Draft of March 2019

# Aggressive reduplication and dissimilation in Sundanese[*]

Juliet Stanton
New York University

**Abstract**

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 more general drive for identity between adjacent syllables. Statistical trends in the lexicon are consistent with this analysis. I suggest that the proposed co-occurrence-based analysis may be preferable to Bennett's (2015a,b) correspondence-based one as it makes more restrictive predictions regarding the typology of locality in dissimilation.

## 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; these characterizations will be equivalent for our purposes). 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* [a**ɾ**a]
/pak-a**ɾ**a/ → [pak-a**ɾ**a] 'break open'

b. *Local dissimilation:* /. . . **ɾ**-a**ɾ**a/ → [. . . **ɾ**-a**t**a]
/ap**ɾ**-a**ɾ**a/ → [ap**ɾ**-a**t**a] 'open, spread'

c. *No non-local dissimilation:* /. . . **ɾ**. . . -a**ɾ**a/ → [. . . **ɾ**. . . -a**ɾ**a]
/a**ɾ**aŋ-a**ɾ**a/ → [a**ɾ**aŋ-a**ɾ**a] 'tear into pieces'

[*] For comments on a draft I am grateful to Gillian Gallagher and Maria Gouskova. 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.

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).

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

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

b. *Local dissimilation:* /…**r**-u**r**i/ → […**r**-u**l**i]
/asu**r**+u**r**i/ → [asu**r**-u**l**i] 'Assyrian'

c. *Long-distance dissimilation:* /…**r**…-u**r**i/ → […**r**…-u**l**i]
/ast'**r**onomia+u**r**i/ → [ast'**r**onomia-u**l**i] 'astronomical'

The third logical possibility I will refer to as *non-local* 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 rhotic dissimilation in Sundanese (Cohn 1992, Bennett 2015a,b)

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

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

c. *Long-distance dissimilation:* /…-a**r**-…**r**…/ → […-a**l**-…**r**…]
/a**r**+ŋumba**r**a/ → [ŋ-a**l**-umba**r**a] '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 treates dissimilation as a way of avoiding similarity-based surface correspondence. Both theories can generate non-local dissimilation, but they do so in different ways. Under the GOCP, non-local 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 dissimilation: cases like (3) can be generated directly, without appealing to pressures that interact with those that motivate dissimilation. The GOCP is, in this way, a more restrictive theory than the SCTD, as it predicts that non-local dissimilation is only possible given the independent activity of constraints that disprefer the result of local dissimilation (whose effects, all else being equal, should be visible elsewhere).

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 $*X\ldots Y$, 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 $*X\ldots Y$ constraint stands for a family of constraints, where "…" denotes intervening material of increasing lengths.[1] To explore the theory's predictions regarding non-local dissimilation, we will consider two constraints belonging to the *[r]…[r] family: one that penalizes co-occurring [r]s separated by only a mora (4), and one that penalizes each pair of [r]s that occur within the word (5).

(4) *[r]-$\mu$-[r]:
Assign one * for each pair of [r]s separated by a mora.

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

A factorial typology of (4–5), together with IDENT-[±lateral] ("assign one violation for each input [$\alpha$lateral] segment whose output correspondent is [-$\alpha$lateral]"), predicts two kinds of dissimliation: local (as in Yimas) and unbounded (as in Georgian). Cases of non-local dissimilation are not predicted, as neither *[r]-$\mu$-[r] nor *[r]…[r] penalizes only non-local co-occurrence. In (6), a comma indicates that the constraints could be ranked in either order, with equivalent results.

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

| *Ranking* | *Pattern* |
|---|---|
| IDENT-[±lateral] ≫ *[r]-$\mu$-[r], *[r]…[r] | No dissimilation<br>(/r-$\mu$-r/ → [r-$\mu$-r]; /r…r/ → [r…r]) |
| *[r]-$\mu$-[r] ≫ IDENT-[±lateral] ≫ *[r]…[r] | Local dissimilation<br>(/r-$\mu$-r/ → [r-$\mu$-l]; /r…r/ → [r…r]) |
| *[r]-$\mu$-[r], *[r]…[r] ≫ IDENT-[±lateral] | Unbounded dissimilation<br>(/r-$\mu$-r/ → [r-$\mu$-l]; /r…r/ → [r…l]) |

Non-local dissimilation is only predicted to occur when *[r]…[r] dominates IO-IDENT[±lateral], and is dominated in turn by a constraint that disprefers the consequences of local dissimilation. An

[1] Suzuki's proposed hierarchy is $*XY \gg *X\text{-}C_0\text{-}Y \gg *X\text{-}\mu\text{-}Y \gg *X\text{-}\mu\mu\text{-}Y \gg *X\text{-}\sigma\sigma\text{-}Y \gg \ldots \gg *X\text{-}\infty\text{-}Y$. For expositional simplicity I assume only two instantiations of each co-occurrence constraint, $*X\text{-}\mu\text{-}Y$ and $*X\ldots Y$, where $*X\ldots Y$ penalizes co-occurrence at all distances. For present purposes this simplified variant makes equivalent predictions.

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 *[$\alpha$lateral]-$\mu$-[-$\alpha$lateral] (7), though nothing rests on this specific formulation; local assimilation could also be analyzed as driven by correspondence (Rose & Walker 2004, Hansson 2010; see also below for correspondence-driven agreement).

(7) *[$\alpha$lateral]-$\mu$-[-$\alpha$lateral] (*[$\alpha$lat]-$\mu$-[-$\alpha$lat]):
Assign one * for each pair of [$\alpha$lateral] and [-$\alpha$lateral] segments separated by a mora.

When *[$\alpha$lat]-$\mu$-[-$\alpha$lat] $\gg$ *[r]-$\mu$-[r], *[r]...[r] $\gg$ IDENT-[±lateral], the resulting system exhibits [r] dissimilation in non-local contexts only. I illustrate here with two hypothetical input-output pairs, /ra-rata/ → [ra-rata], *[la-rata] and /ra-tara/ → [la-tara], *[ra-tara] (8). Note that *[r]-$\mu$-[r] must also be dominated by *[$\alpha$lat]-$\mu$-[-$\alpha$lat], but its ranking with respect to IDENT-[±lateral] is not crucial. (I assume that [r] is [-lateral] and [l] is [+lateral]; all other segments are not specified for this feature.)

(8) Non-local dissimilation in the GOCP

| /ra-rata/ | *[$\alpha$lat]-$\mu$-[-$\alpha$lat] | *[r]-$\mu$-[r] | *[r]...[r] | IDENT-[±lateral] |
|---|---|---|---|---|
| ☞ a. [ra-rata] | | * | * | |
| b. [la-rata] | *! | | | * |
| /ra-tara/ | | | | |
| a. [ra-tara] | | | *! | |
| ☞ b. [la-tara] | | | | * |

In this hypothetical system there would be other evidence for local assimilation beyond its ability to block dissimilation in local contexts. For example, high-ranked *[$\alpha$lat]-$\mu$-[-$\alpha$lat] would compel /ra-lata/ to surface unfaithfully as [la-lata] (9).

(9) More local assimilation

| /ra-lata/ | *[$\alpha$lat]-$\mu$-[-$\alpha$lat] | *[r]-$\mu$-[r] | *[r]...[r] | IDENT-[±lateral] |
|---|---|---|---|---|
| a. [ra-lata] | *! | | | |
| ☞ b. [la-lata] | | | | * |

The main point is that the GOCP cannot generate only non-local dissimilation. Rather, it emerges from an interaction of constraints that prefer unbounded dissimilation with others that disprefer the results of local dissimilation. It is not necessarily the case that non-local dissimilation must coexist with local assimilation, as the role played by *[$\alpha$lat]-$\mu$-[-$\alpha$lat] above can be played by any other constraint that disprefers (9b). To give another example, (9b) could also be ruled out by a positional faithfulness constraint that protects the word-initial segment; in such a case we might expect to find other evidence of positional faithfulness to the word-initial segment. A related case occurs in Zulu (Beckman 1998, Bennett 2015b), where labial palatalization triggered by a suffixed

/w/ (10a-b) fails to apply if the targeted labial is root-initial (10c). 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 et al. 2012, *a.o.* on initial syllable faithfulness).

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

| | | | | |
|---|---|---|---|---|
| a. | /se**ɓ**enz-w-a/ | → | [se**tʃ’**enz-wa] | ‘be worked’ |
| b. | /**ɓ**oŋg-w-a/ | → | [**ɓ**oŋg-wa] | ‘praise (pass.), be thanked’ |

In sum, the GOCP predicts that all existing cases of non-local dissimilation should be analyzable as an interaction between unbounded dissimilation and some other factor, which (all else being equal) should have observable effects elsewhere in the language.

## 1.2 Non-local dissimilation in the SCTD

Bennett’s (2015b) SCTD treats dissimilation as a way of avoiding an otherwise required correspondence relation among consonants. As is true for the closely related Agreement by Correspondence framework (Rose & Walker 2004, Hansson 2010), correspondence between surface segments is required by a set of Corr·*X* constraints, which penalize pairs of segments sharing some featural specification *X* that do not stand in correspondence with one another. Corr·[-lateral], for example, requires that all rhotics within a word stand in correspondence with one another (11).

(11) Corr·[-lateral]:
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]. An example is CC·Ident-[±anterior], which requires corresponding consonants to have identical values for [±anterior] (12).

(12) CC·Ident-[±anterior]:
Assign one * for each pair of corresponding consonants that are [$\alpha$anterior] and [-$\alpha$anterior].

In a grammar where Corr·[-lateral] and CC·Ident-[±anterior] are high-ranked, inputs like /ɹa-ɻata/ with place-distinct rhotics must surface unfaithfully (13).[2] The faithful [ɹ$_x$a-ɻ$_y$ata] (13a), where the place-distinct rhotics do not stand in correspondence, violates Corr·[-lat]. [ɹ$_x$a-ɻ$_x$ata] (13b),

[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 (see McMullin & Hansson 2016b on difficulties that the SCTD faces in generating dissimilation of identical elements). The segmental interactions analyzed here are based very loosely on trends in the lexicon of Yidiɲ; see Stanton (2017).

where the place-distinct rhotics correspond, violates CC·Ident-[±anterior]. The choice between (13c) and (13d) depends on the relative ranking of two input-output faithfulness constraints. If Ident-[±lateral] ≫ Ident-[±anterior], the result is correspondence and place assimilation (13c). If Ident-[±anterior] ≫ Ident-[±lateral], the result is rhotic dissimilation (13d). In this latter case, dissimilation vacuously satisfies Corr·[-lat] by mapping one of the input rhotics to a lateral.

(13) High-ranked Corr·[-lat] and CC·Ident-[±ant] can drive assimilation or dissimilation

| /ɹa-ɻata/ | Corr[-lat] | CC-ID[±ant] | IO-ID[±lat] | IO-ID[±ant] |
|---|---|---|---|---|
| a. [ɹ$_x$a-ɻ$_y$ata] | *! | | | |
| b. [ɹ$_x$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. As a consequence, the SCTD makes a set of predictions about how the typologies of long-distance assimilation and dissimilation should interrelate. Bennett (2015b) terms this property of the theory the "mismatch hypothesis", and argues that it correctly characterizes many aspects of the two typologies. I focus here only on the prediction regarding the role of locality.[3]

Many cases of long-distance consonant assimilation only occur across a single syllable boundary. In Ndonga, for example, suffixal /l/ maps to a nasal if one occupies the previous syllable (/ku**n**-i**l**-a/ → [ku**n**-i**n**-a] 'sow for', but /**n**ik-i**l**-a/ → [**n**ik-i**l**-a], *[**n**ik-i**n**-a] 'season for'; Rose & Walker 2004:479). To formalize this locality restriction, Bennett (2015b) proposes the constraint CC·SyllAdj (itself a modified version of Rose & Walker's 2004 Proximity), which penalizes correspondence between segments that do not belong to adjacent syllables (14).

(14) 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

Cases of local assimilation result when CC·SyllAdj dominates an otherwise active Corr·*X* con-

[3] More detailed support for the mismatch hypothesis is provided in Bennett's (2015b) Ch. 9. For previous critical discussion of its predictions regarding locality see McMullin & Hansson (2016b) and Stanton (2017); for critical discussion of its predictions regarding similarity see Stanton (2016).

straint. An example building on (13) above is in (15). When two place-distinct rhotics are in adjacent syllables, they are compelled to correspond and place-assimilate (15c). When the two rhotics are separated by a syllable, however, the ranking CC·SyllAdj ≫ Corr·[-lateral] renders correspondence impossible (15f). The best option given the ranking Ident-[±anterior] ≫ Ident-[±lateral] is (15e), where the two rhotics do not correspond and do not assimilate. (In (15-16), some non-crucially-ranked constraints are combined in a single column; violations of CC·Ident are subscripted with "CC" and violations of Ident are subscripted with "IO".)

(15) CC-SyllAdj compels local-only assimilation

| /ɹa-ɻata/ | CC·SyllAdj | ID-[±lat]<br>CC·ID-[±ant] | Corr·[-lat] | ID-[±ant] |
|---|---|---|---|---|
| a. [ɹ$_x$a-ɻ$_y$ata] | | | *! | |
| b. [ɹ$_x$a-ɻ$_x$ata] | | *!$_{\text{CC}}$ | | |
| ☞ c. [ɻ$_x$a-ɻ$_x$ata] | | | | * |
| d. [l$_x$a-ɻ$_y$ata] | | *!$_{\text{IO}}$ | | |
| /ɹa-taɻa/ | | | | |
| ☞ e. [ɹ$_x$a-taɻ$_y$a] | | | * | |
| f. [ɻ$_x$a-taɻ$_x$a] | *! | | | * |
| g. [l$_x$a-taɻ$_y$a] | | *!$_{\text{IO}}$ | | |

Given a different ranking, the set of constraints employed in (15) can generate non-local-only dissimilation. Illustrative tableaux for one such ranking are in (16).[4]

(16) CC-SyllAdj compels non-local-only dissimilation

| /ɹa-ɻata/ | CC·SyllAdj | Corr·[-lat] | ID·[±lat] | ID·[±ant]<br>CC·ID-[±ant] |
|---|---|---|---|---|
| a. [ɹ$_x$a-ɻ$_y$ata] | | *! | | |
| ☞ b. [ɹ$_x$a-ɻ$_x$ata] | | | | *$_{\text{CC}}$ |
| ☞ c. [ɻ$_x$a-ɻ$_x$ata] | | | | *$_{\text{IO}}$ |
| d. [l$_x$a-ɻ$_y$ata] | | | *! | |
| /ɹa-taɻa/ | | | | |
| e. [ɹ$_x$a-taɻ$_y$a] | | *! | | |
| f. [ɹ$_x$a-taɻ$_x$a] | *! | | | *$_{\text{CC}}$ |
| g. [ɻ$_x$a-taɻ$_x$a] | *! | | | *$_{\text{IO}}$ |
| ☞ h. [l$_x$a-taɻ$_y$a] | | | * | |

[4] I do not consider candidates here like [ɹ$_x$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 by undominated CC·Ident-[±sonorant], which requires corresponding consonants to have the same specification for [±sonorant].

Here, dissimilation occurs in the non-local context 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 CORR·[-lat] and CC·SYLLADJ vacuously by mapping /ɹ/ to [l]. In the local context, whether or not there is any observable interaction among the rhotics depends on the ranking of IDENT·[±anterior] and CC·IDENT-[±anterior]. Correspondence with place assimilation (16b) is derived given CC·IDENT-[±anterior] ≫ IDENT-[±anterior]; correspondence without place assimilation (16b) is derived given the reverse.

The important point is that non-local dissimilation is derivable within the SCTD, as it is a direct consequence of the analysis of local assimilation. This leads to a difference in the SCTD and GOCP's predictions regarding the character of non-local dissimilation. As discussed above, the GOCP predicts that all cases of non-local dissimilation should be linked to some external factor 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 dissimilation to co-exist with local assimilation, for example (as is the case for the system characterized by (16c,h)), this is not necessary as non-local dissimilation is also predicted to exist on its own (as is the case for the system characterized by (16b,h). This difference in prediction is due to a difference in the type of constraint interactions that generate non-local dissimilation. In the GOCP, non-local 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 is in some ways a less restrictive theory than is the GOCP, as it does not predict that non-local dissimilation is necessarily tied to constraints that disprefer local dissimilation. 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 candidates for a GOCP-based analysis. These would be cases of non-local dissimilation where there is no conceivable reason why local dissimilation should be dispreferred.

## 1.3 Roadmap

The rest of the paper argues that Sundanese non-local 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).[5] 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

[5] See also Hansson (2010) for discussion of an analysis that invokes Base-Reduplicant correspondence. Discussion in Zuraw (2002), Hansson (2010), and Bennett (2015a) suggests that the analysis proposed here resembles an analysis proposed by Suzuki (1999), but the paper containing that analysis is not available to me.

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). [ar] and [al] are exponents of a plural infix that appears before the first vowel in the stem.[6] It is a productive verbal infix 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.

Table 1: Data illustrating the Sundanese pattern

| *Input* | *Output* | *Schematics* | *Comments* |
|---|---|---|---|
| a. /ar-kusut/<br>‘messy (pl.)’ | [k-ar-usut] | [C-ar-VCVC] | **[ar]**: No other [r]s present. |
| b. /ar-gɨlis/<br>‘beautiful (pl.)’ | [g-ar-ɨ**l**is] | [C-ar-V**l**VC] | **[ar]**: No other [r]s present. |
| c. /ar-hormat/<br>‘respect (pl.)’ | [h-al-o**r**mat] | [C-al-V**r**CVC] | **[al]**: Multiple [r]s avoided. |
| d. /ar-combrek/<br>‘cold (pl.)’ | [c-al-omb**r**ek] | [C-al-VCCV**r**V] | **[al]**: Multiple [r]s avoided. |
| e. /ar-ŋumbara/<br>‘go abroad (pl.)’ | [ŋ-al-umba**r**a] | [C-al-VCCV**r**V] | **[al]**: Multiple [r]s avoided. |
| f. /ar-rahɨt/<br>‘wounded (pl.)’ | [**r**-ar-ahɨt] | [**r**-ar-VCVC] | **[ar]**: Preceding onset is /r/. |
| g. /ar-curiga/<br>‘suspicious (pl.)’ | [c-ar-u**r**iga] | [C-ar-V**r**VCV] | **[ar]**: Following onset is /r/. |
| h. /ar-lɨtik/<br>‘little (pl.)’ | [**l**-al-ɨtik] | [**l**-al-VCVC] | **[al]**: First consonant is /l/. |
| i. /ar-liren/<br>‘take a break (pl.)’ | [**l**-al-i**r**en] | [**l**-al-V**r**VC] | **[al]**: First consonant is /l/. |

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

I assume that the preferred allomorph is [ar], as it surfaces when the root contains no other [r] (a-b). When the root contains an [r], however, the preferred allomorph is generally [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 one of the syllables adjacent to the infix's /r/ has an /r/ onset, [ar] surfaces unexpectedly (f-g). The result is agreement among onsets of adjacent syllables for [-lateral]. Second, if the stem-initial onset is [l], [al] surfaces unexpectedly (h-i). The result is agreement between the stem's first two syllable onsets 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 (14) 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.[7]

Arguments for the SCTD-based analysis of Sundanese come in part from its natural ability to derive this complementarity, and in part from difficulties that the data pose to co-occurrence-based theories of dissimilation (such as the GOCP). Namely, it is difficult for theories invoking constraints like *$X\ldots Y$ to explain why [r] co-occurrence is permitted in adjacent syllables but not otherwise. 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 Sundanese assimilation and dissimilation is justified. First, as discussed above, CC·SYLLADJ incorrectly predicts the existence of non-local dissimilation in the absence of restrictions on local co-occurrence. 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, work by Hansson & McMullin (2014 *et seq.*) has 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 dissimilation with concomitant local assimilation. Crucially, participants were unable to learn non-local dissimilation not accompanied by local assimilation, the only pattern type exclusively predicted by the SCTD.

[7] This summary glosses over a number of details, including how the stem-initial restriction on [+lateral] assimilation is derived. See Bennett (2015a,b) for the full analysis.

## 2.1 Co-occurrence-based analysis

I analyze the preference for the [ar] allomorph, visible from forms like [k-ar-usut] and [g-ar-ɨlis] (a-b of Table 1), by assuming that the morpheme's underlying representation is /ar/ and mapping to [al] is penalized by 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), I assume that *[r]. . . [r] (5) dominates 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 interaction is illustrated in (17).

(17) 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] | *! | | * |

Forms like [r-ar-ahɨt] and [c-ar-uriga] show, however, that violations of *[r]. . . [r] are tolerated when the two [r]s belong to adjacent syllable onsets. To explain why this occurs, I propose that in Sundanese there is a more general drive for adjacent syllables to be "coupled" in a reduplication-like structure (as anticipated 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 productive 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 proposes that coupling is motivated by REDUP, a constraint that requires words to contain coupled substrings. While "there is a tendency in the data [. . . ] for coupled substrings to be adjacent syllables" (p. 405), Zuraw does not take a position on whether adjacency should be encoded into the definition of REDUP or captured by constraints that control the size and placement of reduplicants more generally. For expositional simplicity, I assume that the constraint promoting coupling in Sundanese is REDUP-$\sigma\sigma$ (18), which requires words to contain coupled adjacent syllables.

(18) REDUP-$\sigma\sigma$:
Assign one * if a word does not contain adjacent coupled syllables.

Coupled substrings are subjected to additional identity requirements. In Sundanese, I claim that in order for two adjacent substrings to be coupled, their onsets must be identical. This preference is loosely defined as κκ·Ident-[onset] (19).

(19) κκ·Ident-[onset]:
Assign one * for each pair of coupled syllables with non-identical onsets.

To derive [c-ar-uriga] and [r-ar-ahɨt], Redup-$\sigma\sigma$ and κκ·Ident-[onset] must dominate *[r]. . . [r] (20).[8] This captures the generalization that having adjacent coupled syllables with matching onsets is more important than avoiding [r] co-occurrence. (Note that in all tableaux that follow, I omit Ident·Root-[±lateral] and candidates that violate it. Ident·Root-[±lateral]'s ranking with respect to the rest of the constraints introduced here will be included and justified in the summary. In addition, I do not include candidates like [c-a]$_\kappa$[c-u]$_\kappa$riga, where the affixal /r/ maps to a consonant other than [r] or [l]. I assume that the absence of these forms is due to high-ranked faithfulness constraints for features like [±consonantal] and [±approximant].)

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

| /ar-rahɨt/ | κκ·Ident-[onset] | Redup-$\sigma\sigma$ | *[r]. . . [r] |
|---|---|---|---|
| a. r-ar-ahɨt | | *! | * |
| ☞ b. [r-a]$_\kappa$[r-a]$_\kappa$hɨt | | | * |
| c. r-al-ahɨt | | *! | |
| d. [r-a]$_\kappa$[l-a]$_\kappa$hɨt | *! | | |
| /ar-curiga/ | | | |
| e. c-ar-uriga | | *! | * |
| ☞ f. c-a[r-u]$_\kappa$[ri]$_\kappa$ga | | | * |
| g. c-al-uriga | | *! | |
| h. c-a[l-u]$_\kappa$[ri]$_\kappa$ga | *! | | |

Candidates (20a,c,e,g) do not contain adjacent coupled syllables and are eliminated by high-ranked Redup-$\sigma\sigma$. Candidates (20d,h) satisfy Redup-$\sigma\sigma$ but are eliminated by higher-ranked κκ·Ident-[onset], as the onsets of the coupled syllables are not identical. The optimal (20b,f) show that violation of *[r]. . . [r] is acceptable when it allows for satisfaction of higher-ranked κκ·Ident-[onset] and Redup-$\sigma\sigma$. Put differently, violations of *[r]. . . [r] are permitted only when the result is onset identity between adjacent syllables.[9]

[8] κκ·Ident-[onset] $\gg$ Redup-$\sigma\sigma$ is established transitively on the basis of further data; see (27).

[9] Forms like [c-ar-uriga] (Table 1h), [r-ar-ɨwat] 'startled (pl.)', [di-k-ar-irim] 'sent-pass (pl.)' (Cohn 1992:206) suggest that constraints requiring further 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 κκ·Ident-[nucleus] is outranked by input-output faithfulness constraints on vowel quality (Ident-[±high], etc.). di-k-a[r-i]$_\kappa$[rim]$_\kappa$ suggests that the constraint(s) requiring coupled syllables to be identical in their rime structure (see Zuraw 2002:415-417 for evidence of such a pressure in Tagalog) is subordinated to Max and Dep. For evidence that nucleus-matching is gradiently attested in the lexicon, see Section 3.

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]$_\kappa$ (21b) has adjacent [r]-containing onsets but still violates $\kappa\kappa$·IDENT-[onset] because the onsets are not identical. The analysis predicts that the winning candidate should be c-al-ombrek (21c), 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. A further candidate c-a[r-omb]$_\kappa$[rek]$_\kappa$ is presumably ruled out by *COMPLEX-CODA, inviolable in Sundanese.)

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

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

Similarly, the analysis derives the generalization that identical [r]-containing onsets are illicit if they are not adjacent. In the case of /ar-ŋumbara/, for example, ŋ-a[r-um]$_\kappa$ba[ra]$_\kappa$ (22b) satisfies $\kappa\kappa$·IDENT-[onset] but violates REDUP-$\sigma\sigma$; ŋ-a[l-um]$_\kappa$ba[ra]$_\kappa$ violates both. ŋ-al-umbara (22c) is selected as optimal because, unlike (22a), it satisfies *[r]. . . [r].[10]

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

| /ar-ŋumbara/ | $\kappa\kappa$·IDENT-[onset] | REDUP-$\sigma\sigma$ | *[r]. . . [r] |
|---|---|---|---|
| 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]$_\kappa$ | *! | * | |

The current analysis incorrectly predicts that /ar-gɨlis/ should surface as g-a[l-ɨ]$_\kappa$[lis]$_\kappa$. As $\kappa\kappa$·IDENT-[onset] and REDUP-$\sigma\sigma$ 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, *[l]. . . [l] (23), is active in Sundanese.

(23) *[l]. . . [l]:

Assign one * for each pair of [l]s within the word.

[10] A further candidate ŋ-a[r-um]$_\kappa$[ba]$_\kappa$[ra]$_\kappa$ satisfies REDUP-$\sigma\sigma$, given the definition in (18). It is not clear to me however if coupling among more than two syllables should be permitted by GEN; Zuraw (2002) implicitly assumes that only two strings can be coupled and I follow her here. Even if ŋ-a[r-um]$_\kappa$[ba]$_\kappa$[ra]$_\kappa$ were a possible candidate it would be eliminated by $\kappa\kappa$·IDENT-[onset]. The variant ŋ-a[r-um]$_\kappa$[ra]$_\kappa$[ra]$_\kappa$ could be ruled out by constraints demanding faithfulness to the featural content of root segments or just to [±sonorant].

To take effect, *[l]...[l] must dominate REDUP-$\sigma\sigma$. Under this ranking, g-al-ɨlis (26c) and g-a[l-ɨ]$_\kappa$[lis]$_\kappa$ (26d) are eliminated by high-ranked *[l]...[l]; g-ar-ɨlis is selected as optimal as it satisfies both top-ranked constraints, despite its violation of REDUP-$\sigma\sigma$.

(24) *[l]...[l] ≫ REDUP-$\sigma\sigma$ explains /ar-gɨlis/ → [g-ar-ɨlis]

| /ar-gɨlis/ | $\kappa\kappa$·IDENT-[onset] | *[l]...[l] | REDUP-$\sigma\sigma$ | ID-[±lateral] |
|---|---|---|---|---|
| ☞ a. g-ar-ɨlis | | | * | |
| b. g-a[r-ɨ]$_\kappa$[lis]$_\kappa$ | *! | | | |
| c. g-al-ɨlis | | *! | * | * |
| d. g-a[l-ɨ]$_\kappa$[lis]$_\kappa$ | | *! | | * |

This analysis accounts for all of the data in Table 1's (a-g). Left to explain is why /ar-lɨtik/ surfaces as [l-ar-ɨtik] and /ar-liren/ surface as [l-al-iren]; the current analysis predicts that in both cases the infix should surface as [ar], due to undominated *[l]...[l]. To account for these data I propose a constraint demanding coupling between the first two syllables in the word, REDUP-$\sigma_1\sigma_2$ (25).

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

To allow [l]s to co-occur when they are onsets to the first and second syllables, it must be the case that REDUP-$\sigma_1\sigma_2$ dominates *[l]...[l]. Tableau (26) shows that given this ranking, the winning candidate for the input /ar+liren/ is [l-a]$_\kappa$[l-i]$_\kappa$ren (26d); other possible candidates are eliminated by high-ranked constraints.

(26) 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-$\sigma\sigma$ |
|---|---|---|---|---|
| a. l-ar-iren | | *! | | * |
| b. l-a[r-i]$_\kappa$[ren]$_\kappa$ | | *! | | |
| c. l-al-iren | | *! | * | * |
| ☞ d. [l-a]$_\kappa$[l-i]$_\kappa$ren | | | * | |
| e. [l-a]$_\kappa$[r-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 (27) 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 (27c).

(27) $\kappa\kappa$·Ident-[onset] ≫ Redup-$\sigma_1\sigma_2$ is necessary for /ar-curiga/ → [c-ar-uriga]

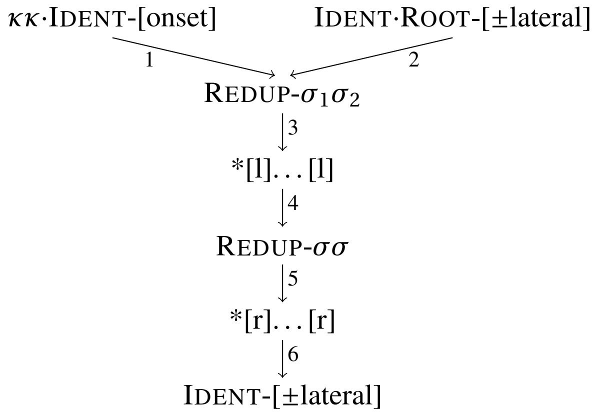
| /ar-curiga/ | Redup-$\sigma_1\sigma_2$ | $\kappa\kappa$·Ident-[onset] | Redup-$\sigma\sigma$ | *[r]…[r] |
|---|---|---|---|---|
| a. c-ar-uriga | *! | | * | * |
| b. [c-a]$_\kappa$[r-u]$_\kappa$riga | | * | | *! |
| 💣 c. [c-a]$_\kappa$[l-u]$_\kappa$riga | | * | | |
| ☹ d. c-a[r-u]$_\kappa$[ri]$_\kappa$ga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

To allow coupling to occur outside of the stem-initial context, then, it is necessary for $\kappa\kappa$·Ident-[onset] to dominate Redup-$\sigma_1\sigma_2$. The analysis can now account for all data in Table 1.

## 2.2 Local summary

The proposed analysis of the Sundanese data is summarized in (28), 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.

(28) Summary of Sundanese analysis



| | |
|---|---|
| [1] $\kappa\kappa$·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-$\sigma\sigma$: | g-ar-ɨlis ≻ *g-a[l-ɨ]$_\kappa$[lis]$_\kappa$ |
| [5] Redup-$\sigma\sigma$ ≫ *[r]…[r]: | c-a[r-u]$_\kappa$[ri]$_\kappa$ga ≻ *c-al-uriga |
| [6] *[r]…[r] ≫ Ident-[±lateral]: | h-al-ormat ≻ *h-ar-ormat |

As mentioned throughout, a number of other constraints regulate aggressive reduplication in these infixed forms. Some of these constraints dominate Redup-$\sigma_1\sigma_2$: constraints demanding faithfulness for features other than [±lateral], for example, are necessary to explain why /ar-kusut/ does not surface as $[\text{k-a}]_\kappa[\text{k-u}]_\kappa$sut. In some cases the position of these constraints in the hierarchy is unclear: MAX, DEP, and COMPLEXCODA must dominate REDUP-$\sigma\sigma$ to explain why /ar-combrek/ is not c-a$[\text{r-om}]_\kappa[\text{rek}]_\kappa$ or c-a$[\text{br-om}]_\kappa[\text{brek}]_\kappa$ or c-a$[\text{r-omb}]_\kappa[\text{rek}]_\kappa$, 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.[11] The ranking in (28) predicts that it should surface instead as [ar-ulur], as *[l]...[l] violations 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] (29c,d) over attested [al-ulur] (29b).

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

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

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] 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 (29b,d) would satisfy REDUP-$\sigma_1\sigma_2$ and a$[\text{l-u}]_\kappa[\text{lur}]_\kappa$ (29d) would be correctly chosen as the winner.[12] It is difficult to know at present which of these solutions is more plausible.

## 3 Evidence from the lexicon

As discussed above, co-occurrence-based theories of dissimilation predict that non-local dissimilation must coexist with a distinct pressure that disprefers the result of local dissimilation. Under the

[11] Bennett (2015b) transcribes this as [(ʔ)-al-ulur], but does not hear the [ʔ], and notes that its inclusion is for consistency with past descriptions. The distribution of [ʔ] is predictable (Robins 1959) and from this I infer that is not part of roots' underlying representations; whether or not and where it surfaces predictably is not important here.

[12] 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 (25), 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.

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 for 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 approximately 960 [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 rhotic dissimilation, liquid dissimilation, and aggressive reduplication. This evidence and its relationship to the analysis is previewed below.

- *Evidence for aggressive reduplication in adjacent syllables:*
  If aggressive reduplication is active in the Sundanese lexicon, then in contexts where coupling is possible, similarity along one dimension should encourage similarity along another. I investigate this prediction by considering 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 disproprtionately likely to have matching nuclei. For non-adjacent syllables, no such correlation exists. These findings are consistent with Section 2's claim that REDUP-$\sigma\sigma$ 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 $\sigma_1$ and $\sigma_2$ 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).

- *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 $\sigma_1\sigma_2$, discussed above).

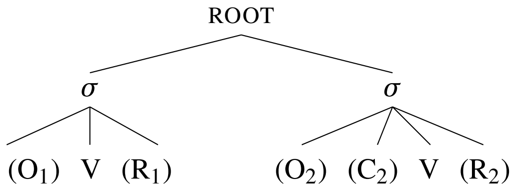
The main point of this section is that trends in the Sundanese lexicon are consistent with each of the markedness constraints proposed in Section 2. These findings are thus consistent with the claim that non-local dissimilation in Sundanese can be analyzed as the interaction between unbounded dissimilation 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 (28) 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 other languages.[13] 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 (30).

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



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)

Following this description meant that a word like *ablag* was syllabified as *a.blag*, a word like *ambacak* was syllabified as *am.ba.cak*, and a word like *aŋgrit* was syllabified as *aŋ.grit*. A small

[13] Lembaga Basa & Sastra Sunda (1985) is a monolingual Sundanese dictionary. Since I cannot read Sundanese, no glosses are provided in what follows.

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 (*tasblaŋ*, for example, was syllabified as *tas.blaŋ*). Unsyllabified and syllabified versions of the wordlist are available on the author's website.

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*). In particular, a large number of the longer forms appear to be fully reduplicated roots (*alangahéléngeh*, *alunalun*, *balataboloto*, *borakborak*; see Van Syoc 1959:78-80 on morphological reduplication in Sundanese). As part of our interest here is in the extent of evidence for the activity of aggressive reduplication, including morphologically reduplicated forms would bias the results.[14]

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.[15] Thus words like *ke.ke.ba*, where all syllables have singleton onsets, are considered for all contexts ($\sigma_1\sigma_2$, $\sigma_2\sigma_3$, and $\sigma_1\sigma_3$). Words like *af.ri.ka*, 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*, 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* 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 (31).

(31) 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*, *af.ri.ka* |
| $\sigma_1\sigma_3$ | 2,933 | *ke.ke.ba*, *ba.i.kot* |

To determine the frequency of onset pairs relative to expectation, loglinear models were fit to each of the datasets in (31). 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

[14] Sundanese also has several types of partial reduplication. These are discussed in Sections 3.2 as they become relevant.

[15] The limitation to native singleton onsets was made largely to simplify the statistics and the data visualizations, but also in part 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, like *kaʔbah*, are likely not due to this predictable pattern. See Cohn 1992:205 for the Sundanese inventory and Robins 1959 for discussion of [ʔ].)

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 (32) 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.

(32) Co-occurrence count encoding for regression analysis

| Combination | Count | Identical | $p_1$ | $t_1$ | $l_1$ | $k_1$ | $p_2$ | $t_2$ | $l_2$ | $k_2$ | $l_{12}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| pp | 2 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| pt | 7 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| pl | 5 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| pk | 4 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| tp | 30 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| tt | 15 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| tl | 44 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| tk | 26 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| lp | 15 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| lt | 13 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 |
| ll | 6 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 |
| lk | 14 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 |
| kp | 3 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 |
| kt | 2 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 |
| kl | 3 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 |
| kk | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |

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 2007) and the quasipoisson link function.[16]

For each context, further evidence for aggressive reduplication was evaluated by determining if nucleus-matching was significantly correlated with onset-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.

## 3.2 Results

The results of the lexicon study are presented by-context below: $\sigma_1\sigma_2$ is in Section 3.2.1, $\sigma_2\sigma_3$ is in Section 3.2.2, and $\sigma_1\sigma_3$ is in Section 3.2.3.

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 provide some external support for the constraints proposed in Section 2. Materials that provide a more complete description of Sundanese lexical statistics are available on the author's web site.

### 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 1, which plots the baseline model's predicted count for a given onset pair against its observed count.[17] Identical pairs are represented with black dots and all other pairs are represented with gray.[18] 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=[ɟ], and y=[j]. The interpretation of all other characters is straightforward.)

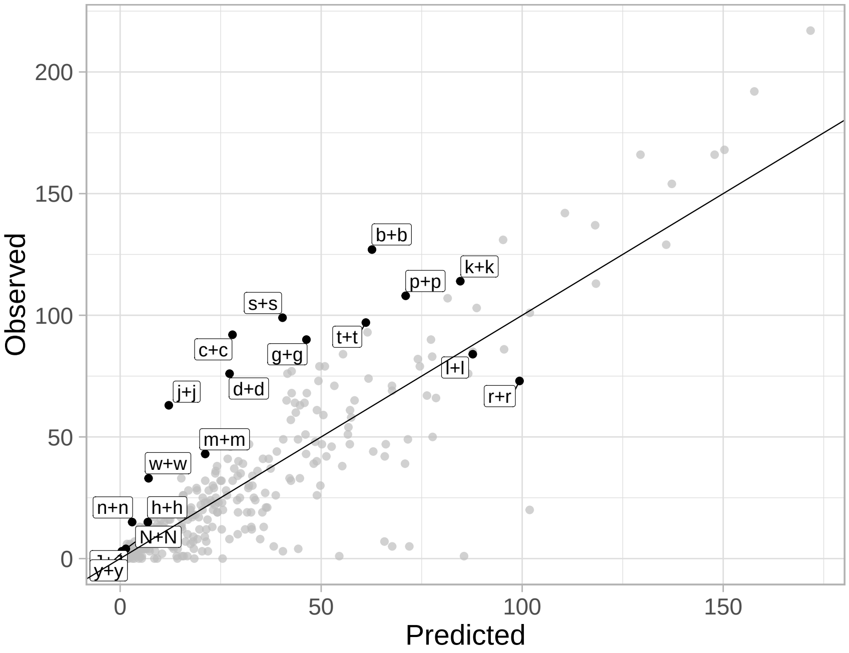
It is clear from Figure 1 that identical $\sigma_1\sigma_2$ onset pairs are overattested relative to expectation: identity is linked to a boost in frequency that cannot be explained only by reference to the independent frequency of the pair's members. In addition, l+l and r+r are underattested relative to other identical pairs. The results of the second loglinear model, which incorporate predictors referencing identity, confirm that these observations are unlikely to be due to chance. The positive coefficient in (33b) confirms that identity is linked to a significant increase in log frequency, and the negative coefficients in (33c-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 (33e-h)) and the general frequency boost

[16] 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 family is appropriate for these data because in all relevant subsets, the variance in frequency is larger than the mean.

[17] All plots in this section were made with R's ggplot2 and gghighlight packages; Wickham 2016, Yutani 2018.

[18] For each figure, interactive plots that label each dot are available on the author's website.

Figure 1: Predicted vs. observed frequencies of $\sigma_1\sigma_2$ onset pairs



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

(33) 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 (34).

(34) Onset-matching encourages nucleus-matching in $\sigma_1\sigma_2$ ($\chi^2$ (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 (see Robins 1959:361-362 for further examples). 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 vast majority (97%) of words in the dictionary are two syllables or longer, suggesting a dispreference for monosyllabic words. 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 (e.g. *boŋ.boŋ* was excluded because *boŋ* independently exists); these exclusions were made to control for the possibility of reduplicated monosyllabic roots, and brought the number of forms considered down from 6,409 to 6,373.[19] Figure 2 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 1.

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 (35).

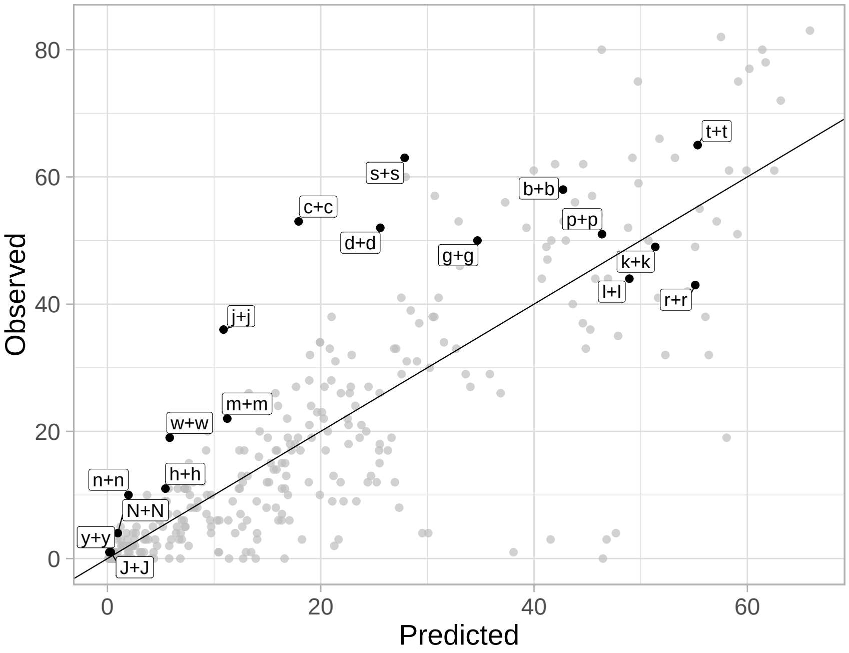
(35) Onset-matching encourages nucleus-matching in $\sigma_1\sigma_2$ ($\chi^2$ (1) = 409.33, $p$ < .001)

| | 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. Furthermore, it is unlikely that the observed preference for self-similarity between $\sigma_1$ and $\sigma_2$ can be attributed to morphological reduplication: the preference is also observed within a

[19] Implicit evidence for a minimal word restriction comes from Robins (1959), who provides no monosyllabic examples, and Cohn (1992), who states that "most roots are disyllabic" (p. 206). Van Syoc (1959), however, provides sporadic examples of monosyllabic content words (e.g. *pok* 'say', p. 55; *prak* 'about to eat', p. 56; *toŋ* 'barrel', p. 149), suggesting that this restriction is not absolute.

Figure 2: Predicted vs. observed frequencies of $\sigma_1\sigma_2$ onset pairs, from disyllabic subset



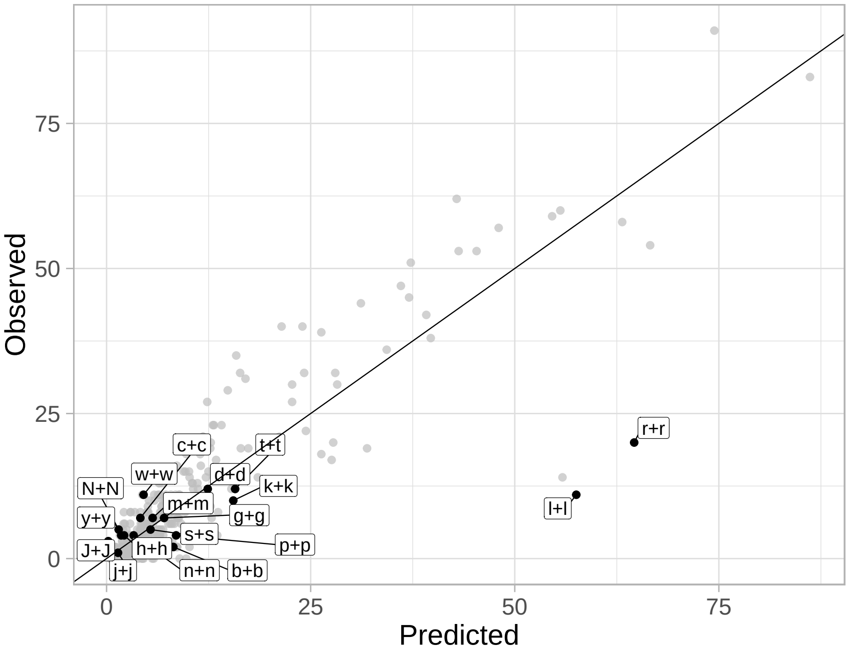
set of forms that are likely not morphologically reduplicated.

### 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 3 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 (36)) confirm the observations made on the basis of Figure 3. The predictor for onset identity is not significant: whether or not a pair of onsets are 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.

Figure 3: Predicted vs. observed frequencies of $\sigma_2\sigma_3$ onset pairs



(36) 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 (37), where 66.7% of syllable pairs with matching onsets but only 49.6% of syllables with mismatching onsets have matching nuclei.

(37) Onset-matching encourages nucleus-matching in $\sigma_2\sigma_3$ ($\chi^2$ (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 that similarity along one dimension encourages 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 $\sigma_2$'s onset is frequently occupied by [l] or [r] ((33f,h); (36e,g)). One might ask if this is due to morphology, and, in particular, to the dictionary's potential inclusion of plural forms (like *k**al**usut*, *h**al**ormat*). 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. *g**ar**alaŋ*, *b**al**ida*). Possible singulars were identified by removing the *al* or *ar* from the potential plural (so *galaŋ*, *bida*) and searching the wordlist again for the result.

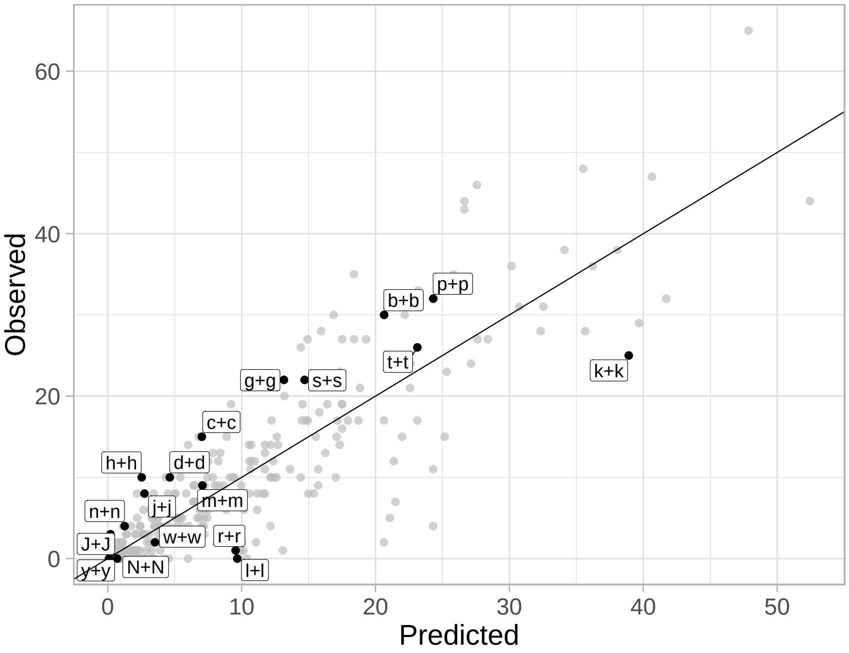
The majority of forms (722/935, or 77%) that qualify as a potential plural do not have a corresponding potential singular in the wordlist. Of the 214 potential plurals that do, 81 do not obey the generalizations regarding the distribution of [al~ar] (these are forms like *c**al**acah*, where *ar* is expected, or *l**ar**uŋ*, where *al* is expected), leaving 133 phonologically plausible plurals with a potential singular pair. Examples are *dapon-d**ar**apon*, *jujur-j**al**ujur*; pairs like *o-**ar**o* 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 pluralized 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 4 plots the observed count for each $\sigma_1\sigma_3$ onset 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).

Figure 4: Predicted vs. observed frequencies of $\sigma_1\sigma_3$ onset pairs



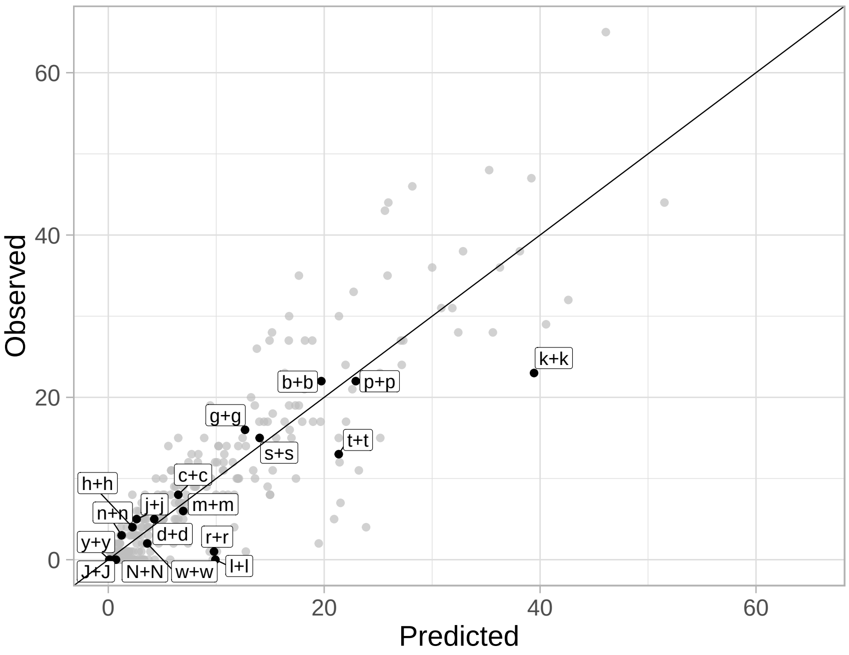
To determine if these trends are meaningful, identity-based predictors were added to the baseline model. The results are consistent with Figure 4: there is a boost in log frequency for identity (38b) and a decrease in log frequency for r+r and l+l (38c-d) relative to other identical pairs and the independent frequencies of [r]s and [l]s (38e-h).

(38) 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*, *corodcod*, *harashas*, *perekpek*). 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).[20] As is 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 5 confirms that they do in fact look quite different.

Figure 5: Predicted vs. observed frequencies of $\sigma_1\sigma_3$ 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] in this context (*rudira*). Yet in Figure 5, the apparent preference for identity has vanished. These trends are confirmed by a second loglinear model fit to the data visualized in Figure 5, which 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.[21] (Full results for this model are included in the appendix.)

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

[20] 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.

[21] 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.

the 74 potentially reduplicated forms are included, the rate of onset-matching is not significantly correlated with the rate of nucleus-matching (39). 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 ($\chi^2$ (1) = 1.00, $p$ > .1).

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

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

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-$\sigma_1\sigma_2$ that corresponding substrings are strictly adjacent: in non-adjacent syllables, similarity along one dimension does not encourage 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-$\sigma\sigma$, REDUP-$\sigma_1\sigma_2$, and $\kappa\kappa$·IDENT constraints). While corroborating evidence from other synchronic processes is limited, I have shown here that each of the constraints proposed in Section 2 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*, *budrabidru*) or partial/aggressive reduplication (*karari*, *rereb*). Many of the remaining 47 are likely loans, though they are not necessarily annotated as such (*kolaborator*, *barometer*, *organisator*). For [l]: 226 forms contain more than one [l], and 204 of these cases can be interpreted as involving total reduplication (*ulangaling*, *lapatlapat*) or partial/aggressive reduplication (*lalab*, *tulalet*). Again, of the remaining 22, many are loans (*alkohol*, *kolonial*). 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.

Another general finding is that in adjacent but not non-adjacent syllables, onset identity encourages nucleus identity (consistent with the activity of REDUP-$\sigma\sigma$ and the IDENT-$\kappa\kappa$ constraints associated with it). In addition, onsets are more likely to match in $\sigma_1\sigma_2$ than is naïvely expected (providing support for 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.

It is worth discussing briefly what the larger analysis of these effects might be, given Section 2's claim that faithfulness to root material supersedes all of the markedness constraints whose effects are investigated here. I follow Martin (2007) in assuming that the grammar's effects on the lexicon are indirect: constraints like *[l]. . . [l] and $\kappa\kappa$·IDENT-[onset] play a role in determining which words are more likely to be coined and accepted by speakers, but do not necessarily modify those words directly. The relative rarity of words containing multiple [r]s and [l]s, then, is due not to any active dissimilatory process but rather to speakers' relative unwillingness to accept and use words that violate these constraints. The willingness of speakers to bring novel words into conformance with the language's phonotactics (exemplifed, for example, by the variation among *rapor*, *lapor*, and *rapot*; Cohn 1992:231 and discussion above) could potentially be explained along these same lines.

### 3.3 Lexical evidence and learnability

The above shows that the constraints proposed in my analysis of the /ar/ allomorphy 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 et al. 2011), so it is not necessarily the case that there will 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 learners would only rarely be exposed to evidence that [ar~al] alternations exist, and in order to account for them 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. This is 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 to appear) 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.[22]

### 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 *l**al**oba* and ***ar**eliŋ* were considered but forms like *region**al*** were not. Potential singulars were identified by removing *ar* or *al* from the plural and searching the wordlist for the resulting singular. Thus for *l**al**oba*, the wordlist was searched for *loba*; for ***ar**eliŋ*, the wordlist was searched for *eling*. A singular-plural pair was recorded if the corresponding singular exists and has a higher token frequency than the plural. This frequency criterion was 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.[23] Several examples with their associated frequencies are in (40).

(40) 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. | *leres* | *laleres* | 'correct (pl.)' | 129, 0 |

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.

[22] 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.

[23] 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 are *budak-barudak* 338/346, *maneh-maraneh* 1083/1425, *manehna-maranehna* 619/749, *manehanana-maranehana* 0/24). The high frequency of these plural forms is unsurprising as the learner would need to be exposed to them in order to learn that they take that particular affix. As far as I am able to tell, none of the forms exhibiting the [ar∼al] alternations are nouns.

### 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-h**al**al* and *tatu-tat**alu*** 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 (*tat**alu***). (I included pairs like *tatu-tat**alu*** because some prefixes can attach outside of /ar/; an example from Robins 1959:344 is *ka-duga*, *ka-d**ar**uga*. Semantics were not considered because glosses are not provided.)

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.[24] 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).

## 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 isomorphy 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-

[24] 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-s**al**alabar* appear to exhibit [l]-assimilation in an unexpected context, but Sundanese has a prefix *sa-* (Robins 1959:352) and so it is possible that [l]-assimilation has applied as expected in stem-initial position. Most of the apparent exceptions have a plausible reanalysis along these lines.

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 dissimilation should only arise given the coexistence of some independent pressure that disprefers the results of local dissimilation. As discussed above, Sundanese – the only known case of non-local dissimilation – fits this description. In addition to cases like Sundanese, the SCTD predicts cases of non-local 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 (2016a) 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). Hansson & McMullin (2014) however show that participants are not able learn to 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 non-local contexts. These findings suggest that the types of non-local dissimilation uniquely predicted by the SCTD are not only unattested but also unlearnable, and that the correct theory of dissimilation should not treat them as part of the learner's hypothesis space. In this way, the GOCP is a more accurately restrictive theory than is the SCTD.

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 (2016) 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 2016:539). 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 (33), the $\sigma_2\sigma_3$ model summarized in (36), the $\sigma_1\sigma_3$ model summarized in (38), and the additional $\sigma_1\sigma_3$ 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 2: Results for $\sigma_1\sigma_2$ (all forms included)

| Predictor | Coefficient | *t* value | Significant? | Predictor | Coefficient | *t* value | Significant? |
|---|---|---|---|---|---|---|---|
| Intercept | 0.14 | – | | $\mathrm{m}_1$ | 0.12 | 0.47 | |
| **Identical** | **0.43** | **9.43** | ******* | $\mathrm{m}_2$ | -0.06 | -0.27 | |
| **$\mathbf{l}_{\mathbf{12}}$** | **-0.48** | **-3.02** | ****** | $\mathrm{n}_1$ | -0.69 | -2.58 | * |
| **$\mathbf{r}_{\mathbf{12}}$** | **-0.64** | **-3.85** | ******* | $\mathrm{n}_2$ | -0.20 | -0.83 | |
| $\mathrm{p}_1$ | 0.45 | 1.84 | . | $\mathrm{ny}_1$ | -1.09 | -3.62 | *** |
| $\mathrm{p}_2$ | 0.16 | 0.68 | | $\mathrm{ny}_2$ | -0.77 | -2.90 | ** |
| $\mathrm{t}_1$ | 0.30 | 1.22 | | $\mathrm{ng}_1$ | -0.85 | -3.05 | ** |
| $\mathrm{t}_2$ | 0.25 | 1.05 | | $\mathrm{ng}_2$ | -0.41 | -1.65 | |
| $\mathrm{c}_1$ | 0.16 | 0.65 | | $\mathrm{s}_1$ | 0.40 | 1.61 | |
| $\mathrm{c}_2$ | 0.02 | 0.10 | | $\mathrm{s}_2$ | -0.04 | -0.18 | |
| $\mathrm{k}_1$ | 0.49 | 2.01 | * | **$\mathbf{l}_{\mathbf{1}}$** | **0.28** | **1.13** | |
| $\mathrm{k}_2$ | 0.20 | 0.85 | | **$\mathbf{l}_{\mathbf{2}}$** | **0.56** | **2.37** | * |
| $\mathrm{b}_1$ | 0.43 | 1.75 | . | **$\mathbf{r}_{\mathbf{1}}$** | **0.29** | **1.17** | |
| $\mathrm{b}_2$ | 0.13 | 0.53 | | **$\mathbf{r}_{\mathbf{2}}$** | **0.64** | **2.74** | ****** |
| $\mathrm{d}_1$ | -0.15 | -0.59 | | $\mathrm{w}_1$ | -0.29 | -1.16 | |
| $\mathrm{d}_2$ | 0.31 | 1.33 | | $\mathrm{w}_2$ | -0.18 | -0.73 | |
| $\mathrm{j}_1$ | -0.04 | -0.16 | | $\mathrm{y}_1$ | -1.91 | -4.06 | *** |
| $\mathrm{j}_2$ | -0.17 | -0.71 | | $\mathrm{y}_2$ | -0.31 | -1.26 | |
| $\mathrm{g}_1$ | 0.28 | 1.12 | | $\mathrm{h}_1$ | -0.17 | -0.67 | |
| $\mathrm{g}_2$ | 0.14 | 0.61 | | $\mathrm{h}_2$ | -0.31 | -1.28 | |

Table 3: Results for $\sigma_2\sigma_3$ (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 4: Results for $\sigma_1\sigma_3$ (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 | |
| $\mathbf{l_{13}}$ | **-1.51** | **-1.95** | **.** | $n_1$ | -0.65 | -2.35 | * |
| $\mathbf{r_{13}}$ | **-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 | ** | $\mathbf{l_1}$ | **0.03** | **0.12** | |
| $k_3$ | 0.25 | 1.07 | | $\mathbf{l_3}$ | **0.33** | **1.43** | |
| $b_1$ | 0.58 | 2.35 | * | $\mathbf{r_1}$ | **0.04** | **0.16** | |
| $b_3$ | 0.06 | 0.27 | | $\mathbf{r_3}$ | **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 5: Results for $\sigma_1\sigma_3$ (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*, vol. 27, 17–32. GLSA.

Anderson, Stephen R. 1993. Wackernagel's Revenge: Clitics, Morphology, and the Syntax of Second Position. *Language* 69. 68–98.

Becker, Michael, Nihan Ketrez & Andrew Nevins. 2011. The Surfeit of the Stimulus: Analytic Biases Filter Lexical Statistics in Turkish Laryngeal Alternatinos. *Language* 87. 84–125.

Becker, Michael, Andrew Nevins & Jonathan Levine. 2012. Asymmetries in generalizing alternations to and from initial syllables. *Language* 88. 231–268.

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

Bennett, William G. 2015a. Assimilation, dissimilation, and surface correspondence in Sundanese. *Natural Language and Linguistic Theory* 33. 371–415.

Bennett, Wm. G. 2015b. *The Phonology of Consonants: Harmony, Dissimilation, and Correspondence*. Cambridge: Cambridge University Press.

Cohn, Abigail C. 1992. The Consequences of Dissimilation in Sundanese. *Phonology* 9. 199–220.

Downing, Laura. 1998. On the prosodic misalignment of onsetless syllables. *Natural Language and Linguistic Theory* 16. 1–52.

Eringa, F. S. 1949. *Loetoeng Kasaroeng. Verhandelingen van het Koninklijk Instituut voor Taal-, Land- en Volkenkunde 8.*

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. 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.

Gallagher, Gillian & Jessica Coon. 2009. Distinguishing total and partial identity: Evidence from Chol. *Natural Language and Linguistic Theory* 27. 545–582.

Gelman, A. & J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gouskova, Maria & Gillian Gallagher. to appear. Inducing Nonlocal Constraints from Baseline Phonotactics. *Natural Language and Linguistic Theory* .

Hansson, Gunnar Ólafur. 2010. *Consonant Harmony: Long-Distance Interaction in Phonology*. Berkeley, CA: University of California Press.

Hansson, Gunnar Ólafur & Kevin McMullin. 2014. Biased learning of long-distance assimilation and dissimilation. Poster presented at the 2014 Annual Meeting of the Linguistics Society of Great Britain.

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, University of Massachusetts Graduate Student Association.

Kroeger, Paul R. 1989. Discontinuous reduplication in vernacular Malay. In *Proceedings of the Berkeley Linguistic Society*, 193–202.

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

MacEachern, Margaret. 1997. *Laryngeal Cooccurrence Restrictions*. Los Angeles: University of California dissertation.

Martin, Andrew Thomas. 2007. *The Evolving Lexicon*. Los Angeles: University of California dissertation.

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

McMullin, Kevin & Gunnar Ólafur Hansson. 2016b. Long-Distance Phonotactics as Tier-Based Strictly 2-Local Languages. In Adam Albright & Michelle A. Fullwood (eds.), *Proceedings of the 2014 Meeting on Phonology*, Washington, DC: Linguistic Society of America.

Myers, Scott. 1997. OCP Effects in Optimality Theory. *Natural Language and Linguistic Theory* 15. 847–892.

Robins, R. H. 1959. Nominal and verbal derivation in Sundanese. *Lingua* 8. 337–369.

Rose, Sharon & Rachel Walker. 2004. A Typology of Consonant Agreement as Correspondence. *Language* 80. 475–531.

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

Stanton, Juliet. 2016. Wm. G. Bennett (2015). The phonology of consonants: harmony, dissimilation, and correspondence. (Cambridge Studies in Linguistics 147.) Cambridge: Cambridge University Press. Pp. xix+ 394. *Phonology* 33. 533–544.

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.

Suzuki, Keiichiro. 1998. *A typological investigation of dissimilation*. Tucson, AZ: The University of Arizona dissertation.

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, H. 2016. *ggplot2: Elegant Graphics for Data Analysis*. New York: Spring-Verlag.

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'. Manual available online at http://CRAN.R-project.org/package=gghighlight.

Zuraw, Kie. 2002. Aggressive reduplication. *Phonology* 19. 395–439.