An IBSP Description of Sanskrit /n/-Retroflexion

Ayla Karakaş
Department of Linguistics
Stony Brook University
Stony Brook, NY 11794, USA
ayla.karakas@stonybrook.edu

Abstract

Graf and Mayer (2018) analyze the process of Sanskrit /n/-retroflexion (*nati*) from a subregular perspective. They show that *nati*, which might be the most complex phenomenon in segmental phonology, belongs to the class of *input-output tier-based strictly local languages* (IO-TSL). However, the generative capacity and linguistic relevance of IO-TSL is still largely unclear compared to other recent classes like the *interval-based strictly piecewise* languages (IBSP: Graf, 2017, 2018). This paper shows that IBSP has a much harder time capturing *nati* than IO-TSL, due to two major shortcomings: namely, the requirement of an upper bound on relevant segments, and a lack of descriptive succinctness.

1 Introduction

Research in computational phonology has determined that all phonological patterns fit in the class of finite-state languages (Kaplan and Kay, 1994). The study of subregular phonology explores tighter characterizations of phonological phenomena in the form of subclasses of the regular languages. This furnishes lower and upper complexity bounds for phonological computations, which in turn provides new insights for typology and learnability — see Heinz 2018 and references therein.

One phenomenon that has proven to be particularly complex is /n/-retroflexion in Sanskrit, also known as *nati*. The nasal /n/ undergoes retroflexion whenever it appears immediately before a sonorant and a retroflex exists somewhere to its left. While this interaction of local and nonlocal factors is already unusual, the true complexity of the process comes from various blocking effects. It has been known since Graf (2010) that *nati* — when viewed as a phonotactic constraint on surface forms — is star-free. Recently, an alternative upper bound has been established in the form of *input-output tier-based strictly local languages* (IO-TSL; Graf and Mayer, 2018).

IO-TSL is an extension of the empirically well-supported class TSL (Heinz et al., 2011). Whereas subclasses of IO-TSL enjoy independent empirical support (De Santo and Graf, 2019; Mayer and Major, 2018), the only empirical motivation for IO-TSL itself is *nati*. The formal properties of IO-TSL are also not well-understood. It is not even known whether IO-TSL is a subclass of the star-free languages. By contrast, the class of *interval-based strictly piecewise* languages (IBSP; Graf, 2017, 2018) is properly star-free, handles a wide range of phonotactic phenomena, and has even been applied to syntax (Shafiei and Graf, 2019). For all these reasons, an IBSP analysis of *nati* would be a valuable addition to the current IO-TSL description, and might furthermore shed light on how these two classes differ.

In this paper, I argue that *nati* belongs to the intersection closure of IBSP, but the resulting grammar is much more convoluted than the IO-TSL analysis. While the basic cases of *nati* are very natural from an IBSP perspective, the interactions of blocking effects are hard to capture due to two limitations of IBSP's notion of *open slots*: the inability to force a segment to always appear in an open slot, and the inability to mark an open slot as optional. These insights might prove useful for a future proof separating IBSP and IO-TSL.

The structure of the paper is as follows: IBSP is formally defined in Sec. 2, adapting the more general format proposed in Graf (2018). Sec. 3 then walks the reader through the *nati* analysis, starting from the simplest case and refining the IBSP grammar with each new complication. Sec. 4 reflects on the status of the analysis and what lim-

itations of IBSP make *nati* so difficult to account for.

2 Preliminaries

Graf (2017) first defined the class of *interval-based strictly piecewise* (IBSP) string languages as an extension of the *strictly piecewise* (SP) languages (Rogers et al., 2010). IBSP enriches SP with locality domains, and the checking of SP-dependencies is limited to these locality domains. IBSP properly subsumes SP, but also the classes SL and TSL, all three of which play a major role in subregular phonology. Graf (2018) further generalizes the format of locality domains to account for phenomena that had previously been analyzed in terms of I-TSL. Only this more general version can handle *nati*.

Intuitively, an IBSP interval involves definitions of I) the left and right *domain edge*, II) a finite number k of *open slots*, and III) the *fillers* that can occur between open slots. Fillers and domain edges are defined through k-*intervals*, also called k-vals. The IBSP grammar also supplies a list of forbidden k-grams. A string is well-formed iff there is no way to instantiate the k-val in such a manner that the configuration of open slots matches a forbidden k-gram.

While IBSP is originally defined in terms of first-order logic (Graf, 2017), I adopt the newer definition of Shafiei and Graf (2019) as it also subsumes the generalized intervals of Graf (2018). Note that $\cdot$ in definition 2.2 denotes string concatenation lifted to sets, i.e. $S \cdot T := \{st \mid s \in S, t \in T\}$.

Definition 2.1 (k-val)**.** A *segmented* k*-interval* ($k \geq 0$) over alphabet Σ, or simply segmented k-val, is a tuple $\langle L, R, F_i \rangle_{0 \leq i \leq k}$ such that:

- $L, R \subseteq \Sigma \cup \{\varepsilon\}$ specify the left edge and right edge, respectively, and
- $F_i \subseteq \Sigma$ specifies the i-th filler slot.

Definition 2.2 (IBSP-k)**.** Let Σ be some fixed alphabet and $\rtimes, \ltimes \notin \Sigma$ two distinguished symbols. An IBSP-k grammar over $\Sigma \cup \{\rtimes, \ltimes\}$ is a pair $G := \langle i, S \rangle$, where i is a segmented k-val over $\Sigma \cup \{\rtimes, \ltimes\}$ and $S \subseteq (\Sigma \cup \{\rtimes, \ltimes\})^k$ is a set of forbidden k-grams. A string $s \in \Sigma^*$ is generated by G iff there is no k-gram $u_1 ... u_k \in S$ such that $\rtimes^k s \ltimes^k$ is a member of the language

$$(\Sigma \cup \{\rtimes, \ltimes\})^* \cdot L \cdot F_0^* \cdot \{u_1\} \cdot F_1^* \cdot \{u_2\} \cdot \\ \ldots \cdot F_{k-1}^* \cdot \{u_k\} \cdot F_k^* \cdot R \cdot (\Sigma \cup \{\rtimes, \ltimes\})^*$$

The language $L(G)$ is the set of all $s \in \Sigma^*$ that are generated by G. A stringset L is IBSP-k iff $L = L(G)$ for some IBSP-k grammar G.

The reader may skip ahead to (1) and (2) for a depiction of a concrete IBSP interval and its application to an illicit string.

In IBSP, all possible instantiations of a locality domain must be evaluated. If at least one of them yields a match for an illicit k-gram, the whole string is discarded. By default, fillers allow each open slot to be arbitrarily far away from the next one. However, adjacency of the i-th and $i+1$-th open slot can be enforced by stipulating $F_{i+1} = \emptyset$. Here, F_{i+1} refers to the subset of Σ that is allowed in the filler between the i-th and $i+1$-th slots. The subset is empty if nothing is allowed in that filler. This is not to be confused with the string language corresponding to the $i+1$-th filler, which is $F_{i+1}^* = \{\varepsilon\}$. Mixing such empty fillers with normal fillers allows IBSP to capture phonotactic constraints in which local and non-local dependencies interact. As we will see next, this is not needed for the simplified version of *nati*, but will be crucial once the full range of facts is considered (Sec. 3.3 and subsequent sections).

3 Data and Analysis

Nati is a left-to-right long-distance assimilation process with a single trigger, a single target, and several conditions for blocking. While *nati* is usually described as a process — i.e. a mapping from underlying forms to surface forms — I treat it as a phonotactic phenomenon. That is to say, *nati* is reanalyzed as a constraint on the distribution of [n] in surface forms, making it a matter of string languages rather than string transductions. This is in line with the previous work done by Graf and Mayer (2018), which will henceforth be referred to as G&M.

The discussion starts with the simplest cases of *nati* and continually refines the IBSP description as new data is considered. The final version is presented in Sec. 3.5.

Several notational conventions will be adopted for the remainder of this paper: Sanskrit examples have their triggers and targets bolded, while

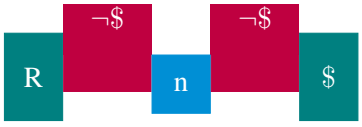
active blockers are underlined. All the examples are taken from G&M and Ryan (2017). Since the phonotactic perspective forgoes any notion of underlying forms, I will only use square brackets to denote surface segments throughout this paper. IBSP interval diagrams are represented in a pictorial fashion: domain edges are large, green rectangles, fillers are vertically offset boxes in red, and open slots are blue squares.

3.1 Long-distance assimilation

Nati starts out with the basic constraint that a nasal target /n/ becomes [ɳ] when preceded arbitrarily far to the left by a non-lateral retroflex continuant in {/ɻ/, /ɻ̝/, /ɻ̞/, /ʂ/}. G&M formalize this as the constraint "no [n] may appear in the context $R \cdots \_$", where R is one of the triggers listed in the preceding sentence.

G&M's constraint is easily expressed in terms of IBSP. Our grammar consists of a single forbidden unigram, which is n. By keeping word edges ($) and string edges ({⋊, ⋉}) distinct, IBSP enables us to instantiate intervals across multiple words in a string, if desired. I will use $ instead of ⋉ for now as this does not commit us as to whether the string consists of a single phonological word or a sequence of words. But as discussed in Sec. 4, it may eventually be necessary to use the string edge ⋉ instead. For now, the use of the word edge $, along with banning the appearance of $ in fillers, captures that *nati* cannot apply across word boundaries.

(1) **IBSP interval (Version 1)**



For the sake of succinctness, the interval above lists the forbidden unigram directly in the open slot. While this is non-standard, I believe it makes the analysis easier to follow once the complexity of the intervals starts to increase.

Tab. 1 lists some data points that are relevant for this base case. The form of the instrumental singular suffix /-eːna/ alternates based on whether the root it attaches to contains a trigger for *nati*. For the sake of exposition, I also include an illicit nonce variation, indicated by the gloss "N/A".

Form	Gloss	*Nati*?	Licit?
káːm-eːna	'by desire'	✗	✓
manuʂj-eːɳa	'by human'	✓	✓
manuʂj-eːna	N/A	✓	✗

Table 1: Forms showing basic *nati* (Ryan, 2017, p. 305)

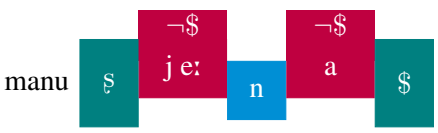
The reader may wonder why an analogous nonce form *káːm-eːɳa* is not included in Tab. 1. In this nonce form, /n/ would undergo *nati* without a suitable trigger, which should be illicit. However, this presupposes a view of *nati* as a process. From the perspective of phonotactics, it is not obvious that this nonce form is actually illicit because [ɳ] can occur independently of *nati*. The phonotactics of *nati* only concern the distribution of [n], not [ɳ], so only the former need to be considered here.

Let us now see how the locality domain in (1) captures the well-formedness of the first two forms in Tab. 1 while also ruling out the illicit nonce form. First, *káːm-eːna* is well-formed because it lacks a retroflex, so there is no suitable left edge for the interval in (1). Hence the locality domain cannot be established at all, so there are no open slot configurations to check against the list of forbidden unigrams. As a result, the string is well-formed.

The second example is *manuʂj-eːɳa*, which does allow for numerous instantiations of the interval. In all instantiations, the interval spans from [ʂ] to the right word edge, and the only difference is what segments make up the fillers and which one ends up in the open slot. Since *manuʂj-eːɳa* does not contain any [n], the open slot never matches the forbidden unigram. Consequently, this string is also deemed well-formed. In contrast to the first example, where well-formedness followed from the inability to instantiate any locality domain, this example allows for many distinct instantiations but none of them yield a forbidden configuration of open slots.

This leaves us with the illicit *manuʂj-eːna*. It works exactly like the second case, except that now there is an instantiation that results in a match with the forbidden unigram n. This particular instantiation is depicted below.

(2) **IBSP interval: manuʂj-eːna**



So far, IBSP has not done anything that could not be accomplished by simpler means, e.g. an SP grammar. As we start adding on conditions and exceptions, though, IBSP intervals will quickly become indispensable.

3.2 Unconditional blocking by intervening coronals

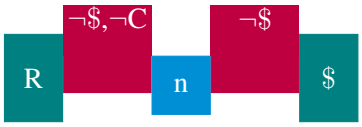
We now turn to the first of the *nati*-blocking effects: /n/-retroflexion can be blocked if a coronal segment appears between trigger and target. The set of relevant coronals includes retroflexes but excludes the glide [j] as the latter is both a sonorant and a coronal — see Ryan (2017) for further discussion. Tab. 2 lists a particular example of coronal blocking, an illicit nonce form, and a nonce form that illustrates what the surface form would look like if coronals were not blockers.

Form	Gloss	*Nati*?	Blocking?	Licit?
vaɻɳ-anaːnam	no gloss	✗	✓	✓
vaɻm-anaːnam	N/A	✗	✗	✗
vaɻɳ-aɳaːnam	N/A	✓	N/A	✓

Table 2: Forms showing blocking by intervening coronals (Hansson, 2001, p. 227)

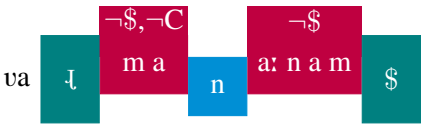
In G&M, the forbidden context for [n] is updated to $R\overline{C}\cdots\_$, where $\overline{C}$ matches every segment that is not a coronal, including [j]. To represent this in IBSP, we modify the first filler in (1) so that it may not contain any coronals either. If a string contains a coronal, it must go in the open slot or the second filler. Either way, no subsequent [n] can appear in the open slot, and consequently the string will be deemed well-formed.

(3) **IBSP interval (Version 2)**



At the same time, strings without coronals will still be judged illicit. This is illustrated below for the nonce form *vaɻm-anaːnam*.

(4) **IBSP interval: vaɻm-anaːnam**



Note that [ɳ] itself is a coronal blocker, so any subsequent [n] in a word loses its eligibility as a target for *nati*. The only exception to this is geminate [nn] sequences where both [n] become retroflexed. However, this could also be treated as a separate process of progressive local assimilation. I put this issue aside for now, but it will be revisited in Sec. 4.

3.3 Mandatory adjacency to sonorant

In order for [n] to undergo *nati*, it must also be immediately followed by a vowel, a glide, [m], or [n] itself. More succinctly, the following segment must be a non-liquid sonorant (Whitney, 1889). For example, in the form *bɻahman*, *nati* does not apply as [n] occurs at the very end of the word without any subsequent sonorant. Similarly, *nati* does not apply in *caɻ-a-n-ti*, in this case because [t] is not a sonorant. Sanskrit has some nasals besides [m] and [n] that are non-liquid sonorants, but since those cannot follow [n] for independent reasons (Emeneau, 1946) they do not matter for the purposes of this paper.

Form	Gloss	*Nati*?	Sonorant?	Licit?
caɻ-a-**n**-ti	'wander (3Pl)'	✗	✗	✓
bɻahma**n**	'brahman'	✗	✗	✓
bɻahma**n**a	N/A	✗	✓	✗

Table 3: Forms showing mandatory adjacency to sonorant; (Hansson, 2001, p.229) and (Ryan, 2017, p. 318)

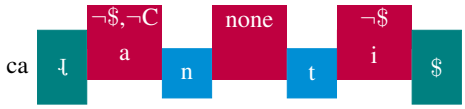
G&M represent the new illicit context for [n] as $R\overline{C}\cdots\_S$, where S is a suitable sonorant. We will use the same definition of S to add a second open slot to the interval in (3). The list of illicit unigrams is now expanded to illicit bigrams. It is no longer just [n] that is forbidden, but rather any bigram of the form nS. Keep in mind that coronal blocking is still active, though.

(5) **IBSP interval (Version 3)**

The descriptor *none* in the second filler of (5) indicates that $F_1 \subset \Sigma$ is $\emptyset$ (and thus $F_1^* = \{\varepsilon\}$). That is to say, this filler cannot contain any symbols at all and the first and second open slot must always be adjacent.

Let us verify that the first two examples in Tab. 3 are still well-formed given the grammar in (5). Below is an example of one possible interval established in *caɻ-a-n-ti*.

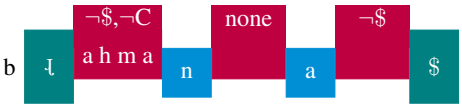
(6) **IBSP interval: caɻ-a-n-ṯi**



This is the only interval that could possibly cause the IBSP grammar to reject the string, since the first open slot is filled by n. However, as the second open slot is not a sonorant, the open slot configuration does not match any of the forbidden bigrams. The well-formedness of *bɻahman* follows for the very same reason: there is no way of instantiating the locality domain so that the two open slots would contain [n] and a sonorant, respectively.

At the same time, *bɻahmana* is correctly ruled out as illicit.

(7) **IBSP interval: bɻahmana**



3.4 Conditional blocking by preceding velar and labial plosives

Coronal consonants are not the only blockers of *nati*: velar and labial plosives also block its application, but only if I) the plosive immediately precedes the target nasal, and II) a left root boundary ($\sqrt{}$) occurs somewhere between the trigger and the plosive. Based on the data given in G&M and Ryan (2017), I assume that for a given word, an interval instantiated within the word never has to contend with more than one $\sqrt{}$ — this will be elaborated on in Sec. 4. Blocking is contingent on both conditions being met, as is exemplified by the data in Tab. 4. In *pɻa-√miːŋ-aː-ti*, *nati* still occurs across a left root boundary due to the absence of a plosive immediately before [n]. In *√ɻug-ŋá*, *nati* can target an n after an immediately preceding velar plosive [g] because the left root boundary does not occur between the triggering retroflex and the plosive. Only in *(abʱi-)pɻa-√gʱn-an-ti* does *nati* fail as there is both a plosive and a root boundary, both of which occur in the relevant positions.

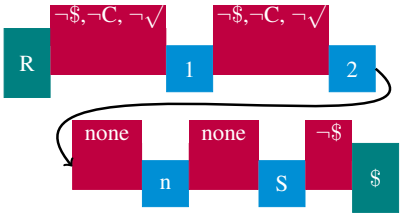
Form	**Gloss**	***Nati*?**	**Licit?**
pɻa-√miːŋ-aː-ti	'vanishes (3s)'	✓	✓
√ɻug-ŋá	'break (pass. part.)'	✓	✓
(abʱi-)pɻa-√gʱn-an-ti	'broken'	✗	✓

Table 4: Forms showing conditional blocking by preceding plosives (Ryan, 2017, p. 319, 321)

In response to this additional complication, G&M update the banned context to $R\alpha\cdots\_$. Here α is any string that neither contains a coronal nor matches $\cdots\sqrt{}\cdots P$, with P denoting a velar or labial plosive. It is at this point that the complexity of our IBSP treatment ramps up significantly. We must now introduce open slots whose only purpose is to be sensitive to the conditional presence of certain segments. By setting up the fillers in such a way that root boundaries and immediately preceding plosives can only go into open slots, we can ensure that the grammar is always aware of these segments if they occur in the string. The list of forbidden k-grams is then set up in such a fashion that open slot configurations that start with a root boundary and a plosive are exempt from *nati*. This is a very unusual use of open slots and fillers, and I am unaware of any other IBSP-analysis that has to resort to this trick.

The concrete steps are as follows. First, two additional open slots must be included between the trigger and target. Open slot 1 detects the presence of a left root boundary somewhere arbitrarily to the left of [n]. Open slot 2 detects the presence of a velar/labial plosive immediately before an [n]. For readability, graphical depictions of longer intervals will now be broken up across two lines.

(8) **IBSP interval (Version 4)**



The filler before the third open slot is set to *none* so that it can only be filled by whatever segment immediately precedes [n]. The fillers surrounding the first open slot are more complex. The ban against coronals is carried over from coronal blocking, but in addition these fillers may not contain a root boundary either. As a result, a root boundary that occurs somewhere between the triggering retroflex and a suitable plosive is forced into the first open slot. The conjunction of all these factors ensures that if a string contains a suitable root boundary and plosive, they will always occur in the first two open slots.

In the next step, we expand the list of forbidden bigrams of the form nS to forbidden 4-grams of the form ϕnS. Here ϕ represents a large number of bigrams. As *nati* is only blocked whenever the

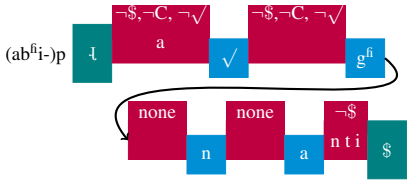
first open slot is a root boundary and the second open slot is a plosive, nS is illicit if:

1. the first open slot is not a root boundary, or
2. the second open slot is not a plosive, or
3. both 1 and 2 hold.

Hence ϕ corresponds to any combination of segments that matches one of the three conditions above.

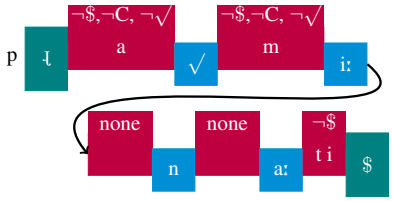
If the first two open slots in an instantiated interval do not match ϕ, *nati* will not be enforced, capturing the described blocking effect. This is illustrated below for *(ab$^{\text{ɦ}}$i-)pɻa-$\sqrt{\text{g}^{\text{ɦ}}}$n-an-ti*.

(9) **IBSP interval: (ab$^{\text{ɦ}}$i-)pɻa-$\sqrt{\text{g}^{\text{ɦ}}}$n-an-ti**



Any configuration where the first two open slots are not $\sqrt{}$ and a plosive will match ϕ, triggering a *nati* violation if the remaining two open slots are filled by [n] and a sonorant. As a concrete example, consider the nonce form *pɻa-√miːn-aː-ti*.

(10) **IBSP interval: pɻa-√miːn-aː-ti**



The reader is urged to verify for themselves that the remaining forms in Tab. 4 are handled correctly by this grammar.

An additional bug arises in that the introduction of new open slots has created an "escape hatch" for coronals. In previous versions, a coronal had to go into the first or second open slot, or the third filler. These are now the third and fourth open slot and the fifth filler. While coronals are still banned in the first and second filler, they could go into the first or second open slot. Since ϕ currently matches coronals, too, we no longer capture coronal blocking. Fortunately, the fix is easy. We further restrict the shape of ϕ so that it does not match any open slot configuration with a coronal. Overall, this leaves the following patterns for ϕ:

1	2
$\sqrt{}$	$\neg$P $\wedge$ $\neg$C
$\neg\sqrt{} \wedge \neg$C	P
$\neg\sqrt{} \wedge \neg$C	$\neg$P $\wedge$ $\neg$C

Figure 1: Open slots in ϕ s.t. nS is illicit

Given a list of suitable list of segments for Sanskrit, ϕ can be compiled out into a list of bigrams. These bigrams are then prefixed with every possible instantiation of nS to arrive the list of forbidden 4-grams.

3.5 Conditional blocking by following retroflex

Even though the grammar in (8) is already fairly complicated, it still does not handle the last layer of *nati*: if a retroflex appears arbitrarily far to the right of the target [n], /n/-retroflexion may be blocked. Blocking only occurs when both of the following two conditions are met: I) a left root boundary intervenes between the trigger and the target, and II) there is no coronal between the target [n] and blocking retroflex. Condition II) is particularly peculiar. Essentially, the appearance of a coronal consonant between [n] and its following retroflex blocks the blocking of *nati* by said retroflex, so that *nati* applies as usual.

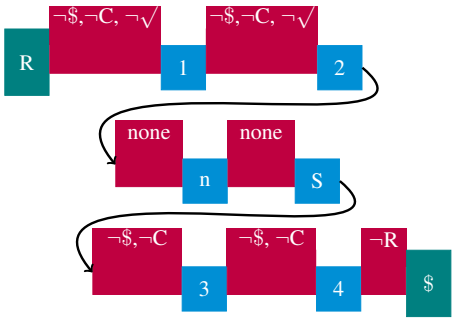
Form	Gloss	*Nati*?	Licit?
pɻa-√**n**aṣ-ʈum	'to vanish (inf.)'	✗	✓
pɻa-√ɳeː-ʈɻ	'leader'	✓	✓
pɻ-ɳa-k-ṣi	'unite (2s)'	✓	✓

Table 5: Forms showing conditional blocking by following retroflex (Ryan, 2017, p. 325)

The form *pɻa-√naṣ-ʈum* in Tab. 5 shows the following retroflex acting as a blocker when a left root boundary intervenes between [ɻ] and [n]. On the other hand, the retroflex is not a blocker in *pɻa-√ɳeː-ʈɻ*, due to the coronal intervening between [n] and [ɻ]. Finally, *pɻ-ɳa-k-ṣi* is a case where the retroflex does not block in the absence of an intervening root boundary.

We can follow the same approach as in Sec. 3.4 to handle this complication. That is to say, we include yet another two conditional slots following the target nasal, and its mandatory adjacent sonorant. As the interval now gets exceedingly long, graphical depictions have to be broken up again across multiple lines.

(11) **IBSP interval (Version 5, Final)**



This time, open slot 3 tracks the presence of a coronal, and open slot 4 indicates whether a retroflex is present. Once again we have to forbid these segments in the neighboring fillers to ensure that if such a segment is present, it must go into one of these open slots.

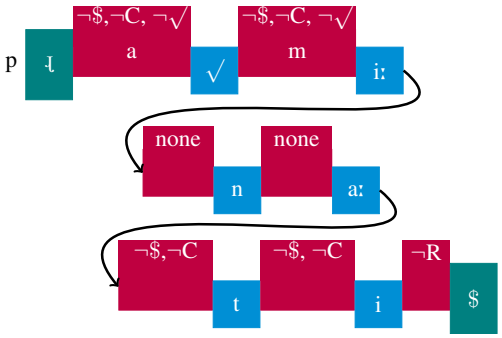
We then expand the list for forbidden 4-grams to forbidden 6-grams. The 4-gram pattern ϕnS is expanded to ϕn$S\phi'$. Just like ϕ describes the illicit segments for 1 and 2, ϕ' handles open slots 3 and 4 in (11). However, ϕ' cannot be described independently of ϕ as the relevance of slots 3 and 4 for blocking depends on the presence of a root boundary in open slot 1. Hence the options for ϕ and ϕ' have to be specified in conjunction in order to represent the conditions needed for *nati* to apply (i.e. cases where it fails to be blocked):

1	2	3	4
$\surd$	$\neg$P $\wedge$ $\neg$C	$\neg$C	$\neg$R
$\neg\surd \wedge \neg$C	$\neg$P $\wedge$ $\neg$C	$\neg$C	$\neg$R
$\neg\surd \wedge \neg$C	P	$\neg$C	$\neg$R
$\surd$	$\neg$P $\wedge$ $\neg$C	C	$\neg$R
$\neg\surd \wedge \neg$C	$\neg$P $\wedge$ $\neg$C	C	$\neg$R
$\neg\surd \wedge \neg$C	P	C	$\neg$R
$\surd$	$\neg$P $\wedge$ $\neg$C	C	R
$\neg\surd \wedge \neg$C	$\neg$P $\wedge$ $\neg$C	C	R
$\neg\surd \wedge \neg$C	P	C	R

Figure 2: Open slots in $\phi \wedge \phi'$ s.t. nS is illicit

The interval in (11), with the list of forbidden 6-grams above in Figure 2, is the final version of the IBSP grammar for *nati* (although other potential variants are discussed in Sec. 4). This is a good point to reevaluate some of the earlier data points. For example, we can model some examples that illustrate conditional blocking of intervening velar/labial plosives like so:

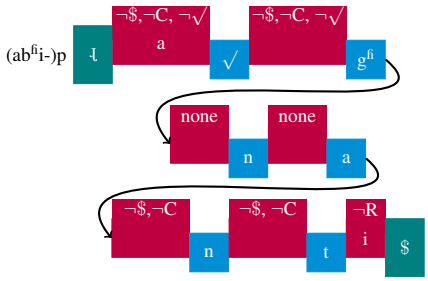
(12) **IBSP interval: pɻa√miːnaːti**



The instantiated locality domain looks quite similar to its previous iteration in (10). The main difference is that rather than having [t] and [i] in the filler following the nS sequence, those segments are pushed into the open slots that check for the presence of an anti-blocking coronal and/or blocking retroflex. The configuration of conditional slots matches $\surd$, $\neg P \wedge \neg C$, $\neg C$, $\neg R$, which is one that enforces *nati*. Consequently, the presence of an [n] in the open slot where it is forbidden causes the string to be rejected. If [n] had undergone *nati* as required, the string would not have been deemed illicit by the grammar.

The string *pɻa-√g$^{\text{ɦ}}$n-an-ti*, on the other hand, is still well-formed. Even when [n] appears in the open slot, this does not yield an illicit configuration of open slots due to the presence of a root boundary in open slot 1 and a plosive in open slot 2.

(13) **IBSP interval: (ab$^{\text{ɦ}}$i-)pɻa-√g$^{\text{ɦ}}$n-an-ti**



4 Discussion and conceptual remarks

The IBSP analysis developed over the course of Sec. 3 is with a doubt convoluted, much more so than the analysis in terms of IO-TSL. In contrast to IO-TSL, it also hinges on several idealizations that cannot be eliminated without further complicating the grammar. I will briefly sketch the most important issues here, in particular those that highlight the shortcomings of IBSP relative to IO-TSL.

At a high level of abstraction, the strategy employed in this paper boils down to a few simple tricks:

1. Furnish an open slot for every type of segment that can potentially matter for the dependency.

2. If an open slot needs to track the presence of some segment of type X, do not allow the surrounding fillers to contain X.

3. Whatever implicational relations hold between the relevant segments are compiled out into a list of forbidden k-grams.

While each step is conceptually simple, the sheer number of open slots and potential combinations of segments make proving that this approximation of *nati* is IBSP a daunting task. In addition, the first two strategies have serious drawbacks as they respectively impose a lower bound on the number of segments in the string, and an upper bound on how many segments of a specific type may occur in a specific part of the interval.

Let us consider the problem of a lower bound first. As more and more factors were incorporated into the analysis, more and more open slots had to be added to make the interval sensitive to the presence of any segments that might affect well-formedness. However, as the number of open slots grows, shorter strings are automatically considered well-formed. This is because IBSP trivially allows any string in which the interval cannot be instantiated. An interval with 6 open slots, for example, cannot be instantiated in a string that only consists of 5 symbols. In IBSP, a high number of interacting factors makes it difficult to regulate short strings.

As a remedy, Graf (2017) allows strings to be padded out by additional edge markers to enforce the required minimal length of strings. We could take a similar approach, and modify the right interval boundary to be the string edge rather than the word edge. As long as each string only represents a single phonological word rather than a string of words, the string edge is a viable replacement for the word edge. It is still far from obvious, though, that padding out can solve the problem of words where only one segment occurs between the retroflex trigger and the targeted [n]. Recall that the current interval posits two open slots, and hence at least two segments between them. While there might be some way to add even more open slots so that [n] can be "shifted" to the left and also occur in one of the first two open slots, this would render the account entirely opaque to human intuition.

In the other direction, IBSP also runs into an undesirable upper bound limit. For instance, coronals cannot go into the first or second filler, leaving only the first open slots as an option for a coronal that is somewhere to the left of [n] but not adjacent to it. If a string contains two coronals, neither one of which is adjacent to [n], the interval cannot be instantiated at all. In this case, this is unproblematic since coronals would block *nati* anyways, so either way the string is deemed well-formed. The situation is reversed, however, with coronals after [n], which undo blocking of *nati* by a retroflex. If a string contains two coronals between [n] and such a retroflex, the interval will not be instantiated and the string will incorrectly be treated as well-formed. Similarly, if more than one retroflex occurs between the sonorant following target [n] and the right interval boundary, the interval cannot evaluate the string. Again, one could fix these issues by adding more open slots and modifying the list of forbidden k-grams, but this would exacerbate the lower bound problem with short strings. It once again would make the grammar unintelligible.

Whether *nati* is actually IBSP thus cannot be answered definitively — it depends on how one generalizes from the finite data to an infinite sample. For the available data, it is certainly possible to construct the interval and the list of k-grams in a suitable manner, although it may be very difficult to verify the correctness of the analysis by hand. Once one generalizes from the data to allow an arbitrary number of coronals and retroflexes, IBSP may prove insufficient.

The latter point also holds for the intersection closure of IBSP. Suppose that each case of *nati* is given its own IBSP grammar, and that these grammars are arranged in such a fashion that the intervals for simpler cases cannot be established in the more complex cases. For instance, the interval in (5) could be amended so that the first filler may not contain a left root boundary and the last filler may not contain any retroflex. The interval then cannot be instantiated in any strings where these complicating factors are present, limiting it only to simple cases of *nati*. This solves the lower

bound problem, because shorter strings are now regulated by one of the IBSP grammars for simpler cases of *nati*. At full generality, however, the upper bound problem remains. For instance, sensitivity to retroflexes requires that retroflexes may not be fillers, and thus the interval's ability to accommodate retroflexes depends on its number of open slots. As there can be only a finite number of open slots, the number of retroflexes is finitely bounded. Intersection closure can increase that bound to any desired k, but it will always be bounded. Consequently, the intersection closure of IBSP can handle the attested *nati* data, but not necessarily the most natural generalization of this data.

There are also several minor issues of data analysis, such as the status of geminates. As mentioned in Sec. 3.2, geminate [n] becomes geminate [ɳ] under *nati*. This is not captured by the current grammar, but corresponding modifications could be made. If geminate [nː] is modeled as underlying /nn/, the list of forbidden 6-grams can be modified to also block [ɳn]. Then, [ɳɳ] would be the only possible surface form. On the other hand, if [nː] is a single symbol, then the 6-grams must be modified such that [nː] is forbidden even if the following segment is not a sonorant, since the geminate acts as its own sonorant (metaphorically speaking). These are minor issues compared to the much more substantive problem of how conditional sensitivity to a segment may sometimes entail an upper bound on the number of those segments in IBSP.

For all these reasons, IBSP does not provide an insightful or elegant perspective of *nati*, in particular compared to G&M's IO-TSL treatment. Nonetheless, the IBSP of view of *nati* has identified several issues that are relevant for subregular research, most prominently the specific shortcomings of IBSP in comparison to IO-TSL. These have not been noticed before because most phonological phenomena only require sensitivity to two or three segments. We now face the question of how one should treat analyses that diverge depending on how one generalizes from the finite data sample. The intersection closure of IBSP can handle all generalizations of *nati* as long as there is an upper bound on the number of relevant segments (retroflexes, coronals, left root boundaries), whereas IO-TSL requires no such upper bounds. Which one of the two is a more appropriate characterization? It may be the case that the bounds we find in the available data are not an artifact of a finite data sample, but indicators of a principled bound to the limits of IBSP (see Joshi (2000) for a similar argument in syntax).

Finally, there is the issue of succinctness and elegance and to what extent they should be a criterion in the classification of empirical phenomena. This is a long-standing debate: if X is computationally simpler than Y, but only Y provides for a natural description, which one of the two is a better model of the relevant linguistic factors? Of course, formal language theory is well-served by having both X and Y as descriptions of the phenomenon, but if we regard subregular complexity as an abstract gauge of the cognitive machinery (cf. Rogers and Pullum, 2011), X and Y may embody vastly different claims.

5 Conclusion

I have argued that a phonotactic pattern as complex as *nati*, which can be viewed as an interaction between local and non-local dependencies with intervening material that provides blocking effects, can be modeled with the intersection closure of IBSP. However, the details depend on specific assumptions about the data, and the proposed account is fairly complicated and lacks linguistic naturalness. These drawbacks highlight specific limitations of IBSP relative to IO-TSL, and might be useful for future work on the relation between the two.

Future work could revisit my findings along two dimensions. On a formal level, it might be possible to extend IBSP grammars with mechanisms that allow for more succinct descriptions without increasing generative capacity. From a linguistic perspective, one might try to reassess the empirical status of *nati* with respect to which of its components are most natural under an IBSP-analysis. If these aspects turn out to be on empirically solid ground, this might provide indirect evidence for IBSP as a model of natural language phonotactics.

Acknowledgements

The work reported in this paper was supported by the National Science Foundation under Grant No. BCS-1845344. I would like to express my deepest gratitude to Thomas Graf, for his endless insights and patient guidance. I am thankful to my colleagues, Aniello De Santo and Alëna

Aksënova, who provided valuable feedback and are a constant source of inspiration. My appreciation is further extended to the anonymous reviewers whose comments helped improve this paper. Any errors in this work are entirely my own.

References

Aniello De Santo and Thomas Graf. 2019. Structure sensitive tier projection: Applications and formal properties. In *Formal Grammar*, pages 35–50, Berlin, Heidelberg. Springer.

M.B. Emeneau. 1946. The nasal phonemes of Sanskrit. *Language*, page 22(2):86–93.

Thomas Graf. 2010. Comparing incomparable frameworks: A model theoretic approach to phonology. *University of Pennsylvania Working Papers in Linguistics*, 16(2):Article 10.

Thomas Graf. 2017. The power of locality domains in phonology. *Phonology*, 34:385–405.

Thomas Graf. 2018. Locality domains and phonological c-command over strings. In *Proceedings of "NELS" 2017*.

Thomas Graf and Connor Mayer. 2018. Sanskrit n-retroflexion is input-output tier-based strictly local. In *Proceedings of the Fifteenth Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 151–160, Brussels, Belgium. Association for Computational Linguistics.

Gunnar Ólafur Hansson. 2001. *Theoretical and Typological Issues in Consonant Harmony*. Ph.D. thesis, University of California, Berkeley.

Jeffrey Heinz. 2018. The computational nature of phonological generalizations. *Phonological Typology, Phonetics and Phonology*, pages 126–195.

Jeffrey Heinz, Chetan Rawal, and Herbert G. Tanner. 2011. Tier-based strictly local constraints in phonology. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, pages 58–64.

Aravind K. Joshi. 2000. Relationship between strong and weak generative power of formal systems. In *Proceedings of the Fifth International Workshop on Tree Adjoining Grammar and Related Frameworks (TAG+5)*, pages 107–114, Université Paris 7.

Ronald M. Kaplan and Martin Kay. 1994. Regular models of phonological rule systems. *Computational Linguistics*, 20(3):331–378.

Connor Mayer and Travis Major. 2018. A challenge for tier-based strict locality from Uyghur backness harmony. volume 10950 of *Lecture Notes in Computer Science*, pages 62–83. Formal Grammar, Springer, Berlin, Heidelberg.

James Rogers, Jeffrey Heinz, Gil Bailey, Matt Edlefsen, Molly Vischer, David Wellcome, and Sean Wibel. 2010. On languages piecewise testable in the strict sense. 6149:255–265.

James Rogers and Geoffrey K. Pullum. 2011. Aural pattern recognition experiments and the subregular hierarchy. *Journal of Logic, Language and Information*, 20:329–342.

Kevin Ryan. 2017. Attenuated spreading in Sanskrit retroflex harmony. *Linguistic Inquiry*, 48 (2):299–340.

Nazila Shafiei and Thomas Graf. 2019. The subregular complexity of syntactic islands. Ms., Stony Brook University.

William Dwight Whitney. 1889. *Sanskrit Grammar*. Oxford University Press, London.