*	*

The main point is that non-local dissimilation is not a basic prediction of the GOCP. Rather, it emerges from an interaction of constraints that prefer unbounded dissimilation with others that disprefer the result of local dissimilation. Note that these other constraints need not promote local assimilation, as the role played by *[αlat]μ[-αlat] above can be played by any other constraint that disprefers (8b). To give another example, (8b) could also be ruled out by a positional faithfulness constraint that protects the root-initial segment; in such a case we might expect to find other evidence of positional faithfulness to the root-initial segment. A related case occurs in Zulu (Beckman 1998; Bennett 2015b), where labial palatalization triggered by a suffixed /w/ (11a) fails to apply if the targeted labial is root-initial (11b). External evidence suggesting that root-initial consonants are privileged comes from the larger inventory of consonants licensed initially and the fact that long-distance laryngeal harmony is controlled by the root-initial consonant (Hansson 2010:122–126; see Beckman 1998; Becker, Nevins & Levine 2012 on initial syllable faithfulness).

(11) Positional faithfulness blocks labial dissimilation in Zulu (Bennett 2015b:225, 237)

a. /seɓenz-w-a/ → [setʃ'enz-w-a] 'be worked'
b. /ɓoNg-w-a/ → [ɓoNg-w-a] 'praise (pass.), be thanked'

In sum, the GOCP predicts that existing cases of non-local-only dissimilation must result from an interaction between unbounded dissimilation and a dispreference for the result of local dissimilation. This is because, as shown in (6), the GOCP cannot generate non-local-only dissimilation on its own.

1.2 Non-local dissimilation in the SCTD

In Bennett's (2015b) SCTD, dissimilation avoids an otherwise required correspondence relation among consonants. Correspondence between surface segments is required by a set of CORR·[F] constraints, which penalize pairs of segments sharing some featural specification [F] that do not stand in correspondence with one another (see also Rose & Walker 2004; Hansson 2010). CORR·[-lateral, +liquid], for example, requires that all rhotics within a word stand in correspondence with one another.

(12) CORR·[-lateral] (CORR·[-lat]):
Assign one * for each pair of [-lateral] segments that do not stand in correspondence with one another.

Identity among surface correspondents is regulated by a family of CC·IDENT-[F] constraints, which require identity for the feature [F]. One example is CC·IDENT-[±anterior], which requires corresponding consonants to agree for [±anterior] (13).

(13) CC·IDENT-[±anterior] (CC-ID·[±ant]):
Assign one * for each pair of corresponding consonants that are [αanterior] and [-αanterior].

In a grammar where CORR·[-lateral] and CC·IDENT-[±anterior] are high-ranked, inputs like /ra-ɻata/ with [+anterior] /r/ and [-anterior] /ɻ/ must surface unfaithfully (14).[2] The faithful [r$_x$a-ɻ$_y$ata] (14a), where the rhotics do not correspond, violates CORR·[-lat]. The faithful [r$_x$a-ɻ$_x$ata] (14b), where the rhotics correspond,

[2] I switch here to inputs containing two distinct rhotics, rather than two identical rhotics or a lateral and a rhotic, as this allows for a formally simple illustration of how Bennett's proposal unites the analysis of assimilation and dissimilation. The segmental interactions analyzed here are hypothetical and not based closely on any real language. The interactions in (16–17) are an example of consonant harmony (see Rose & Walker 2004; Hansson 2001 for examples and further analysis of these cases), but the rest of the interactions represent types of languages that are, to the best of my knowledge, unattested.

violates CC·IDENT-[±anterior]. The choice between unfaithful (14c) and (14d) depends on the relative ranking of two input-output faithfulness constraints. If IDENT-[±lateral] ≫ IDENT-[±anterior], the result is place assimilation (14c); if IDENT-[±anterior] ≫ IDENT-[±lateral], the result is rhotic dissimilation (14d).

(14) High-ranked CORR·[-lat] and CC·IDENT-[±ant] can drive assimilation or dissimilation

/ra-ɻata/	CORR·[-lat]	CC·ID-[±ant]	IDENT-[±lateral]	IDENT-[±anterior]
a. [r$_x$a-ɻ$_y$ata]	*!			
b. [r$_x$a-ɻ$_x$ata]		*!		
☞ c. [ɻ$_x$a-ɻ$_x$ata]				*
☞ d. [l$_x$a-ɻ$_y$ata]			*	

Under this theory, long-distance consonant assimilation and dissimilation are two sides of the same coin: the same constraints that generate assimilation also generate dissimilation. The SCTD thus predicts a set of relationships between the typologies of long-distance assimilation and dissimilation (Bennett 2015b: Ch. 9). I focus here only on the prediction regarding the role of locality (for previous critical discussion on this point, see McMullin & Hansson 2019).

Many cases of long-distance consonant assimilation only occur across a single syllable boundary. In Ndonga, for example, suffixal /l/ maps to a nasal only if one occupies the previous syllable (/kun-il-a/ → [kun-in-a] 'sow for', but /nik-il-a/ → [nik-il-a], *[nik-in-a] 'season for'; Rose & Walker 2004:479). To formalize this locality restriction, Bennett (2015b) proposes the constraint CC·SYLLADJ (a modified version of Rose & Walker's 2004 PROXIMITY), which penalizes correspondence between segments that do not belong to adjacent syllables (15).

(15) CC·SYLLADJ (definition from Bennett 2015b:61):
'Cs in the same correspondence class must inhabit a contiguous span of syllables'
(≈ 'correspondence cannot skip across an inert intervening syllable')
For each distinct pair of output consonants X and Y, assign a violation if:

a. X and Y are in the same surface correspondence class
b. X and Y are in distinct syllables, Σx and Σy
c. there is some syllable Σz that precedes Σy, and is preceded by Σx
d. Σz contains no members of the same surface correspondence class as X and Y

Local assimilation results when CC·SYLLADJ dominates an otherwise active CORR·[F] constraint. The example in (16–17) builds on (14). When two place-distinct rhotics are in adjacent syllables, they correspond and place-assimilate (16). When the two rhotics are separated by a syllable, however, the ranking CC·SYLLADJ ≫ CORR·[-lateral] makes correspondence impossible (17). The best option given the ranking in (16–17) is (17a), where the two rhotics do not correspond and do not assimilate.

(16) CC·SYLLADJ compels local assimilation

/ɻa-rata/	CC·SYLLADJ	ID-[±lat]	CC·ID-[±ant]	CORR·[-lat]	ID-[±ant]
a. [ɻ$_x$a-r$_y$ata]				*!	
b. [ɻ$_x$a-r$_x$ata]			*!		
☞ c. [ɻ$_x$a-ɻ$_x$ata]					*
d. [ɻ$_x$a-l$_y$ata]		*!			

(17) CC·SyllAdj does not compel non-local assimilation

/ɻa-tara/	CC·SyllAdj	ID-[±lat]	CC·ID-[±ant]	Corr·[-lat]	ID-[±ant]
☞ a. [ɻ$_x$a-tar$_y$a]				*	
b. [ɻ$_x$a-taɻ$_x$a]	*!				*
c. [ɻ$_x$a-tal$_y$a]		*!			

Given a different ranking, the set of constraints employed in (16) can generate non-local-only dissimilation. Illustrative tableaux for one such ranking are in (18–19).[3]

(18) CC·SyllAdj compels non-local dissimilation

/ɻa-tara/	CC·SyllAdj	Corr-[-lat]	ID-[±lat]	ID-[±ant]	CC·ID-[±ant]
e. [ɻ$_x$a-tar$_y$a]		*!			
f. [ɻ$_x$a-tar$_x$a]	*!				*
g. [ɻ$_x$a-taɻ$_x$a]	*!			*	
☞ h. [ɻ$_x$a-tal$_y$a]			*		

(19) CC·SyllAdj does not compel local dissimilation

/ɻa-rata/	CC·SyllAdj	Corr·[-lat]	ID-[±lat]	ID-[±ant]	CC·ID-[±ant]
a. [ɻ$_x$a-r$_y$ata]		*!			
☞ b. [ɻ$_x$a-r$_x$ata]					*
c. [ɻ$_x$a-ɻ$_x$ata]				*!	
d. [ɻ$_x$a-l$_y$ata]			*!		

Dissimilation occurs in (18) because correspondence among rhotics is both required (by Corr·[-lat]) and prohibited (by CC·SyllAdj). As Ident-[±lateral] is relatively low-ranked, the optimal solution is to satisfy both Corr·[-lat] and CC·SyllAdj by mapping /r/ to [l]. In (19), dissimilation does not occur because CC·SyllAdj is not active; assimilation does not occur because ID·[±ant] dominates CC·ID-[±ant]. The system in (18–19) is thus one which exhibits non-local-only dissimilation, in the absence of any extrinsic factor that prevents local dissimilation. This is the type of system that the GOCP does not predict.

In addition to the type of system in (18–19), the SCTD – like the GOCP – predicts a range of systems in which constraints promoting unbounded dissimilation interact with those that exert (dis)preferences in local contexts. To give one example: a system differing from (18–19) in that CC·ID-[±ant] ≫ ID-[±ant] yields the mappings /ɻa-rata/ → [ɻ$_x$a-ɻ$_x$ata], /ɻa-tara/ → [ɻ$_x$a-tal$_y$a]; this is a system in which non-local dissimilation co-exists with local assimilation. To give another: it is possible to analyze the mappings in (18–19) as an interaction between unbounded rhotic dissimilation and a positional faithfulness constraint (here Ident-σ_1, after Becker et al. 2012) protecting root-initial syllables, as shown in (20–21).

(20) Positional faithfulness blocks local dissimilation

/ɻa-rata/	Ident-σ_1	Corr·[-lat]	CC·ID-[±ant]	ID-[±ant]	ID-[±lat]
☞ a. [ɻ$_x$a-r$_y$ata]		*			
☞ b. [ɻ$_x$a-r$_x$ata]			*		
c. [ɻ$_x$a-ɻ$_x$ata]	*!			*	
d. [ɻ$_x$a-l$_y$ata]	*!				*

[3] In (18) I do not consider candidates like [r$_x$a-t$_x$aɻ$_x$a], where Corr·[-lat] is satisfied by placing the intermediate syllable's onset into correspondence with the two rhotics. Such a candidate could be ruled out in a number of ways; one is to assume (*contra* Bennett 2015b) that Corr·[-lat] penalizes correspondence between a [-lateral] segment and one not specified for [±lateral] (like [t]). This move rules out the possibility of blocking-by-bridging (Bennett 2015b:65–70), which is likely desirable: Bennett's only use of the mechanism is to account for a description of Latin [l]-dissimilation whose empirical basis is questionable (Stanton 2016a).

(21) Positional faithfulness does not block unbounded dissimilation

/ɻa-tara/	IDENT-σ_1	CORR·[-lat]	CC·ID-[±ant]	ID-[±ant]	ID-[±lat]
a. [ɻ$_x$a-tar$_y$a]		*!			
b. [ɻ$_x$a-tar$_x$a]			*!		
c. [ɻ$_x$a-taɻ$_x$a]				*!	
☞ d. [ɻ$_x$a-tal$_y$a]					*

The important point here is a difference in the SCTD and GOCP's predictions regarding the character of non-local-only dissimilation. As discussed above, the GOCP predicts that non-local-only dissimilation must be linked to some interacting constraint that disprefers the consequences of local dissimilation. The SCTD, by contrast, makes no such prediction. While it is possible for a case of non-local-only dissimilation to co-exist with some external factor, this is not necessary, as non-local-only dissimilation is also predicted to exist on its own (as in (18–19)). This difference in prediction is due to a difference in the type of constraint interactions that generate non-local-only dissimilation. In the GOCP, non-local-only dissimilation occurs when local dissimilation is penalized; in the SCTD it occurs when local dissimilation is not motivated.

With respect to locality effects in dissimilation, the SCTD predicts a superset of those systems predicted by the GOCP, as non-local dissimilation can occur in both the presence and absence of external constraints that hold in local contexts. To show that the SCTD's comparative lack of restrictiveness in this domain is justified, it would be necessary to find cases of non-local dissimilation that are not obvious candidates for a GOCP-based analysis. One example of such a case could be a language with the mappings in (18–19) and (20–21) where there is no external evidence for initial-syllable faithfulness. More broadly, these would be cases of non-local dissimilation where there is no apparent reason why local dissimilation should fail.

1.3 Roadmap

The rest of the paper argues that Sundanese non-local-only dissimilation does not uniquely support the SCTD's predictions regarding locality, as a GOCP-based analysis is available. Developing a suggestion by Zuraw (2002:433), I show that the full pattern can be analyzed as resulting from the interaction of two distinct pressures: unbounded co-occurrence restrictions on [r]s and [l]s, whose effects in local contexts are obscured by a language-wide desire for identity between adjacent syllables (Section 2). Building on results presented by Cohn (1992), I show that statistical trends in the lexicon are consistent with this analysis: words containing multiple [r]s and [l]s are underattested relative to naïve expectations, and identity between adjacent syllables is overattested relative to naïve expectations (Section 3). Given the success of a GOCP-based analysis in accounting for the Sundanese pattern, the extant typology of locality in dissimilation provides us with little reason to adopt the less restrictive SCTD. Some implications for the analysis of long-distance consonant interactions more generally are discussed in the conclusions (Section 4).

2 Sundanese assimilation and dissimilation: Data and analysis

Sundanese exhibits a complex pattern of liquid assimilation and dissimilation, manifested primarily as allomorphy between [ar] and [al] (though see Section 3 for discussion of related effects in the lexicon). The allomorphs [ar] and [al] are exponents of a plural affix that appears before the first vowel in the stem.[4] It is a productive verbal affix and is also used with a small, likely closed class of nouns (Robins 1959:343). As discussed by Cohn (1992), Bennett (2015a,b) and others, the choice between [ar] and [al] depends on the presence of other liquids ([r] and [l]) within the word, as well as their location respective to the affixal liquid. The data considered throughout most of this section are in Table 1; the presentation follows Bennett (2015b:315), but with some reordering and my comments.

[4] Cohn (1992:218) notes that Ewing (1991) has shown that /ar/ is technically a distributive affix, but I follow her and Bennett (2015a,b) in continuing to refer to it as plural.

Table 1: Data illustrating the Sundanese pattern

	Input	*Output*	*Schematics*	*Comments*
a.	/ar-kusut/ ‘messy (pl.)’	[k-ar-usut]	[C-ar-VCVC]	[ar]: No other [r]s present.
b.	/ar-gɨlis/ ‘beautiful (pl.)’	[g-ar-ɨlis]	[C-ar-VlVC]	[ar]: No other [r]s present.
c.	/ar-hormat/ ‘respect (pl.)’	[h-al-ormat]	[C-al-VrCVC]	[al]: Multiple [r]s avoided.
d.	/ar-combrek/ ‘cold (pl.)’	[c-al-ombrek]	[C-al-VCCVrV]	[al]: Multiple [r]s avoided.
e.	/ar-ŋumbara/ ‘go abroad (pl.)’	[ŋ-al-umbara]	[C-al-VCCVrV]	[al]: Multiple [r]s avoided.
f.	/ar-lɨtik/ ‘little (pl.)’	[l-al-ɨtik]	[l-al-VCVC]	[al]: First consonant is /l/.
g.	/ar-liren/ ‘take a break (pl.)’	[l-al-iren]	[l-al-VrVC]	[al]: First consonant is /l/.
h.	/ar-rahɨt/ ‘wounded (pl.)’	[r-ar-ahɨt]	[r-ar-VCVC]	[ar]: Preceding onset is /r/.
i.	/ar-curiga/ ‘suspicious (pl.)’	[c-ar-uriga]	[C-ar-VrVCV]	[ar]: Following onset is /r/.

I follow Cohn (1992:207) in assuming that the affix’s underlying form is /ar/, as [ar] surfaces when the root contains no other [r] (a–b). When the root contains an [r], however, the affix generally surfaces as [al] (c–e). These alternations suggest a general process of [r]-dissimilation: co-occurrence of two [r]s is avoided by mapping the affixal /r/ to [l]. There are two kinds of exception to this pattern, both of which suggest processes of local liquid assimilation. First, if the stem-initial onset is [l], [al] surfaces unexpectedly (f–g). The result is agreement between the stem’s first two syllable onsets for [+lateral]. Second, if one of the syllables adjacent to the affixal /r/ has an /r/ onset (and the root-initial consonant is not [l]), [ar] surfaces unexpectedly (h–i). The result is agreement among onsets of adjacent syllables for [-lateral].

Bennett (2015a,b) proposes an analysis of these facts within the SCTD. The premise of the analysis is that correspondence among liquids is only possible when the liquids inhabit adjacent syllable onsets. This requirement is enforced by CC·SyllAdj (15) as well as CC·SRole, which requires corresponding consonants to have the same syllabic role. In adjacent syllable onsets, where liquids must correspond, they are forced to assimilate for [±lateral] by CC·Ident-[±lateral]. In all other contexts, liquids cannot correspond, so satisfaction of the relevant Corr constraints dictates that they must dissimilate for [±lateral]. The overall analysis is one in which the complementarity between assimilation and dissimilation observable in Table 1 is derived by constraints that limit the contexts in which liquids can correspond.

Arguments for the SCTD-based analysis of Sundanese come in part from its ability to derive this complementarity, and in part from difficulties that the data pose to co-occurrence-based theories of dissimilation (like the GOCP). Namely, it is difficult for theories invoking constraints like **X*…*Y* to explain why [r] co-occurrence is permitted only in adjacent syllables. Bennett (2015a:375) notes that “with enough wrangling, the co-occurrence constraint approach can be made to accommodate the Sundanese data”, but that “such elaborations require extra stipulations beyond the theoretical machinery of co-occurrence constraints, and they miss a significant insight about Sundanese: the connection between assimilation and dissimilation.”

Even granting these advantages, there are reasons why pursuing a co-occurrence based analysis of Sun-

danese assimilation and dissimilation is justified. First, as discussed above, the SCTD makes less restrictive predictions regarding the character of non-local-only dissimilation. Second, there is evidence that the SCTD's predictions fail to line up with the types of long-distance consonant interactions that are learnable. As Section 4 summarizes in more detail, a series of artificial grammar learning experiments by McMullin & Hansson (2016, 2019) have shown that the types of dissimilatory patterns learned by participants in artificial grammar studies correspond to the types of patterns predicted by the GOCP: local and unbounded dissimilation, plus non-local-only dissimilation with concomitant local assimilation. Crucially, participants had difficulty learning non-local-only dissimilation when not accompanied by local assimilation, the only pattern type exclusively predicted by the SCTD.

2.1 Co-occurrence-based analysis

The analysis of the Sundanese data proceeds in three parts. First (Section 2.1.1), I provide an analysis of [r]-dissimilation, as observed in Table 1's c–e. Second (in Section 2.1.2), I provide an analysis of [r]-assimilation and [l]-assimilation (Table 1's f–i) in terms of aggressive reduplication (Zuraw 2002). Third (Section 2.1.3), I fix a problem with the analysis by positing an additional process of [l]-dissimilation. Similarities and differences between the proposed analysis and related co-occurrence-based analyses (Suzuki 1999; Hansson 2001) are discussed in Section 2.3.

2.1.1 Analyzing [r]-dissimilation

The preference for the [ar] allomorph, visible from forms like [k-ar-usut] and [g-ar-ɨlis] (a–b of Table 1), follows from assuming that the morpheme's underlying representation is /ar/ and that mapping it to [al] violates IDENT-[±lateral]. To formalize the dispreference for [r] co-occurrence, visible from forms like [h-al-ormat] and [c-al-ombrek] (*[h-ar-ormat], *[c-ar-ombrek]; c–e of Table 1),*[r]…[r] (5) must dominate IDENT-[±lateral]. The fact that the affixal liquid alternates, rather than the root liquid, suggests that a root-specific version of IDENT-[±lateral], IDENT·ROOT-[±lateral], is active as well. This part of the analysis follows Hansson (2001:368–369, 2010:283–284) and is illustrated in (22).

(22) Rhotic dissimilation; /ar-hormat/ → [h-al-ormat]

/ar-hormat/	IDENT·ROOT-[±lateral]	*[r]…[r]	IDENT-[±lateral]
a. [h-ar-ormat]		*!	
☞ b. [h-al-ormat]			*
c. [h-ar-olmat]	*!		*

This ranking, however, incorrectly predicts that /ar+curiga/ should surface as *[c-al-uriga] (instead of the attested [c-ar-uriga]) and that /ar+rahɨt/ should surface as *[r-al-ahɨt] (instead of [r-ar-ahɨt]). In order to solve this problem, it is necessary to explain why violations of *[r]…[r] should be tolerated when the two [r]s belong to adjacent syllable onsets.

2.1.2 Aggressive reduplication, [r]-assimilation, and [l]-assimilation

To explain the problem posed by [c-ar-uriga] and [r-ar-ahɨt], I propose that in Sundanese there is a more general drive for adjacent syllables to be "coupled" in a reduplication-like structure (as suggested by Zuraw 2002:433). Zuraw (2002) argues that such a drive, which she terms aggressive reduplication, encourages a heightening of self-similarity between adjacent, phonologically similar constituents. For example: Zuraw interprets the frequent misspellings of English *pompon* as *pompom*, and *sherbet* as *sherbert* (among others), as the result of aggressive reduplication. In the case of *pompon*, the misspelling *pompom* results in total identity between the word's two syllables. In the case of *sherbet*, the misspelling *sherbert* results in nucleus identity. Beyond English, a desire to preserve word-internal self-similarity in Tagalog can impede an otherwise pro-

ductive word-final vowel raising process, if the result of raising would be a reduction in similarity between the final and penultimate syllables (see Zuraw 2002:410ff for more details).

Zuraw's proposal has two crucial components. The first is REDUP, which promotes word-internal coupling. While Zuraw's (p. 405) definition of REDUP is deliberately simple – "A word must contain some substrings that are coupled" – I adopt, for expositional reasons, a more specific definition that requires coupled substrings to be adjacent syllables (23).

(23) REDUP:
Assign one * if a word does not contain adjacent coupled syllables.

Whether or not a candidate has coupled substrings, and where these coupled substrings are located, is determined by GEN. Again for expositional simplicity I make two limitations to the candidates that GEN can produce. First, a coupled substring must be isomorphic with a syllable: given /pabada/, for example, $[pa]_{\kappa}[ba]_{\kappa}da$ and $pa[ba]_{\kappa}[da]_{\kappa}$ are possible candidates but $p[a]_{\kappa}[ba]_{\kappa}da$ and $[pab]_{\kappa}[ada]_{\kappa}$ are not. Second, a word may have no more than two coupled substrings. This, together with the adjacency requirement imposed by (23), limits the possible coupling structures. For /pabada/, the only licit candidates that would satisfy REDUP are $pa[ba]_{\kappa}[da]_{\kappa}$ and $[pa]_{\kappa}[ba]_{\kappa}da$. The candidate $[pa]_{\kappa}ba[da]_{\kappa}$ does not satisfy REDUP because its coupled syllables are not adjacent; $[pab]_{\kappa}[ada]_{\kappa}$ and $[pa]_{\kappa}[ba]_{\kappa}[da]_{\kappa}$ are not admissible by GEN because of the size and number of coupled substrings.[5]

The second component of Zuraw's proposal is a set of faithfulness constraints that hold between corresponding segments in coupled substrings. These faithfulness constraints are assumed to belong to the same families of constraints as those that govern identity along other correspondence dimensions, e.g. IDENT-[F] and MAX. The candidate $[pa]_{\kappa}[ba]_{\kappa}dra$, for example, with the correspondence structure $[p_1a_2]_{\kappa}[b_1a_2]_{\kappa}dra$, would incur a violation of $\kappa\kappa$·IDENT-[±voice] because the corresponding [p] and [b] differ in [±voice]. The candidate $pa[ba]_{\kappa}[dra]_{\kappa}$, with the correspondence structure $pa[b_1a_3]_{\kappa}[d_1r_2a_3]_{\kappa}$, would incur a violation of $\kappa\kappa$·MAX because [r] is present in one substring but absent from the other. The $\kappa\kappa$ faithfulness constraints can promote identity among coupled substrings, if $\kappa\kappa$·IDENT $\gg$ IO·IDENT; they can also inhibit coupling among dissimilar substrings, if $\kappa\kappa$·IDENT $\gg$ REDUP (similarity is a prerequisite to coupling) and IO·IDENT $\gg$ $\kappa\kappa$·IDENT (input specifications cannot be changed to enhance self-similarity). Crucially, these constraints do not assign violations in the absence of coupling; thus [pabada] does not violate $\kappa\kappa$·IDENT-[±voice] and [pabrada] does not violate $\kappa\kappa$·MAX because there are no coupled substrings. (Violations of $\kappa\kappa$·DEP and $\kappa\kappa$·MAX are equivalent; I use MAX, following Zuraw 2002.)

In this section, what will be of interest is whether or not the onsets of coupled syllables are identical. While it would be possible to encode these requirements as the combination of $\kappa\kappa$·MAX (ensuring that onsets contain the same number of segments) and a set of $\kappa\kappa$·IDENT-[F] constraints (ensuring that onsets contain the same segments), I depart from Zuraw in assuming that faithfulness constraints along the $\kappa\kappa$ dimension can also evaluate entire syllabic constituents. Thus the requirement for onset identity among coupled syllables is formalized as $\kappa\kappa$·IDENT-[onset] ((24); see Suzuki 1999 for a similar proposal).

(24) $\kappa\kappa$·IDENT-[onset]:
Assign one * if the onsets of coupled syllables are not identical.

With this in place, we can continue with the analysis of [c-ar-uriga] and [r-ar-ahɨt]. To derive the fact that co-occurring [r]s are permitted in adjacent syllable onsets, I propose that Sundanese prioritizes coupling over

[5] As suggested by Zuraw (in a more general way, p. 405), these limitations could be derived through the interaction of a more general REDUP with constraints that govern reduplicant size and placement. Sundanese has initial-syllable partial reduplication; reduplicants are always adjacent to their bases, and there are no instances of multiple reduplication that I am aware of (see Robins 1959 on partial reduplication in Sundanese, and Hansson 2010:289 for previous discussion of the connection between partial reduplication and /ar/ allophony). As the nature of REDUP is not the focus of this paper, I do not explore this alternative.

avoiding words with multiple [r]s (REDUP ≫ *[r]…[r]), but that adjacent syllables must have identical onsets to be coupled ($\kappa\kappa$·IDENT-[onset] ≫ REDUP). A tableau for [c-ar-uriga] illustrates the analysis. (In all tableaux that follow, I omit IDENT·ROOT-[±lateral] and candidates that violate it. I also do not consider candidates in which the affixal /r/ maps to anything other than [r] or [l]; this amounts to a claim that faithfulness constraints for all consonant features, except IDENT-[±lateral], dominate $\kappa\kappa$·IDENT-[onset].)

(25) Aggressive reduplication results in unexpected realization of [ar] allomorph

/ar-curiga/	$\kappa\kappa$·IDENT-[onset]	REDUP	*[r]…[r]	IDENT-[±lateral]
a. c-ar-uriga		*!	*	
☞ b. c-a[r-u]$_\kappa$[ri]$_\kappa$ga			*	
c. c-al-uriga		*!		*
d. c-a[l-u]$_\kappa$[ri]$_\kappa$ga	*!			*

Candidates (25a,c) do not contain adjacent coupled syllables and are eliminated by REDUP. Candidate (25d) satisfies REDUP but is eliminated by higher-ranked $\kappa\kappa$·IDENT-[onset] (see Section 2.2 for this ranking argument), as the onsets of the coupled syllables are not identical. The optimal (25b) shows that violation of *[r]…[r] is acceptable when it allows for satisfaction of higher-ranked REDUP and $\kappa\kappa$·IDENT-[onset]. Put differently, violation of *[r]…[r] is permitted when the result is onset identity between adjacent syllables. Note however that in these data, the drive for self-similarity is limited to onsets: forms like [r-ar-ɨwat] 'startled (pl.)' and [di-k-ar-irim] 'sent-PASS (pl.)' (Cohn 1992:206), in addition to [c-ar-uriga], suggest that further constraints requiring identity among coupled syllables are inactive. In c-a[r-u]$_\kappa$[ri]$_\kappa$ga and [r-a]$_\kappa$[r-ɨ]$_\kappa$wat, the nuclei of coupled syllables are not identical, suggesting that input-output faithfulness constraints on vowel quality dominate $\kappa\kappa$·IDENT-[nucleus]. (But for evidence of gradient nucleus-matching in the lexicon, see Section 3.) The non-identical rime structure of di-k-a[r-i]$_\kappa$[rim]$_\kappa$ suggests that MAX and DEP dominate $\kappa\kappa$·IDENT-[coda].

In this way, the current analysis derives the generalization that adjacent [r]-containing onsets are not allowed if they are not identical. In the case of /ar-combrek/, c-a[r-om]$_\kappa$[brek]$_\kappa$ (26b) has adjacent [r]-containing onsets but still violates $\kappa\kappa$·IDENT-[onset] because the onsets are not identical. The analysis correctly predicts that the winning candidate should be c-al-ombrek (26c), as unlike c-ar-ombrek it satisfies *[r]…[r]. (In the tableau below I do not consider the candidates c-a[r-om]$_\kappa$[rek]$_\kappa$ and c-a[br-om]$_\kappa$[brek]$_\kappa$, with deletion and insertion. I assume that these are ruled out by undominated MAX and DEP.[6])

(26) Aggressive reduplication not possible for /ar-combrek/ due to mismatched onsets

/ar-combrek/	$\kappa\kappa$·IDENT-[onset]	REDUP	*[r]…[r]	IDENT-[±lateral]
a. c-ar-ombrek		*	*!	
b. c-a[r-om]$_\kappa$[brek]$_\kappa$	*!		*	
☞ c. c-al-ombrek		*		*
d. c-a[l-om]$_\kappa$[brek]$_\kappa$	*!			*
e. [c-a]$_\kappa$[l-om]$_\kappa$brek	*!			*

Similarly, the analysis derives the generalization that identical [r]-containing onsets are only licit if they are adjacent. In the case of /ar-ŋumbara/, for example, ŋ-a[r-um]$_\kappa$ba[ra]$_\kappa$ (27b) satisfies $\kappa\kappa$·IDENT-[onset] but violates REDUP; ŋ-a[l-um]$_\kappa$ba[ra]$_\kappa$ (27d) violates both. ŋ-al-umbara (27c) is selected as optimal because, unlike ŋ-ar-umbara (27a), it satisfies *[r]…[r]. (I do not include candidates like ŋ-a[l-um]$_\kappa$[ba]$_\kappa$ra; these satisfy REDUP, but violate high-ranked $\kappa\kappa$·IDENT-[onset].)

[6] Assuming undominated MAX and DEP means that we expect to find dissimilation for hypothetical inputs like /ar-krenda/. It is not clear that any cluster-initial roots form the plural with [ar] (Bennett 2015b:143), so this prediction cannot be tested.

(27) Aggressive reduplication not possible for /ar-ŋumbara/ due to non-adjacent onsets

/ar-ŋumbara/	ᴋᴋ·IDENT-[onset]	REDUP	*[r]…[r]	IDENT-[±lateral]
a. ŋ-ar-umbara		*	*!	
b. ŋ-a[r-um]$_{\kappa}$ba[ra]$_{\kappa}$		*	*!	
☞ c. ŋ-al-umbara		*		*
d. ŋ-a[l-um]$_{\kappa}$ba[ra]$_{\kappa}$	*!	*		*

Finally, for forms like /ar-hormat/ (22), the analysis correctly predicts that h-al-ormat is the winner. This is because [l] and [r] cannot correspond: the liquids are contained in the same syllable (ha.lor.mat), and I assume that coupled substrings must be isomorphic to a syllable (*ha[lo]$_{\kappa}$[r]$_{\kappa}$mat).

2.1.3 Analyzing [l]-dissimilation

The current analysis incorrectly predicts that /ar-gɨlis/ should surface as g-a[l-i]$_{\kappa}$[lis]$_{\kappa}$. Because ᴋᴋ·IDENT-[onset] and REDUP dominate IDENT-[±lateral], mapping /ar/ to [al] should occur when the result would be satisfaction of the constraints promoting aggressive reduplication. To solve this problem, I propose that a second co-occurrence constraint, *[+lateral]…[+lateral] (or *[l]…[l], (28)), is active in Sundanese.[7]

(28) *[+lateral]…[+lateral] (*[l]…[l]):
Assign one * for each pair of [l]s within the word.

To take effect, *[l]…[l] must dominate REDUP; the intuition is that avoiding [l] co-occurrence is more important than having coupled syllables. Under this ranking, g-al-ɨlis (29d) and g-a[l-ɨ]$_{\kappa}$[lis]$_{\kappa}$ (29e) are eliminated by high-ranked *[l]…[l]; g-ar-ɨlis (29a) is selected as optimal because it satisfies both top-ranked constraints (unlike (29b,c)), despite its violation of REDUP.

(29) *[l]…[l] ≫ REDUP predicts /ar-gɨlis/ → [g-ar-ɨlis]

/ar-gɨlis/	ᴋᴋ·IDENT-[onset]	*[l]…[l]	REDUP	IDENT-[±lateral]
☞ a. g-ar-ɨlis			*	
b. g-a[r-ɨ]$_{\kappa}$[lis]$_{\kappa}$	*!			
c. [g-a]$_{\kappa}$[r-i]$_{\kappa}$lis	*!			
d. g-al-ɨlis		*!	*	*
e. g-a[l-ɨ]$_{\kappa}$[lis]$_{\kappa}$		*!		*

The next part of the pattern to explain is why /ar-lɨtik/ surfaces as [l-al-ɨtik] and /ar-liren/ surface as [l-al-iren]. As shown in (30) for /ar-liren/, the current analysis incorrectly predicts that the infix should surface as [ar] (30b), because *[l]…[l] dominates *[r]…[r].

(30) *[l]…[l] ≫ *[r]…[r] predicts the wrong output for /ar-liren/

/ar-liren/	ᴋᴋ·IDENT-[onset]	*[l]…[l]	REDUP	*[r]…[r]
a. l-ar-iren			*!	*
💣 b. l-a[r-i]$_{\kappa}$[ren]$_{\kappa}$				*
☹ c. l-al-iren		*!	*	
☹ d. [l-a]$_{\kappa}$[li]$_{\kappa}$ren		*!		

To account for these data, I propose that coupling is preferred between the first two syllables of the word, and

[7] A reviewer asks why *[r]…[r] and *[l]…[l] hold, and not other similar constraints. While it would be attractive to claim that one constraint, *[αlateral]…[αlateral], is responsible for both patterns, this is not possible as *[r]…[r] and *[l]…[l] need to be ranked at different places in the hierarchy. One possible reason why we find both [l]- and [r]-dissimilation in this language is simply that they are both frequent kinds of dissimilation; both are described as 'robustly attested' by Bennett 2015b:331.

that this context-sensitive preference for coupling overrides the prohibition on [l] co-occurrence. I formalize the preference for initial coupling as a context-sensitive version of REDUP, REDUP-$\sigma_1\sigma_2$ (31), which requires that the stem's first two syllables be coupled.

(31) REDUP-$\sigma_1\sigma_2$:
Assign one * if the first two syllables of the stem are not coupled.

To derive the result that /ar-liren/ surfaces as [l-a]$_\kappa$[l-i]$_\kappa$ren, REDUP-$\sigma_1\sigma_2$ must dominate *[l]…[l]. Tableau (32) confirms that, with this ranking in place, [l-a]$_\kappa$[l-i]$_\kappa$ren (32d) is correctly selected as optimal. The overall effect is that [l] co-occurrence is permitted only if it results in onset identity between the first two syllables.

(32) REDUP-$\sigma_1\sigma_2$ ≫ *[l]…[l] explains /ar-liren/ → [l-al-iren]

/ar-liren/	$\kappa\kappa$·IDENT-[onset]	REDUP-$\sigma_1\sigma_2$	*[l]…[l]	REDUP
a. l-ar-iren		*!		*
b. l-a[r-i]$_\kappa$[ren]$_\kappa$		*!		
c. l-al-iren		*!	*	*
☞ d. [l-a]$_\kappa$[l-i]$_\kappa$ren			*	

The ranking $\kappa\kappa$·IDENT-[onset] ≫ REDUP-$\sigma_1\sigma_2$ is motivated by further consideration of forms like [c-ar-uriga], where this analysis assumes that [r] co-occurrence is permitted because the second and third syllables are coupled. Tableau (33) demonstrates that if the ranking between these two constraints were reversed, as REDUP-$\sigma_1\sigma_2$ ≫ $\kappa\kappa$·IDENT-[onset], the analysis would incorrectly select [c-a]$_\kappa$[l-u]$_\kappa$riga (33c).

(33) $\kappa\kappa$·IDENT-[onset] ≫ REDUP-$\sigma_1\sigma_2$ is necessary for /ar-curiga/ → [c-ar-uriga]

/ar-curiga/	REDUP-$\sigma_1\sigma_2$	$\kappa\kappa$·IDENT-[onset]	REDUP	*[r]…[r]
a. c-ar-uriga	*!		*	*
b. [c-a]$_\kappa$[r-u]$_\kappa$riga		*		*!
💣 c. [c-a]$_\kappa$[l-u]$_\kappa$riga		*		
☹ d. c-a[r-u]$_\kappa$[ri]$_\kappa$ga	*!			*

To allow coupling to occur outside of the stem-initial context, then, $\kappa\kappa$·IDENT-[onset] must dominate REDUP-$\sigma_1\sigma_2$. With this final ranking in place, the analysis can now account for all data in Table 1.

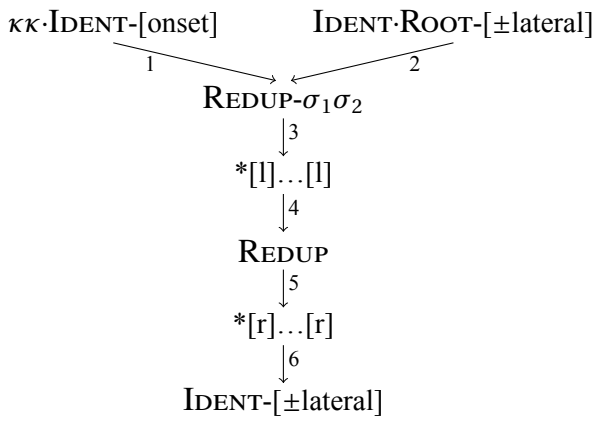
2.2 Summary of analysis

The proposed analysis of the Sundanese data is summarized in Figure 1, with winner-loser pairs provided to illustrate each ranking argument. IDENT·ROOT-[±lateral] ≫ REDUP-$\sigma_1\sigma_2$ was not established in the analysis above but is included and justified below for completeness's sake.

As mentioned throughout, a number of other constraints regulate aggressive reduplication in these infixed forms. Some of these constraints dominate $\kappa\kappa$·IDENT-[onset]: constraints demanding faithfulness to features other than [±lateral], for example, are necessary to explain why /ar-kusut/ does not surface as [k-a]$_\kappa$[k-u]$_\kappa$sut. Some of these constraints are lower-ranked: both $\kappa\kappa$·IDENT-[nucleus] and $\kappa\kappa$·IDENT-[coda] must be dominated by REDUP, as coupling is allowed word-medially despite a lack of nucleus and coda identity. Finally, in some cases the position of these constraints in the hierarchy is unclear: MAX and DEP must dominate REDUP to explain why /ar-combrek/ is not c-a[r-om]$_\kappa$[rek]$_\kappa$ or c-a[br-om]$_\kappa$[brek]$_\kappa$, but a lack of relevant forms that contain clusters make their rankings with respect to REDUP-$\sigma_1\sigma_2$ impossible to establish.

One form provided by Bennett (2015b:142), [al-ulur] 'lower on a rope (pl.)', poses a problem for this analysis.[8] The ranking in Figure 1 predicts that it should surface instead as [ar-ulur], as *[l]…[l] violations

[8] Bennett (2015b) transcribes this as [(ʔ)-al-ulur], but does not hear the [ʔ], and notes that its inclusion is for consistency with



[1] κκ·IDENT-[onset] ≫ REDUP-$\sigma_1\sigma_2$: c-a[r-u]$_\kappa$[ri]$_\kappa$ga ≻ *[c-a]$_\kappa$[l-u]$_\kappa$riga
[2] IDENT·ROOT-[±lateral] ≫ REDUP-$\sigma_1\sigma_2$: liren ≻ *[ri]$_\kappa$[ren]$_\kappa$
[3] REDUP-$\sigma_1\sigma_2$ ≫ *[l]…[l]: [l-a]$_\kappa$[l-ɨ]$_\kappa$tik ≻ *l-ar-ɨtik
[4] *[l]…[l] ≫ REDUP: g-ar-ɨlis ≻ *g-a[l-ɨ]$_\kappa$[lis]$_\kappa$
[5] REDUP ≫ *[r]…[r]: c-a[r-u]$_\kappa$[ri]$_\kappa$ga ≻ *c-al-uriga
[6] *[r]…[r] ≫ IDENT-[±lateral]: h-al-ormat ≻ *h-ar-ormat

Figure 1: Summary of analysis

are only tolerated when the [l]s occupy the first two syllables' onsets. Since this is not a possibility for /ar-ulur/, whose root has no initial consonant, the ranking *[l]…[l] ≫ *[r]…[r] prefers unattested *[ar-ulur] (34a) over attested [al-ulur] (34c,d).

(34) Current analysis predicts wrong output for /ar-ulur/

/ar-ulur/	κκ·IDENT-[onset]	*[l]…[l]	REDUP	*[r]…[r]
💣 a. ar-ulur			*	*
b. a[r-u]$_\kappa$[lur]$_\kappa$	*!			*
☹ c. al-ulur		*!	*	
☹ d. a[l-u]$_\kappa$[lur]$_\kappa$		*!		

There are at least two ways to make sense of this apparent exception. One is to treat it as just that – an exception – and to claim that /ulur/ must exceptionally be realized with the plural allomorph [al]. Such a provision must be part of the analysis in any case, as lexical exceptions exist: Robins (1959:344) notes that [gəde] 'to be big' forms its plural as [g-al-əde], and Cohn's (1992:219) discussion strongly implies that there are others. It is also possible, however, to capture the appearance of [al] in [al-ulur] by revising the definition of REDUP-$\sigma_1\sigma_2$. [al-ulur] is unlike all other forms considered here in that it is vowel-initial, and the affix /ar/ surfaces as a prefix. If REDUP-$\sigma_1\sigma_2$ were revised to demand that the first two syllables containing root material must be coupled, candidates (34b,d) would satisfy REDUP-$\sigma_1\sigma_2$ and a[l-u]$_\kappa$[lur]$_\kappa$ (34d) would be correctly chosen as the winner.[9] It is difficult to know at present which of these solutions is more plausible.

past descriptions. The distribution of [ʔ] is predictable (Robins 1959) and from this I infer that is not part of a root's underlying representation; whether or not and where it surfaces predictably is not important here.

[9] A variant of this would be to claim that REDUP-$\sigma_1\sigma_2$ requires coupling between the stem's first and second syllables, as claimed in (31), but that onsetless syllables cannot function as stem-initial syllables (for typological evidence supporting this idea as well as a formal implementation, see Downing 1998). I have not pursued this idea here as I have not found corroborating evidence from other phonological or morphological processes in Sundanese.

An anonymous reviewer asks what this analysis predicts regarding the realization of /arar/, an augmentative affix (examples in Anderson 1997:16). While there is little information regarding the phonology of this affix, the current analysis correctly predicts that /arar-amprok/ should be realized as [arar-amprok] ('everybody meeting together', Bennett 2015a:391): this is because a[ra]$_\kappa$[r-am]$_\kappa$prok (with two violations of *[r]…[r]) is preferable to *a[la]$_\kappa$[l-am]$_\kappa$prok (with a violation of *[l]…[l]) and other admissible candidates. The analysis predicts that /arar/ should surface as [alar] given a lateral-initial root (hypothetical /alar-liren/ → [l-alar-iren]), but no data that I am aware of bears on this prediction.

2.3 Comparison with alternatives

The Sundanese data discussed here have been analyzed several times before, and the analysis proposed in this section bears some resemblance to prior analyses by Suzuki (1999) and Hansson (2001). Some major similarities and differences among these analyses are highlighted below.

The proposed analysis shares with Suzuki (1999) and Hansson (2001) an appeal to *[r]…[r], to account for the realization of /ar/ as [al] in forms like [h-al-ormat]. It also shares with these analyses an appeal to some form of surface correspondence that holds between adjacent syllables, as well as a constraint requiring identity among correspondents. The specifics of how this is accomplished differ. Suzuki's (1999) proposal is closer to the present one: he assumes that any two adjacent syllables can stand in correspondence, and appeals to IDENT$\sigma_1\sigma_2$[ONS] ("Adjacent syllables have an identical onset specifications [sic]") to derive [r] co-occurrence in forms like [c-ar-uriga]. Hansson (2001) appeals to two constraints which together derive [r] co-occurrence: CORR-[lat]$_{\text{Ons}(\sigma 1\text{-}\sigma 2)}$ ("liquids in adjacent syllable onsets must correspond") and IDENT[lat]-CC ("corresponding consonants must agree for [±lateral]").

The proposed analysis differs from prior work in how [l] co-occurrence is analyzed. Suzuki (1999) and Hansson (2001) account for this aspect of the pattern by proposing that two distinct types of correspondence are active in Sundanese: base-reduplicant correspondence (which holds between the first two syllables) and another form of surface correspondence (which can hold between all pairs of adjacent syllables). The limitation of [l] co-occurrence to initial position is then derived by assuming that different faithfulness constraints hold within these two correspondence dimensions. Under Hansson's (2001) proposal, the existence of [l] and [r] co-occurrence in initial position shows us that IDENT[±lat]-BR is high-ranked; the existence of only [r] co-occurrence elsewhere shows us that IDENT[-lat]-CC is high-ranked but IDENT[+lat]-CC is not.

The proposed analysis, by contrast, appeals to only one dimension of correspondence. It accounts for the positional limitation of [l] co-occurrence by positing *[l]…[l] and ranking REDUP-$\sigma_1\sigma_2$ above it. To the best of my knowledge, this part of the analysis does not have a clear precedent. (Bennett 2015a,b is probably the closest in positing a position-specific markedness constraint, CC·IDENT-Initial-[lateral], which requires agreement among corresponding consonants for [±lateral] if one of the consonants is stem-initial.) Thus the proposed analysis is novel in two ways. First, it attributes the special behavior of $\sigma_1\sigma_2$ to a distinct pressure for correspondence in this context. Second, it attributes the absence of [l] co-occurrence outside of $\sigma_1\sigma_2$ to a co-occurrence constraint, *[l]…[l]. The overall characterization of the pattern is one in which marked configurations (here co-occurring [r]s and [l]s) are licensed in order to enhance the self-similarity of adjacent syllables. This is distinct from the characterizations argued for in prior work.

3 Evidence from the lexicon

As discussed above, co-occurrence-based theories of dissimilation predict that non-local-only dissimilation must coexist with an interacting pressure that disprefers the result of local dissimilation. Under the analysis above, Sundanese instantiates this prediction: dissimilation of [r]s and [l]s is obscured in local contexts by a general desire for identity between adjacent syllables.

Previous work suggests independent evidence consistent with aspects of this analysis. Regarding rhotic dissimilation, Cohn (1992:213) notes that loanwords with multiple [r]s often undergo optional dissimilation

(*rapor*, *lapor*, or *rapot* for 'report'; *direktur* or *dalektur* for 'director'). In addition, she attributes to Eringa (1949) the observation that other morphologically complex forms optionally exhibit rhotic dissimilation as well (e.g. *pira(ŋ)+kadar* 'type+fate' optionally maps to *pilakadar* 'only'). These facts are consistent with a system in which *[r]…[r] is active. Regarding aggressive reduplication: Cohn's (1992:213-214) investigation of Lembaga Basa & Sastra Sunda (1985), a large Sundanese dictionary, reveals that 105 of the dictionary's [r]-initial entries have co-occurring [r]s. In 87 of these, the [r]s are onsets of adjacent syllables that also have identical nuclei (e.g. *rara* 'braid', *rorod* 'pull in (as a string of a kite)', *ragrag* 'fall'). Zuraw (2002:433) notes that the observed correlation between [r] co-occurrence and nucleus-matching is consistent with an interaction between dissimilation and aggressive reduplication, as "successive liquid onsets that escape a general dissimilation process are likely to belong to strings that are similar in other ways".

This section replicates and expands on Cohn's findings by providing evidence that trends in the Sundanese lexicon are consistent with the activity of liquid dissimilation and aggressive reduplication. This evidence and its relationship to the analysis is previewed below.

- *Evidence for aggressive reduplication in adjacent syllables:*
  If onset identity is a prerequisite for coupling in Sundanese, we might expect that syllables with identical onsets are more likely than expected to be similar in other ways. This is because if an adjacent pair of syllables satisfies $\kappa\kappa$·IDENT-[onset], coupling is compelled by REDUP. These coupled syllables are then evaluated by further $\kappa\kappa$·IDENT constraints, like $\kappa\kappa$·IDENT-[nucleus] and $\kappa\kappa$·IDENT-[coda]. (While input-output faithfulness constraints prevented us from seeing the effects of these further constraints in Section 2, we might expect to find effects in the larger lexicon, as faithfulness constraints do not play a role in lexical innovation; see discussion in Section 3.2.4.) I show that this prediction is borne out in the relationship between onset and nucleus identity. When syllables are adjacent, there is a statistically significant correlation between onset-matching and nucleus-matching: syllables with matching onsets are disproportionately likely to have matching nuclei. For non-adjacent syllables, no such correlation exists. These findings are consistent with Section 2's claim that REDUP requires coupling only between adjacent syllables, and that a family of $\kappa\kappa$·IDENT constraints promotes identity among coupled syllables.

- *Evidence for aggressive reduplication in $\sigma_1\sigma_2$:*
  Evidence consistent with the claim that aggressive reduplication is specifically preferred between the first two syllables (formalized in Section 2 as REDUP-$\sigma_1\sigma_2$) comes from patterns of onset-matching. Namely, the onsets of σ_1 and σ_2 are more likely to be identical than is predicted by the frequency of individual onsets in these positions. Importantly, this preference for onset-matching does not hold in $\sigma_2\sigma_3$ or $\sigma_1\sigma_3$ (when other processes promoting identity are controlled for; see Section 3.2.3).

- *Evidence for restrictions on multiple [r]s and multiple [l]s:*
  If there are active co-occurrence restrictions on multiple [r]s and [l]s (formalized in Section 2 as *[r]…[r] and *[l]…[l]) we should find words containing multiple [r]s or [l]s to be significantly less frequent than expected. I show that this is true throughout the Sundanese lexicon, even in contexts where identity is otherwise preferred (like the onsets of $\sigma_1\sigma_2$; see discussion in Section 3.2.1).

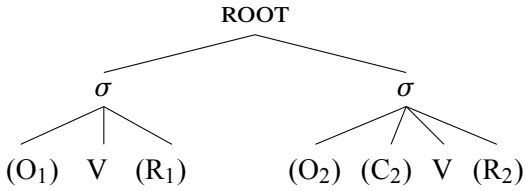
The main point of this section is that trends in the Sundanese lexicon are consistent with each of the markedness constraints proposed in Section 2. These findings are thus consistent with the claim that non-local dissimilation in Sundanese can be analyzed as the interaction between unbounded dissimilation and a preference for identity between adjacent syllables.

Section 3.1 discusses methodological aspects of this study, including information about the data source and the statistical models. Context-by-context results are presented in Section 3.2. Section 3.3 provides a potential learnability-based reason why we should take seriously these links between /ar/ allomorphy and the lexicon. A further corpus study suggests that /ar/-affixed forms supporting the crucial rankings in Figure

1 are likely rare, yet Sundanese children have no problem acquiring the correct grammar: the pattern has been stable for decades. The trends established in Section 3.2 raise the possibility that the relevant constraints and their ranking are discoverable from the lexicon, and that successful acquisition of /ar/ allomorphy may not require much exposure to /ar/-affixed forms.

3.1 Methods

The lexicon study discussed in this section is based on a wordlist that contains 11,913 headwords from Lembaga Basa & Sastra Sunda (1985), excluding only those that were explicitly marked as borrowings (from Arabic, English, Javanese, Malay, and a number of other languages). Each word in this list was syllabified according to Cohn's (1992:205) description of cluster phonotactics in the Sundanese root pattern. For clarity, her description of the canonical Sundanese root is replicated in Figure 2.



O = onset, R = rhyme
O_1, O_2 = any consonant
R_1 = nasal homorganic to the following stop, /r/ and (rarely) others
R_2 = most consonants, except palatal [-continuant] consonants
C_2 = /r/, /l/ after a stop (rare)

Figure 2: Cohn's (1992:205) description of the canonical Sundanese root pattern

Following this description meant that a word like *ablag* 'large, spacious' was syllabified as *a.blag* and a word like *ambacak* 'scattered' was syllabified as *am.ba.cak*. A small number of words contained triconsonantal or longer clusters not explicitly described by Cohn; in these cases, the first consonant was assigned to a syllable coda and the rest to the following onset (*tasbla*ŋ 'finished study / nothing more to learn; from dusk to dawn (awake all night)', for example, was syllabified as *tas.bla*ŋ). Unsyllabified and syllabified versions of the wordlist are available as supplementary materials.

The lexicon analysis takes into account only disyllabic and trisyllabic words. This limitation was made because most quadrisyllabic or longer forms in Lembaga Basa & Sastra Sunda (1985) appear to be morphologically complex or are likely unmarked borrowings (e.g. *afghanistan* 'Afghanistan').[10] In particular, a large number of the longer forms appear to be fully reduplicated roots (e.g. *alangahélengeh* 'shy smile', *alunalun* 'square, plaza'; see Van Syoc 1959:78-80 on morphological reduplication in Sundanese). Part of our interest here is in the extent of evidence for the activity of aggressive reduplication, so including morphologically reduplicated forms would bias the results.

Because each word was maximally three syllables, there were a total of three syllabic contexts to investigate: the first and second syllables ($\sigma_1\sigma_2$), the second and third ($\sigma_2\sigma_3$), and the first and third ($\sigma_1\sigma_3$). For each context, the forms considered were only those that had a native (i.e. not /f v z ʔ/) singleton onset in both positions.[11] Thus words like *ke.ke.ba* 'a bag/container made out of bamboo', where all syllables have

[10] Regarding the issue of loanwords, an anonymous reviewer asks about the influence of Javanese loans on Sundanese lexical statistics. Words marked as Javanese loans in the dictionary (e.g. *kecut* 'sour') have been excluded, as per the discussion above. Abby Cohn (p.c.) notes that Javanese loans are common in the high register of Sundanese, but that many speakers do not command the high register, and that it would be surprising if speakers were aware which words are of Javanese origin and which aren't.

[11] The limitation to native singleton onsets was made largely to simplify the statistics and the data visualizations, but also in part

singleton onsets, are considered for all contexts ($\sigma_1\sigma_2$, $\sigma_2\sigma_3$, and $\sigma_1\sigma_3$). Words like *am.ba.cak*, where one syllable has no onset, are only considered for a subset of the contexts (here only $\sigma_2\sigma_3$). Words like *ke.de.plik* 'very thick', where one syllable has a complex onset, are also only considered for a subset of the contexts (here only $\sigma_1\sigma_2$). Finally, words like *ka.ri* 'leftovers' are only considered for $\sigma_1\sigma_2$, as they lack a third syllable. The number of forms considered per context, with examples, is in (35).

(35) Number of forms considered per context

Context	Number	Examples
$\sigma_1\sigma_2$	9,604	*ke.ke.ba*, *ka.ri*
$\sigma_2\sigma_3$	3,030	*ke.ke.ba*, *am.ba.cak*
$\sigma_1\sigma_3$	2,933	*ke.ke.ba*, *pa.i.don* 'a tool used to spit'

To determine the frequency of onset pairs relative to expectation, loglinear models were fit to each of the datasets in (35). Loglinear models were chosen as they are a statistically sound way of analyzing count data (see Wilson & Obdeyn 2009 for discussion). For each model, the dependent variable was the number of times a particular onset-onset pair was attested. The independent variables included a predictor for identity (is the onset-onset pair composed of two identical consonants?) and one predictor per onset segment per position. For example, if the possible syllable onsets for a given language are /p t l k/, this results in eight segmental predictors: four for the segments in first position (p_1, t_1, l_1, k_1) and four for the segments in second position (p_2, t_2, l_2, k_2). Each predictor assigned a 1 if that segment was present in the specified position and a 0 if it wasn't. In addition, one predictor was included for each identical onset pair of interest (e.g. l_{12}). The schematic example in Table 2 illustrates the structure of the model inputs for a made-up language whose possible onsets are /p t l k/ and where the rate of [l] co-occurrence is of interest.

Two models were fit to each subset of the data. In the baseline model, the counts were modeled as a function of only the segmental predictors (p_1, p_2, etc.). This model was then queried for a set of fitted values (with R's *fitted.values* function) that reflect how frequent each pair is predicted to be, given no constraints on onset-onset combination. If the pair is more frequent than predicted, it is overattested relative to naïve expectation; if it is less frequent than predicted, it is underattested. Following this, predictors that reference identity (above as Identity, l_{12}) were added to the model. The Identity predictor was included to let the model assess whether or not pairs of identical onsets, as a class, are overattested or underattested. The predictors for identical onset combinations (like l_{12}) were included to let the model determine if individual pairs of identical onsets are overattested or underattested, relative to the expectations set by the frequency of identical pairs (as a class) and the independent frequency of each member of the pair. In this way, these models allow us to evaluate evidence for a potential identity preference (which would manifest as significant overattestation of identical pairs) as well as evidence for co-occurrence restrictions on [r]s and [l]s (which would manifest as underattestation of those specific pairs). All loglinear models were fit with the *bayesglm* function of R's *arm* package (Gelman & Hill 2006) and the *quasipoisson* link function.[12]

For each context, further evidence for aggressive reduplication was evaluated by determining if onset-matching was significantly correlated with nucleus-matching. This was done by splitting the forms into four groups, according to (i) whether or not their onsets match and (ii) whether or not their nuclei match, and performing chi-squared tests on the resulting contingency tables.

because non-native and cluster onsets are infrequent. Widening the corpus to contain these forms does not qualitatively change the results or any or the conclusions drawn from them. (Note that while [ʔ] is a native Sundanese phone, its distribution is predictable and it is not written. Instances of it in the dictionary are likely not due to this predictable pattern; Abby Cohn (p.c.) notes that *kaʔbah* 'a place in Mecca', for example, is likely an Arabic loan. See Robins, 1959, for further discussion of [ʔ].)

[12] The *bayesglm* function was selected as Bayesian regression was found to be uniquely capable of accommodating the numerous 0s in the Sundanese count data. The *quasipoisson* link function is appropriate for these data because in all relevant subsets, the variance in frequency is larger than the mean.

Table 2: Co-occurrence count encoding for regression analysis

Combination	Count	Identity	p_1	t_1	l_1	k_1	p_2	t_2	l_2	k_2	l_{12}
p+p	2	1	1	0	0	0	1	0	0	0	0
p+t	7	0	1	0	0	0	0	1	0	0	0
p+l	5	0	1	0	0	0	0	0	1	0	0
p+k	4	0	1	0	0	0	0	0	0	1	0
t+p	30	0	0	1	0	0	1	0	0	0	0
t+t	15	1	0	1	0	0	0	1	0	0	0
t+l	44	0	0	1	0	0	0	0	1	0	0
t+k	26	0	0	1	0	0	0	0	0	1	0
l+p	15	0	0	0	1	0	1	0	0	0	0
l+t	13	0	0	0	1	0	0	1	0	0	0
l+l	6	1	0	0	1	0	0	0	1	0	1
l+k	14	0	0	0	1	0	0	0	0	1	0
k+p	3	0	0	0	0	1	1	0	0	0	0
k+t	2	0	0	0	0	1	0	1	0	0	0
k+l	3	0	0	0	0	1	0	0	1	0	0
k+k	0	1	0	0	0	1	0	0	0	1	0

3.2 Results

The results of the lexicon study are presented by-context below (first $\sigma_1\sigma_2$, then $\sigma_2\sigma_3$, then $\sigma_1\sigma_3$). Note that the goal of this subsection is not to provide a comprehensive description and analysis of all trends in the Sundanese lexicon; the goal is only to discuss results that bear on the analysis in Section 2. Materials that provide a more complete picture of Sundanese lexical statistics are available as supplementary materials.

3.2.1 Results for $\sigma_1\sigma_2$

Results of the loglinear models for the $\sigma_1\sigma_2$ context suggest a dispreference for co-occurring [r]s and [l]s modulated by a co-existing preference for identity. This is visible in Figure 3, which plots the baseline model's predicted count for a given onset pair against its observed count.[13] Identical pairs are represented with black dots and all other pairs are represented with gray.[14] Dots above the identity line denote pairs that are more frequent than expected, given the individual probabilities of each onset; dots below the line denote pairs that are less frequent than expected. (In these figures, N=[ŋ], J=[ɲ], j=[ʄ], and y=[j]. The interpretation of all other characters is straightforward.)

It is clear from Figure 3 that identical $\sigma_1\sigma_2$ onset pairs are overattested relative to expectation: identity is linked to a boost in frequency that cannot be explained only by reference to the independent frequency of the pair's members. In addition, l+l and r+r are underattested relative to other identical pairs. The results of the second loglinear model, which incorporates predictors referencing identity, confirm that these observations are unlikely to be due to chance. The positive coefficient in (36b) confirms that identity is linked to a significant increase in log frequency, and the negative coefficients in (36c–d) confirm that the log frequencies of l+l and r+r are lower than expected, relative to their position-specific frequencies (controlled for in (36e–h)) and the general frequency boost for identical segments. Thus in $\sigma_1\sigma_2$, evidence for a similarity preference among adjacent syllables comes from the overattested status of identical onsets. Evidence for a restriction

[13] All plots in this section were made with R's *ggplot2* and *gghighlight* packages (Wickham 2016; Yutani 2018).

[14] For each figure, interactive plots that label each dot are available as supplementary materials.

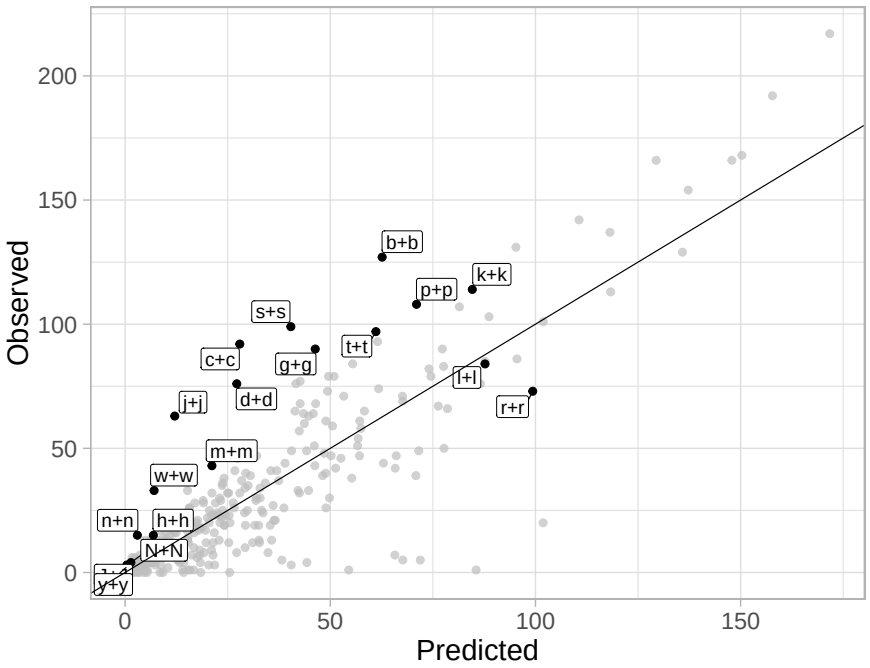


Figure 3: Predicted vs. observed frequencies of $\sigma_1\sigma_2$ onset pairs

on r+r and l+l comes from the fact that these specific pairs are underattested, relative to expectation.

(36) Partial results of loglinear model for $\sigma_1\sigma_2$ onset pairs (full results in the appendix)

	Predictor	Coefficient	*t* value	Significant?
a.	Intercept	0.14	–	–
b.	Identity	0.43	9.43	Yes ($p < .001$)
c.	l_{12}	-0.48	-3.02	Yes ($p < .01$)
d.	r_{12}	-0.64	-3.85	Yes ($p < .001$)
e.	l_1	0.28	1.13	No ($p > .1$)
f.	l_2	0.56	2.37	Yes ($p < .05$)
g.	r_1	0.29	1.17	No ($p > .1$)
h.	r_2	0.64	2.74	Yes ($p < .01$)

More evidence for aggressive reduplication comes from a positive correlation between the rates of onset-matching and nucleus-matching: while 84.5% of syllables with matching onsets have matching nuclei, only 38.2% of syllables without matching onsets have matching nuclei (37).

(37) Onset-matching encourages nucleus-matching in $\sigma_1\sigma_2$ (χ^2 (1) = 875.00, $p < .001$)

	Nucleus match	Nucleus mismatch
Onset match	962	176
Onset mismatch	3231	5235

Before moving on to address the patterns in $\sigma_2\sigma_3$, it is necessary to address a potential confound. Sundanese employs partial reduplication in a variety of morphological contexts, as attested in pairs like *basa* 'language' *ba-basan* 'proverb', *saur* 'to speak' *sa-sauran* 'to talk together', *tani* 'agriculture' *ta-tanen* 'to farm', and others (Robins 1959:360–361). It is possible that the preference for adjacent syllable identity in this context

could be due to the dictionary's inclusion of a large number of morphologically reduplicated forms.

To determine whether or not this alternative interpretation of the results is plausible, I limited the Sundanese roots under investigation to those of the shape CV*x*.CV*x*, where *x* is an optional coda. The majority (97%) of words in the dictionary are two syllables or longer, suggesting a dispreference for monosyllabic words.[15] Given this, it is reasonable to expect that most disyllabic words are not morphologically reduplicated. Words consisting of two identical syllables were however excluded if the repeated syllable was recorded as a monosyllabic word; these exclusions brought the number of forms considered down from 6,409 to 6,373. Figure 4 demonstrates that, in this subset of the data, identical onsets are still overattested. A loglinear model similarly finds a boost in frequency for identical pairs ($p < .001$) and a decrease in frequency for r+r ($p < .05$) and l+l ($p = .06$). These findings suggest that morphological reduplication is not responsible for the preference for identical onsets apparent in Figure 3.

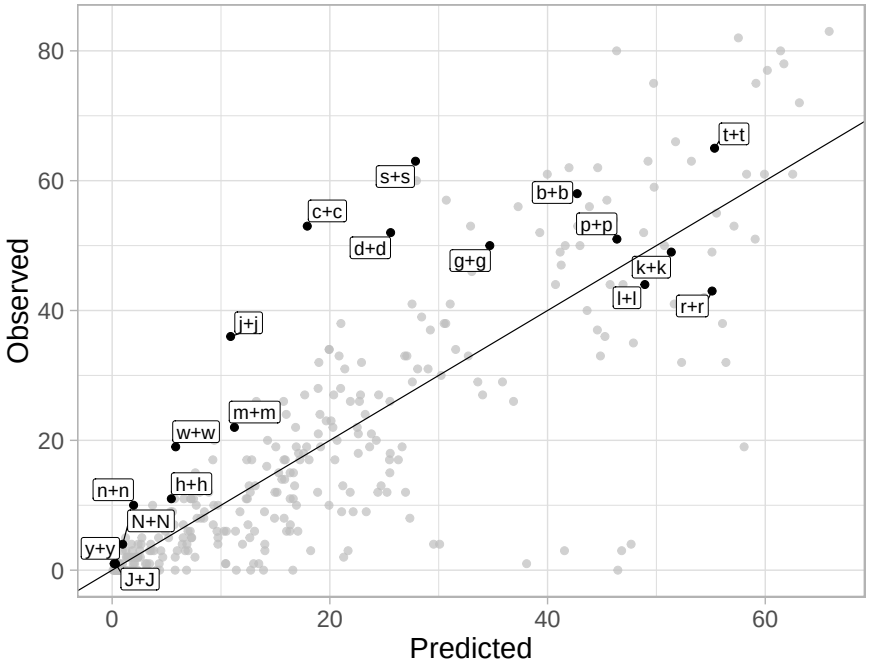


Figure 4: Predicted vs. observed frequencies of $\sigma_1\sigma_2$ onset pairs, from disyllabic subset

Similarly, morphological reduplication is likely not responsible for the link between onset-matching and nucleus-matching. Even when we focus on the subset of disyllabic forms, syllables with matching onsets are still disproportionately likely to have matching nuclei (38).

(38) Onset-matching encourages nucleus-matching in $\sigma_1\sigma_2$ (χ^2 (1) = 409.33, $p < .001$)

	Nucleus match	Nucleus mismatch
Onset match	473	159
Onset mismatch	1932	3809

In short, the properties of $\sigma_1\sigma_2$ investigated in this section are consistent with the analysis proposed in Section 2: we see a preference for identical onsets, as well as a dispreferrence for [l]…[l] and [r]…[r]. (Note however that the interrelation between the preference for identical onsets and the co-occurrence constraints

[15] Abby Cohn (p.c.) confirms this dispreference, noting that monosyllabic content words are not frequent in Sundanese. When they do occur, they are usually minimally CVC. For discussion of similar facts from Indonesian, see Cohn (2005).

is not predicted by the analysis, as the analysis is silent on how these pressures should interact in gradient lexical data.) Furthermore, it is unlikely that the observed preference for self-similarity between σ_1 and σ_2 can be attributed to morphological reduplication: the preference is also observed within a set of forms that are likely not morphologically reduplicated.

3.2.2 Results for $\sigma_2\sigma_3$

The patterns observed in $\sigma_2\sigma_3$ differ from those in $\sigma_1\sigma_2$ as a function of the rate of onset-matching. Figure 5 makes it clear that in this context there is no preference for identity among adjacent syllable onsets. But like the patterns for $\sigma_1\sigma_2$, r+r and l+l behave differently than the rest of the identical pairs. While most identical pairs are fairly close to the identity line – their frequency is predictable given the independent frequencies of their members – r+r and l+l are well below it.

These two findings were confirmed by adding identity-related predictors to the baseline model. The results (in (39)) confirm the observations made on the basis of Figure 5. The predictor for onset identity is not significant: whether or not a pair of onsets is identical has no independent effect on its log frequency. The r_{23} and l_{23} predictors are however both significant, and the negative coefficients indicate that these pairs are less frequent than expected.

(39) Partial results of loglinear model for $\sigma_2\sigma_3$ onset pairs (full results in the appendix)

	Predictor	Coefficient	*t* value	Significant?
a.	Intercept	0.19	–	–
b.	Identity	0.01	9.43	No ($p > .1$)
c.	l_{23}	-0.94	-3.02	Yes ($p < .001$)
d.	r_{23}	-0.71	-3.85	Yes ($p < .001$)
e.	l_2	0.94	1.13	No ($p > .1$)
f.	l_3	0.38	2.37	Yes ($p < .05$)
g.	r_2	1.01	1.17	No ($p > .1$)
h.	r_3	0.37	2.74	Yes ($p < .01$)

The results for $\sigma_2\sigma_3$ are similar to those for $\sigma_1\sigma_2$ in that syllables with identical onsets are disproportionately likely to have matching nuclei. This is evident in (40), where 66.7% of syllable pairs with matching onsets but only 49.6% of syllables with mismatching onsets have matching nuclei.

(40) Onset-matching encourages nucleus-matching in $\sigma_2\sigma_3$ (χ^2 (1) = 13.77, $p < .001$)

	Nucleus match	Nucleus mismatch
Onset match	86	43
Onset mismatch	1438	1463

In sum, underattestation of r+r and l+l is consistent with the activity of *[r]…[r] and *[l]…[l]. The observation in (40) that similarity along one dimension is correlated with similarity along another is consistent with a preference for self-similarity between all adjacent pairs of syllables and not just $\sigma_1\sigma_2$. Finally, the preference for onset-matching in $\sigma_1\sigma_2$ but not $\sigma_2\sigma_3$ is potentially attributable to a higher drive for self-similarity for $\sigma_1\sigma_2$; this is consistent with the activity of REDUP-$\sigma_1\sigma_2$.

One generalization evident from the properties of $\sigma_1\sigma_2$ and $\sigma_2\sigma_3$ is that σ_2's onset is frequently occupied by [l] or [r] ((36f,h); (39e,g)). One might ask if this is due to morphology, and, in particular, to the dictionary's potential inclusion of plural forms (like *karusut*, *halormat*). A search through Lembaga Basa & Sastra Sunda (1985) for potential singular-plural pairs, however, suggests that the dictionary does not record plurals. To identify potential plural forms, I created a list containing the subset of words considered here that have [a] as the rime of the first syllable and [l] or [r] as the onset of the second (e.g. *garala*ŋ 'a long scar', *balida* 'knife fish'). Possible singulars were identified by removing the *al* or *ar* from the potential plural (resulting

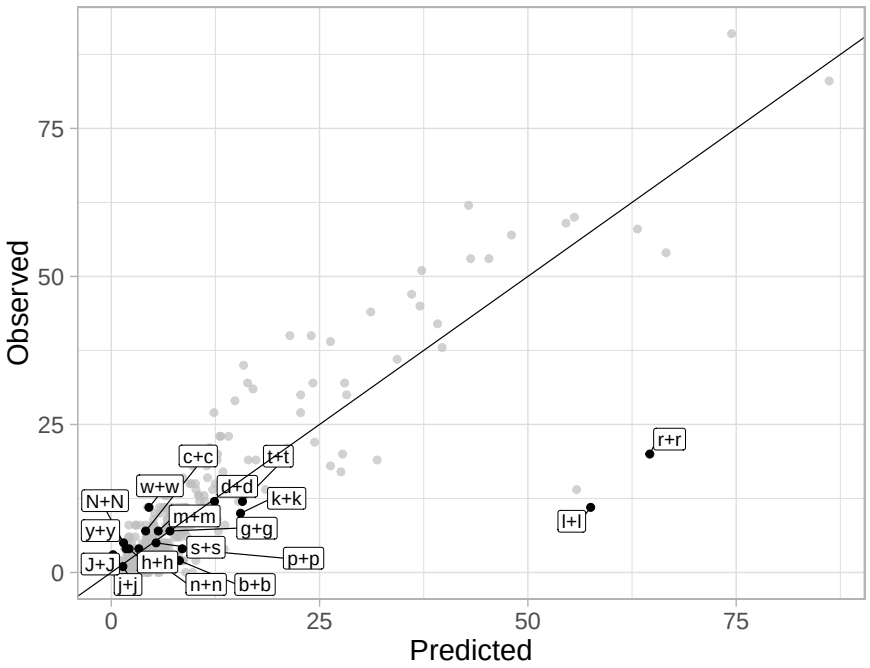


Figure 5: Predicted vs. observed frequencies of $\sigma_2\sigma_3$ onset pairs

in *galaŋ*, *bida*) and searching the wordlist again for the result.

The majority of forms (722/935, or 77%) that qualify as a potential plural do not have a corresponding potential singular in the wordlist. Of the 214 potential plurals that do, 81 do not obey the generalizations regarding the distribution of [al~ar] (these are forms like *calacah* 'cigarette ash', where *ar* is expected, or *laruŋ* 'missed', where *al* is expected), leaving 133 phonologically plausible plurals with a potential singular pair. Examples are *dapon* 'not determined, careless' and *darapon* 'at random', *jujur* 'honest' and *jalujur* 'sewing with hand before using sewing machine'; pairs like *o* 'sound like about to vomit' and *aro* 'fly' were included even though the singular is likely subminimal. Given the small number of these forms relative to the size of the overall corpus (11,913 forms), it is unlikely that plurals are regularly recorded.

Nonetheless, I reran the statistics for $\sigma_1\sigma_2$ and $\sigma_2\sigma_3$ while excluding these 133 plausibly plural forms. There were no resulting qualitative changes. For $\sigma_1\sigma_2$, identical pairs are still overattested ($p < .001$), l+l and r+r are still underattested ($p < .01$ for both), and the presence of [l] or [r] in the second syllable's onset is still associated with an increase in log frequency ($p < .05$ for both). For $\sigma_2\sigma_3$, there is still no effect of identity on log frequency ($p > .1$), l+l and r+r are still underattested ($p < .001$ for both), and the presence of [l] or [r] in the second syllable is still associated with an increase in log frequency ($p < .001$ for both). Even if the 133 forms identified as plausible plurals are in fact plurals, it cannot be the case that their inclusion is responsible for the high frequency of [l] and [r] as the second syllable's onset. It is not clear to me that there is an insightful explanation for the high frequency of liquids in this position beyond some arbitrary phonotactic preference.

3.2.3 Results for $\sigma_1\sigma_3$

The $\sigma_1\sigma_3$ context differs from $\sigma_1\sigma_2$ and $\sigma_2\sigma_3$ in that it involves non-adjacent syllables. The analysis predicts that in this non-adjacent context there should be no drive for self-similarity, and (as a result) that combinations of [r]s and [l]s should be significantly underattested. Figure 6 plots the observed count for each $\sigma_1\sigma_3$ onset

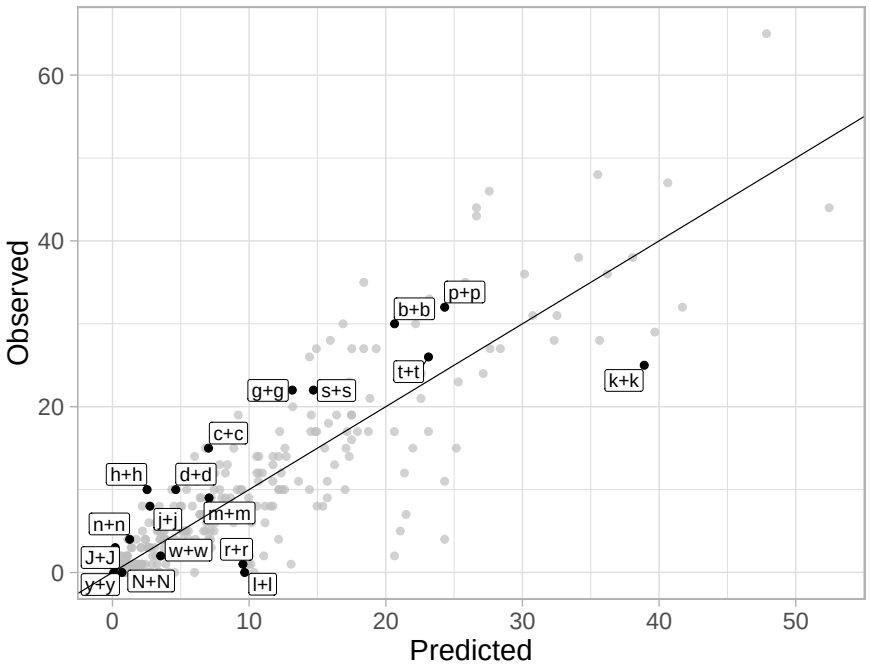


Figure 6: Predicted vs. observed frequencies of $\sigma_1\sigma_3$ onset pairs

pair against its predicted count. The shape of the $\sigma_1\sigma_3$ data looks similar to the shape of the $\sigma_1\sigma_2$ data: there is a preference for onset identity, with a concomitant dispreference for r+r and l+l (and additionally in this context, k+k).

To determine if these trends are meaningful, identity-based predictors were added to the baseline model. The results are consistent with Figure 6: there is a boost in log frequency for identity (41b) and a decrease in log frequency for r+r and l+l (41c–d) relative to other identical pairs and the independent frequencies of [r]s and [l]s (41e–h).

(41) Partial results of loglinear model for $\sigma_1\sigma_3$ onset pairs (full results in the appendix)

	Predictor	Coefficient	*t* value	Significant?
a.	Intercept	0.19	–	–
b.	Identity	0.16	2.84	Yes ($p < .01$)
c.	l_{13}	-1.51	-1.95	Trending ($p = .052$)
d.	r_{13}	-1.06	-1.99	Yes ($p < .05$)
e.	l_1	0.03	0.12	No ($p > .1$)
f.	l_3	0.33	1.43	No ($p > .1$)
g.	r_1	0.04	0.16	No ($p > .1$)
h.	r_3	0.30	1.31	No ($p > .1$)

The effect of identity is surprising, as the analysis does not predict a preference for self-similarity between non-adjacent syllables. A closer look at the 231 forms with identical onsets, however, suggests that this number is likely inflated by a type of discontinuous reduplication. Of these 231 forms, 74 have a third syllable that is composed of the first syllable's onset and the second syllable's rime (e.g. *balingbing* 'starfruit', *corodcod* 'shaky leg', *harashas* 'dried palm leaf', *perekpek* 'take a beating / got beaten'). While it is unclear if this process is synchronically active, similar patterns of discontinuous reduplication are attested in dialects

of closely related Malay (see Kroeger 1989).[16] As it is possible that the self-similarity in these cases is enforced by some morphophonological process, it is worthwhile to consider what the data would look like were these 74 forms excluded. Figure 7 confirms that they do in fact look quite different.

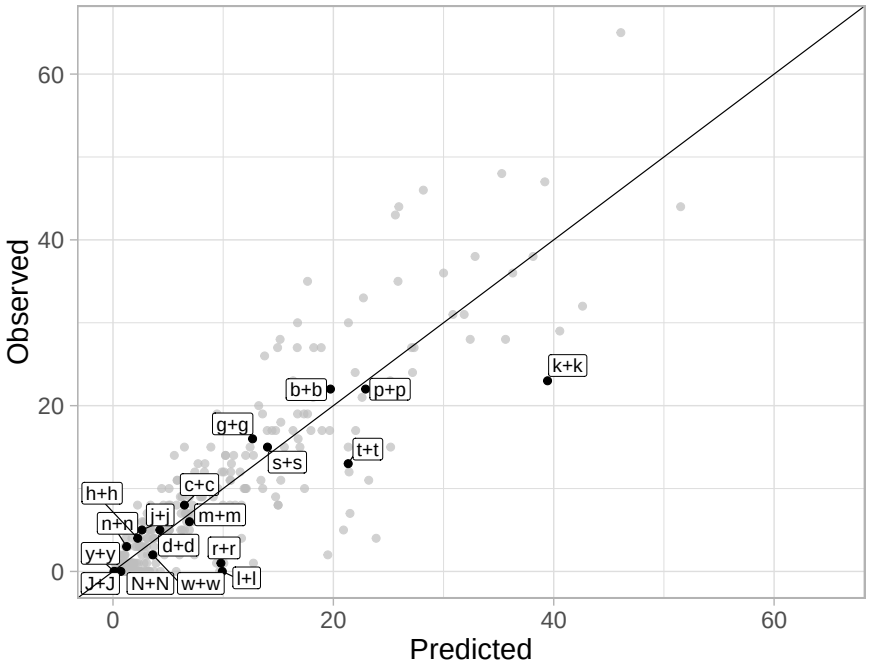


Figure 7: Predicted vs. observed frequencies of $\sigma\sigma$-13 onset pairs, without reduplicated forms

Restrictions on [r] and [l] co-occurrence are still apparent: there are no forms with [l+l] in the first and third syllables, and only one with [r+r] (*rudira* 'blood'). Yet in Figure 5, the apparent preference for identity has vanished. A second loglinear model, fit to the data visualized in Figure 5, finds no increase or decrease in frequency associated with identity ($p > .1$) and near-significant frequency decrements associated with r+r ($p < .1$ for both); it is likely that the lack of significance in these cases is due to a lack of statistical power.[17] These results are consistent with the assumptions of Section 2's analysis: the co-occurrence restrictions, but not the drive for identity, hold in the non-local $\sigma_1\sigma_3$ context.

The suggestion that there is no drive for identity in non-adjacent contexts is supported by the lack of a relationship between onset-matching and nucleus-matching in this context. Even when the 74 potentially reduplicated forms are included, the rate of onset-matching is not significantly correlated with the rate of nucleus-matching (42). When these 74 forms are excluded, the number of forms with matching onsets decreases (nucleus match, n=71; nucleus mismatch, n=74) and the correlation remains insignificant (χ^2 (1) = 1.00, $p > .1$).

(42) Onset-matching does not encourage nucleus-matching in $\sigma_1\sigma_3$ (χ^2 (1) = 1.48, $p > .1$)

	Nucleus match	Nucleus mismatch
Onset match	107	112
Onset mismatch	1204	1510

[16] Further evidence that the Sundanese forms were at one point morphologically complex comes from their phonotactics: the sources I have found (e.g. Van Syoc 1959; Cohn 1992) do not list [dc] (in *corodcod*) or [kp] (in *perekpek*), among other clusters attested in these forms, as licit morpheme-internal clusters.

[17] Excluding these 74 forms does not change any of the results from $\sigma_1\sigma_2$ and $\sigma_2\sigma_3$, so I don't revisit them.

In sum, the $\sigma_1\sigma_3$ data provide further evidence for co-occurrence restrictions on [r]s and [l]s: [l]s do not co-occur and [r]s co-occur only rarely. Furthermore, the lack of a relationship between onset-matching and nucleus-matching is consistent with the assumption encoded in REDUP that corresponding substrings must be adjacent: in non-adjacent syllables, similarity along one dimension is not correlated with similarity along another. This conclusion is further supported by the lack of onset-matching in $\sigma_1\sigma_3$, visible when 74 potentially reduplicated forms are excluded.

3.2.4 Local summary

The markedness constraints proposed in Section 2 to account for /ar/ allomorphy potentially predict language-wide effects of liquid dissimilation (driven by *[r]…[r], *[l]…[l]) and aggressive reduplication (driven by REDUP, REDUP-$\sigma_1\sigma_2$, and $\kappa\kappa$·IDENT). While corroborating evidence from other synchronic processes is limited, I have shown here that each of these constraints has echoes in the Sundanese lexicon.

One general finding is that l+l and r+r are dispreferred relative to their expected frequencies in all positions within the word, as would be expected if *[l]…[l] and *[r]…[r] were active. While the above discussion focuses only on their co-occurrence in onset position, co-occurrence is likely underattested in all contexts (as the analysis predicts). To examine the rates of co-occurrence more broadly, I searched through all forms in Lembaga Basa & Sastra Sunda (1985) (n=16,238) for words that contain more than one [r] or more than one [l]. For [r]: the dictionary contains only 247 forms with multiple [r]s, and 200 can be interpreted as involving total reduplication (*biribiri* ‘thiamin deficiency’) or partial/aggressive reduplication (*rereb* ‘stay overnight on the road’). Many of the remaining 47 are likely loans, though they are not necessarily annotated as such (*kolaborator* ‘a person who helps the opponent’, *barometer* ‘barometer’, *organisator* ‘a person capable of setting a meeting’). For [l]: 226 forms contain more than one [l], and 204 of these cases can be interpreted as involving total reduplication (*lapatlapat* ‘blurry vision because the object is too far’) or partial/aggressive reduplication (*lalab* ‘vegetables served raw / salad’, *loloco* ‘mashing, pounding’). Again, of the remaining 22, many are loans (*kolonial* ‘invasion’). The low frequency of [r] and [l] co-occurrence outside of reduplicative contexts is consistent with an analysis that treats /ar/ allomorphy as resulting in part from co-occurrence constraints on [r]s and [l]s. It is worth noting that while the existence of a dispreference for co-occurring [r]s is consistent with most analyses of Sundanese /ar/ allomorphy, the existence of a dispreference for co-occurring [l]s is uniquely consistent with the proposed analysis, as it is the only analysis I am aware of that posits *[l]…[l] (see Section 2.3).

Another general finding is that in adjacent but not non-adjacent syllables, onset identity encourages nucleus identity (consistent with the activity of REDUP and the IDENT-$\kappa\kappa$ constraints it activates). In addition, onsets are more likely to match in $\sigma_1\sigma_2$ than is naïvely expected (an observation consistent with the position-specific REDUP-$\sigma_1\sigma_2$). These findings hold even when potentially reduplicated forms are excluded from the analysis, underscoring the point that in Sundanese there exists an entirely phonological drive for self-similarity between adjacent syllables.

A word is necessary here regarding the relationship between these lexical trends and the analysis of /ar/ allomorphy. The analysis proposed in Section 2 uses categorical constraints, but the trends in the lexicon are gradient. There are at least two possible ways to understand why this difference exists. One possibility is that the right analysis of the alternations in Section 2 is actually a probabilistic analysis that makes gradient predictions about both alternations and the lexicon. Motivation for such a claim could come from variation in the realization of /ar/: perhaps /ar-liren/ is realized as [l-al-iren] most of the time, but less frequently as the more self-similar [l-il-iren]; perhaps /ar-curiga/ is usually realized as [c-ar-uriga] but occasionally as [c-ar-iriga]. Knowing whether or not this is the correct approach would require more data on speaker judgments and productions than is currently available. The second possibility is that the relationship between the grammar and the lexicon is indirect, in the way outlined by Martin (2007). Under this scenario, constraints like *[l]…[l], $\kappa\kappa$·IDENT-[onset], $\kappa\kappa$·IDENT-[nucleus] play a role in determining which words are more likely

to be coined and accepted by speakers, but do not act on those words directly. Thus the relative rarity of words containing multiple [r]s and [l]s, as well as the prevalence of words with self-similar adjacent syllables, is due not to any active phonological process but rather to speakers' relative unwillingness to accept and continue to use words that violate active markedness constraints. As many other pressures likely help shape the lexicon (e.g. a desire to faithfully render loanwords from languages with different phonotactics), it would be surprising if each active markedness constraint held categorically.

3.3 Lexical evidence and learnability

The discussion in Section 3.2 shows that the constraints proposed in my analysis of /ar/ allomorphy (Section 2) have echoes in the Sundanese lexicon. Recent work shows however that speakers are not always aware of statistically significant trends in the lexicon (e.g. Becker, Ketrez & Nevins 2011), so it is not necessarily the case that there should be a correlation between the constraints apparently implicated by statistical trends in the lexicon and the constraints that drive phonological alternations. The question then is why we should take seriously the lexical evidence outlined above as support for the analysis in Section 2.

This subsection outlines a potential learnability-based argument. One striking fact about descriptions of Sundanese /ar/ allomorphy is that, despite being a complex process limited to a single morphological context, it appears to be reliably acquired: descriptions of the pattern by Robins (1959), Van Syoc (1959), Cohn (1992), and Bennett (2015a,b) are mutually reinforcing (and all appear to rely at least in part on their own primary data). As it is a stable, reliably acquired pattern, its analysis should be easily learnable given the input available to a child. Based on evidence from a large Sundanese corpus, I suggest that ~.02% of a learner's input would provide them with evidence that [ar~al] alternations exist. In order to acquire these alternations, the learner would thus need to posit a complicated set of rankings based on comparatively few forms. Links between morphophonology and the lexicon would make the child's task easier, as the constraints and potentially the rankings among them could be induced at least in part from the larger lexicon.

This subsection focuses on quantifying the evidence for [ar~al] alternations and stops short of implementing a computational learner to demonstrate that the necessary constraints and their ranking can be induced from the lexical evidence. The discussion remains speculative in this way because there is no currently implemented phonotactic learner that can induce the representations and constraints assumed by Zuraw (2002), nor is there a currently implemented learner that can find non-local-only dissimilation. While the Inductive Phonotactic Learner (Gouskova & Gallagher 2020) can discover non-local restrictions, to do so it must first discover a restriction that holds within a trigram (e.g. *X[]X). But the evidence for *r[]r and *l[]l is muted in Sundanese and thus not discoverable by any algorithm that requires evidence for a local co-occurrence restriction to justify searching for a non-local one.[18]

3.3.1 Corpus and methodology

To approximate the frequency of words containing plural /ar/, I extracted all potential singular-plural pairs from the Sundanese An Crúbadán corpus (Scannell 2007), which comprises 713,970 tokens. Potential plurals were forms with an *al* or *ar* sequence that is both followed by and not preceded by a vowel; forms like *laloba* 'many, abundant, plenty *pl.*' were considered but forms like *regional* 'regional' were not. Potential singulars were identified by removing *ar* or *al* from the plural and searching the wordlist for the resulting singular. Thus for *laloba*, the wordlist was searched for *loba*. A singular-plural pair was recorded if the corresponding singular exists and has a higher token frequency than the plural. This frequency criterion was

[18] I confirmed this with an IPL simulation on the full list of native Sundanese words, with a gain of 150 and a goal of discovering 100 constraints. The baseline simulation discovers a number of constraints that suggest the existence of local vowel harmony (like *[-tense][][+tense], *[+high][][-high, -low]), but none that suggest the existence of local co-occurrence restrictions on liquids. Since no relevant trigram placeholder constraints were discovered in the baseline simulation, the learner does not know to look for constraints on non-local co-occurrence.

established based on a preliminary search through the corpus for singular-plural pairs identified in the extant literature on Sundanese /ar/ allomorphy (Robins 1959; Cohn 1992; Bennett 2015a,b). The findings indicate that productively derived plurals are less frequent than the singulars; the mean token frequency for words containing a singular form was 173.9 and the mean token frequency for words containing a plural form was 52.4.[19] Several examples with their associated frequencies are in (43).

(43) Existing singular-plural pairs and their token frequencies

	Singular	Plural	Gloss	Frequencies
a.	*kusut*	*karusut*	'messy (pl.)'	8, 1
b.	*dahar*	*dalahar*	'eat (pl.)'	580, 28
c.	*poho*	*paroho*	'forget (pl.)'	128, 5

Given the general trend for singulars to be more frequent than the corresponding plurals, a search that limits plausible singular-plural pairs to those with a more frequent singular is justifiable.

3.3.2 Findings

The search discovered a total of 991 plausible singular-plural pairs. The token frequency of the plural forms sums to 6,239, meaning that approximately 0.1% of the tokens in the corpus are plausibly pluralized forms. This is a conservative estimate, as neither /ar/'s location nor the semantics of the plural were considered when deciding whether or not a pair was plausible. In other words, pairs like *hal* 'thing' and *halal* 'halal', *tatu* 'a wound from war or accident' and *tatalu* 'hitting the tip of one's fingers or palms against any hard surface to make sounds (music)' were counted as plausible pairs, even when the "plural" is likely a simplex word (*halal*) or the affix does not occur before the initial vowel (*tatalu*). (I included pairs like *tatu-tatalu* because some prefixes can attach outside of /ar/; a prefixed form from Robins 1959:344, where the affix does not occur before the initial vowel, is *di-bawa – di-barawa*, 'to be carried (pl.)'. Semantics were not considered because glosses are not provided in the corpus.)

Not all of these 991 plausible plurals are informative about the ranking governing /ar/ allomorphy, as most lack another liquid. Recall from Section 2 that in roots that do not contain a liquid, IO·IDENT-[±lateral] prohibits /ar/ from being realized as anything but [ar]. The number of plausible plurals whose stem contains a liquid is much smaller, at 353, and their frequency amounts to 1,179 tokens. Assuming that the An Crúbadán corpus is broadly representative of the types of words that the Sundanese learner encounters, the implication is that only .02% of words the learner encounters would provide evidence as to the ranking of the various constraints proposed in Section 2.[20] While this may well be enough information for the learner to arrive at the correct ranking – alternations are salient and .02% of a child's input is likely still a large number of words – the links established here between phonology and the lexicon mean that the child's acquisition of /ar/ allomorphy may be bolstered by trends discoverable in the lexicon. In other words, it may be easier for the Sundanese learner to discover the proposed analysis than an alternative that treats the /ar/ allomorphy as an idiosyncratic property divorced from the larger lexicon (cf. Anderson 1993:78; see also Pierrehumbert

[19] There is a small, apparently closed class of nouns that exceptionally take [ar~al] as the plural morpheme, and in four cases the plural is more frequent than the singular. These exceptions make sense when their meanings are considered: *budak-barudak* 'child-children' (where 'children' can be used generally to refer to young people) 338/346, *maneh-maraneh* '2nd person (low register) sg-pl' 1083/1425, *manehna-maranehna* '3rd person (low register) sg-pl' 619/749, *manehanana-maranehana* '3rd person (low register) alternate form sg-pl' 0/24 (thanks to Abby Cohn, p.c., for the glosses). As far as I am able to tell, none of the forms exhibiting the [ar~al] alternations are nouns.

[20] I do not include breakdowns of how many tokens would support each ranking because there are a number of apparent exceptions (35 plausible plurals, or 151 tokens) to the distribution of /ar/'s allomorphs described in the literature. In most cases this is likely due to prefixation: pairs like *salabar* 'making announcement' and *salalabar* 'making announcement (pl.)' appear to exhibit [l]-assimilation in an unexpected context, but Sundanese has a nominal prefix *sa-* (Robins 1959:352) and so it is probable that [l]-assimilation has applied as expected in stem-initial position. Most of the apparent exceptions have a plausible reanalysis along these lines.

2003 on how learners do not draw morpholophonological generalizations from small amounts of data).

4 Discussion

This paper has shown that Sundanese /ar/ allomorphy can be analyzed as resulting from unbounded co-occurrence restrictions on [r]s and [l]s, whose effects in local contexts are obscured by a general desire for identity between adjacent syllables. Statistical trends from the lexicon are consistent with this analysis. I have suggested that this connection between /ar/ allomorphy and the lexicon may function as an argument for the proposed analysis, as the evidence that would be required for a learner to acquire the crucial rankings governing /ar/ allomorphy is otherwise likely infrequent.

Recall that our interest in the Sundanese data is in how they bear on the predictions of two competing theories of dissimilation: Suzuki 1998's GOCP, in which dissimilation is motivated by co-occurrence constraints; and Bennett's (2015a,b) SCTD, in which dissimilation is a way of avoiding similarity-based surface correspondence. The GOCP predicts that non-local-only dissimilation should only arise given the coexistence of some independent pressure that disprefers the results of local dissimilation. As discussed above, Sundanese – which has the only known case of non-local-only dissimilation – fits this description. In addition to cases like Sundanese, the SCTD predicts cases of non-local-only dissimilation that cannot be analyzed by invoking constraints that disprefer the results of local dissimilation. This prediction is not supported by the typological data. Furthermore, results from artificial grammar learning experiments parallel the typological data. McMullin & Hansson (2016) show that participants are able to acquire the kinds of non-local dissimilation predicted by both the GOCP and the SCTD, where a non-local restriction on identical liquids (*lVCVl, *rVCVr; lVCVr, rVCVl) accompanies a restriction on local non-identical liquids (lVl, rVr; *lVr, *rVl). McMullin & Hansson (2019) however show that participants are not able to reliably learn non-local dissimilation when it is not accompanied by local assimilation, regardless of whether or not they are presented with overt evidence for non-alternation in local contexts. These findings suggest that the type of non-local dissimilation uniquely predicted by the SCTD is not only unattested but also unlearnable, and that the correct theory of dissimilation should not treat it as part of the learner's hypothesis space.

Prior work has shown that the SCTD fails to make accurately restrictive predictions in other domains as well. For example, Stanton (2017) shows that the GOCP predicts a more restricted typology of blocking in long-distance dissimilation than does the SCTD, and that all known relevant cases are consistent with the GOCP's predictions. In addition, Stanton (2016b) shows that the SCTD fails to derive a generalization regarding the role of similarity in dissimilation. Generally speaking, if a language disprefers co-occurrence of two less similar segments it also disprefers co-occurrence of more similar segments (the only exceptional cases in this respect involve fully identical segments; see e.g. MacEachern 1997; Gallagher & Coon 2009; Gallagher 2013 for discussion and analysis). But the SCTD predicts the opposite similarity implication: all else being equal, dissimilation of two more similar segments should imply dissimilation of less similar segments. To give a concrete example, the SCTD can generate a system in which /p p/ and /p f/ can co-occur, but /p v/ is banned (Stanton 2016b:539). The typology of dissimilation suggests no cases with this character.

An argument offered by Bennett (2015a,b) for the SCTD is that it unifies the analysis of long-distance assimilation and dissimilation: the theory's predictions regarding the typology of dissimilation follow directly from its analysis of the typology of assimilation. The work reported in this paper and cited above, however, suggests that the SCTD's predictions in the domains of locality and similarity are not sufficiently restrictive. These results, in turn, raise the question of whether Bennett's theoretically elegant unification of two disparate typologies should come at the expense of restrictiveness. My position is that it should not, and that the facts reviewed here support co-occurrence-based theories of dissimilation over available correspondence-based alternatives.

Appendix: full results of statistical models

This appendix contains full results for four statistical models: the $\sigma_1\sigma_2$ model summarized in (36), the $\sigma_2\sigma_3$ model summarized in (39), the $\sigma_1\sigma_3$ model summarized in (41), and the additional $\sigma_1\sigma_3$ model in which the 74 forms that plausibly exhibit discontinuous reduplication have been excluded (see Section 3.2.3 for discussion). Further variations on these models (like those that exclude plausible plurals) are not reported here as the results did not differ qualitatively from those presented below.

Significance codes can be interpreted as follows: . = $p < .1$, * = $p < .05$, ** = $p < .01$, *** = $p < .001$. Lack of a significance code denotes a non-significant result.

Table 3: Results for $\sigma\sigma$-12 (all forms included)

Predictor	Coefficient	*t* value	Significant?	Predictor	Coefficient	*t* value	Significant?
Intercept	0.14	–		m_1	0.12	0.47	
Identical	**0.43**	**9.43**	***	m_2	-0.06	-0.27	
l_{12}	**-0.48**	**-3.02**	**	n_1	-0.69	-2.58	*
r_{12}	**-0.64**	**-3.85**	***	n_2	-0.20	-0.83	
p_1	0.45	1.84	.	ny_1	-1.09	-3.62	***
p_2	0.16	0.68		ny_2	-0.77	-2.90	**
t_1	0.30	1.22		ng_1	-0.85	-3.05	**
t_2	0.25	1.05		ng_2	-0.41	-1.65	
c_1	0.16	0.65		s_1	0.40	1.61	
c_2	0.02	0.10		s_2	-0.04	-0.18	
k_1	0.49	2.01	*	**l_1**	**0.28**	**1.13**	
k_2	0.20	0.85		**l_2**	**0.56**	**2.37**	*
b_1	0.43	1.75	.	**r_1**	**0.29**	**1.17**	
b_2	0.13	0.53		**r_2**	**0.64**	**2.74**	**
d_1	-0.15	-0.59		w_1	-0.29	-1.16	
d_2	0.31	1.33		w_2	-0.18	-0.73	
j_1	-0.04	-0.16		y_1	-1.91	-4.06	***
j_2	-0.17	-0.71		y_2	-0.31	-1.26	
g_1	0.28	1.12		h_1	-0.17	-0.67	
g_2	0.14	0.61		h_2	-0.31	-1.28	

Table 4: Results for $\sigma\sigma$-23 (all forms included)

Predictor	Coefficient	*t* value	Significant?	Predictor	Coefficient	*t* value	Significant?
Intercept	0.19	–		m_2	0.16	0.67	
Identical	**0.016**	**0.18**		m_3	-0.15	-0.64	
l_{23}	**-0.94**	**-3.72**	***	n_2	-0.17	-0.67	
r_{23}	**-0.71**	**-3.48**	***	n_3	-0.09	-0.37	
p_2	0.15	0.63		ny_2	-0.96	-3.15	**
p_3	0.06	0.27		ny_3	-0.68	-2.63	**
t_2	0.11	0.46		ng_2	-0.50	-1.89	.
t_3	0.41	1.76	.	ng_3	-0.16	-0.66	
c_2	0.04	0.16		s_2	0.01	0.06	
c_3	-0.18	-0.77		s_3	-0.03	-0.11	
k_2	0.26	1.08		**l_2**	**0.94**	**3.94**	***
k_3	0.25	1.08		**l_3**	**0.38**	**1.65**	
b_2	0.14	0.56		**r_2**	**1.01**	**4.23**	***
b_3	0.06	0.25		**r_3**	**0.37**	**1.59**	
d_2	0.21	0.88		w_2	-0.19	-0.75	
d_3	0.19	0.81		w_3	0.09	0.37	
j_2	-0.40	-1.54		y_2	-0.38	-1.47	
j_3	-0.29	-1.18		y_3	-0.18	-0.77	
g_2	0.10	0.41		h_2	-0.32	-1.27	
g_3	0.02	0.09		h_3	-0.15	-0.62	

Table 5: Results for $\sigma\sigma$-13 (all forms included)

Predictor	Coefficient	*t* value	Significant?	Predictor	Coefficient	*t* value	Significant?
Intercept	-2.39	–		m_1	0.27	1.10	
Identical	**0.16**	**2.84**	**	m_3	-0.16	-0.66	
l_{13}	**-1.51**	**-1.95**	**.**	n_1	-0.65	-2.35	*
r_{13}	**-1.06**	**-1.99**	*	n_3	-0.09	-0.40	
p_1	0.66	2.69	**	ny_1	-1.00	-3.27	**
p_3	0.06	0.25		ny_3	-0.65	-2.55	*
t_1	0.28	1.15		ng_1	-0.89	-3.02	**
t_3	0.41	1.79	.	ng_3	-0.14	-0.60	
c_1	0.30	1.21		s_1	0.49	2.00	*
c_3	-0.19	-0.78		s_3	-0.02	-0.07	
k_1	0.70	2.86	**	**l_1**	**0.03**	**0.12**	
k_3	0.25	1.07		**l_3**	**0.33**	**1.43**	
b_1	0.58	2.35	*	**r_1**	**0.04**	**0.16**	
b_3	0.06	0.27		**r_3**	**0.30**	**1.31**	
d_1	-0.32	-1.24		w_1	-0.33	-1.26	
d_3	0.23	0.98		w_3	0.10	0.42	
j_1	-0.09	-0.37		y_1	-1.85	-3.69	***
j_3	-0.26	-1.08		y_3	-0.16	-0.67	
g_1	0.40	1.61		h_1	-0.22	-0.87	
g_3	0.02	0.10		h_3	-0.17	-0.71	

Table 6: Results for $\sigma\sigma$-13 (potentially reduplicated forms excluded)

Predictor	Coefficient	*t* value	Significant?	Predictor	Coefficient	*t* value	Significant?
Intercept	-2.33	–		m_1	0.28	1.13	
Identical	**-0.06**	**-0.85**		m_3	-0.15	-0.65	
r_{13}	**-0.87**	**-1.68**	.	n_1	-0.65	-2.36	*
l_{13}	**-1.33**	**-1.75**	.	n_3	-0.09	-0.40	
p_1	0.66	2.72	**	ny_1	-1.11	-3.50	***
p_3	0.06	0.28		ny_3	-0.68	-2.69	**
t_1	0.28	1.14		ng_1	-0.88	-3.01	**
t_3	0.41	1.79	.	ng_3	-0.14	-0.59	
c_1	0.29	1.19		s_1	0.49	2.02	*
c_3	-0.20	-0.86		s_3	-0.02	-0.07	
k_1	0.72	2.96	**	**l_1**	**0.04**	**0.14**	
k_3	0.28	1.22		**l_3**	**0.33**	**1.44**	
b_1	0.58	2.38	*	**r_1**	**0.05**	**0.19**	
b_3	0.07	0.30		**r_3**	**0.30**	**1.32**	
d_1	-0.34	-1.32		w_1	-0.31	-1.19	
d_3	0.22	0.97		w_3	0.10	0.45	
j_1	-0.10	-0.38		y_1	-1.84	-3.76	***
j_3	-0.27	-1.12		y_3	-0.16	-0.66	
g_1	0.40	1.64		h_1	-0.25	-0.98	
g_3	0.03	0.11		h_3	-0.19	-0.80	

References

Alderete, John. 1997. Dissimilation as local conjunction. In Kiyomi Kusumoto (ed.), *Proceedings of the North East Linguistics Society (NELS)*, vol. 27, 17–32. Amherst, MA: Graduate Linguistics Student Association.

Anderson, Edmund A. 1997. The use of speech levels in Sundanese. In M. Clark (ed.), *Papers in Southeast Asian linguistics No. 16*, 1–45. Canberra: Pacific Linguistics. https://doi.org/10.15144/PL-A90.1.

Anderson, Stephen R. 1993. Wackernagel's revenge: Clitics, morphology, and the syntax of second position. *Language* 69. 68–98. https://doi.org/10.2307/416416.

Becker, Michael, Nihan Ketrez & Andrew Nevins. 2011. The surfeit of the stimulus: Analytic biases Filter lexical statistics in Turkish laryngeal alternations. *Language* 87. 84–125. https://doi.org/10.1353/lan.2011.0016.

Becker, Michael, Andrew Nevins & Jonathan Levine. 2012. Asymmetries in generalizing alternations to and from initial syllables. *Language* 88. 231–268. https://doi.org/10.1353/lan.2012.0049.

Beckman, Jill N. 1998. *Positional faithfulness*. Amherst, MA: UMass Amherst dissertation.

Bennett, William G. 2015a. Assimilation, dissimilation, and surface correspondence in Sundanese. *Natural Language & Linguistic Theory* 33. 371–415. https://doi.org/10.1007/ s11049-014-9268-2.

Bennett, William G. 2015b. *The phonology of consonants: Harmony, dissimilation, and correspondence*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139683586.

Cohn, Abigail C. 1992. The consequences of dissimilation in Sundanese. *Phonology* 9. 199–220. https://doi.org/10.1017/S0952675700001585.

Cohn, Abigail C. 2005. Truncation in Indonesian: Evidence for violable minimal words and ANCHOR-

RIGHT. In Keir Moulton & Matthew Wolf (eds.), *North East Linguistic Society (NELS)*, vol. 34, 372–381. Amherst, MA: Graduate Linguistics Student Association.

Downing, Laura. 1998. On the prosodic misalignment of onsetless syllables. *Natural Language & Linguistic Theory* 16. 1–52. https://doi.org/10.1023/A:1005968714712.

Eringa, Fokko Siebold. 1949. *Loetoeng Kasaroeng: Een mythologisch verhaal uit West-Java. Bijdrage tot de Soendase taal- en letterkunde*. 'S-Gravenhage: Martinus Nijhoff.

Ewing, Michael. 1991. Plural concord in Sundanese. Paper presented at the 6th International Conference on Austronesian Linguistics, Honolulu.

Fallon, Paul D. 1993. Liquid dissimilation in Georgian. In Andreas Kathol & Michael Bernstein (eds.), *Proceedings of the 10th Eastern States Conference on Linguistics (ESCOL)*, 105–116. Ithaca, NY: DMLL Publications.

Foley, William A. 1991. *The Yimas language of New Guinea*. Stanford, CA: Stanford University Press.

Gallagher, Gillian. 2013. Learning the identity effect as an artificial language: Bias and generalization. *Phonology* 31. 1–43. https://doi.org/10.1017/S0952675713000134.

Gallagher, Gillian & Jessica Coon. 2009. Distinguishing total and partial identity: Evidence from Chol. *Natural Language & Linguistic Theory* 27. 545–582. https://doi.org/10.1007/ s11049-009-9075-3.

Gelman, Andrew & Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511790942.

Gouskova, Maria & Gillian Gallagher. 2020. Inducing nonlocal constraints from baseline phonotactics. *Natural Language & Linguistic Theory* 38. 77–116. https://doi.org/10.1007/s11049-019-09446-x.

Hansson, Gunnar Ólafur. 2001. *Theoretical and typological issues in consonant harmony*. Berkeley, CA: University of California, Berkeley dissertation.

Hansson, Gunnar Ólafur. 2010. *Consonant harmony: Long-distance interaction in phonology*. Berkeley, CA: University of California Press.

Holton, David. 1995. Assimilation and dissimilation of Sundanese liquids. In Jill Beckman, Laura Walsh-Dickey & Suzanne Urbanczyk (eds.), *Papers in Optimality Theory*, vol. 18 (University of Massachusetts Occasional Papers). Amherst, MA.: Graduate Student Linguistics Association.

Kroeger, Paul R. 1989. Discontinuous reduplication in vernacular Malay. *Proceedings of the Berkeley Linguistic Society (BLS)* 15. 193–202. https://doi.org/10. 3765/bls.v15i0.1742.

Lembaga Basa & Sastra Sunda. 1985. *Kamus umum basa Sunda*. Indonesia: Penerbit Tarate Bandung.

MacEachern, Margaret. 1997. *Laryngeal cooccurrence restrictions*. Los Angeles: University of California, Los Angeles dissertation.

Martin, Andrew Thomas. 2007. *The evolving lexicon*. Los Angeles: University of California, Los Angeles dissertation.

McMullin, Kevin & Gunnar Ólafur Hansson. 2016. Computational and learnability properties of conflicting long-distance dependencies. Talk presented at the 24th Manchester Phonology Meeting.

McMullin, Kevin & Gunnar Ólafur Hansson. 2019. Inductive learning of locality relations in segmental phonology. *Laboratory Phonology* 10. 14. https://doi.org/10.5334/labphon.150.

Myers, Scott. 1997. OCP effects in Optimality Theory. *Natural Language & Linguistic Theory* 15. 847–892. https://doi.org/10.1023/A:1005875608905.

Pierrehumbert, Janet. 2003. Probabilistic phonology: Discrimination and robustness. In Rens Bod, Jennifer Hay & Stefanie Jannedy (eds.), *Probabilistic linguistics*, 177–228. Cambridge, MA: MIT Press.

Robins, R. H. 1959. Nominal and verbal derivation in Sundanese. *Lingua* 8. 337–369. https://doi.org/10.1016/0024-3841(59)90035-X.

Rose, Sharon & Rachel Walker. 2004. A Typology of consonant agreement as correspondence. *Language* 80. 475–531. https://doi.org/10.1353/lan.2004.0144.

Scannell, Kevin P. 2007. The Crúbadán Project: Corpus building for under-resourced languages. *Cahiers du Cental* 5. 1–10.

Stanton, Juliet. 2016a. Latin -ālis/-āris and segmental blocking in dissimilation. Ms., MIT, Cambridge, MA.
Stanton, Juliet. 2016b. Review of Bennett (2015b). *Phonology* 33. 533–544. https://doi.org/10.1017/S0952675716000233.
Stanton, Juliet. 2017. Segmental blocking in dissimilation: An argument for co-occurrence constraints. In Karen Jesney, Charlie O'Hara, Caitlin Smith & Rachel Walker (eds.), *Proceedings of the 2016 Meeting on Phonology*. Washington, DC: Linguistic Society of America. https://doi.org/10.3765/amp.v4i0.3972.
Suzuki, Keiichiro. 1998. *A typological investigation of dissimilation*. Tucson, AZ: The University of Arizona dissertation. https://doi.org/doi:10.7282/T3NC601H.
Suzuki, Keiichiro. 1999. Identity avoidance vs. identity preference: The case of Sundanese. Paper presented at the 73rd Annual Meeting of the Linguistic Society of America, Los Angeles.
Van Syoc, Wayland Bryce. 1959. *The phonology and morphology of the Sundanese language*. Ann Arbor, MI: University of Michigan dissertation.
Wickham, Hadley. 2016. *ggplot2: Elegant graphics for data analysis*. New York: Springer-Verlag. https://doi.org/10.1007/978-0-387-98141-3.
Wilson, Colin & Marieke Obdeyn. 2009. Simplifying subsidiary theory: Statistical evidence from Arabic, Muna, Shona, and Wargamay. Ms., Johns Hopkins University.
Yutani, Hiroaki. 2018. gghighlight: Highlight lines and points in 'ggplot2'. Software package. https://cran.r-project.org/web/packages/gghighlight/index.html.
Zuraw, Kie. 2002. Aggressive reduplication. *Phonology* 19. 395–439. https://doi.org/10.1017/S095267570300441X.

Juliet Stanton
Department of Linguistics
New York University
10 Washington Place
New York, NY 10003
stanton@nyu.edu