Non-Local Allomorphy in a Strictly Local System
(draft version)

Jurij Božič
McGill University
jurij.bozic@mail.mcgill.ca

May 23, 2017

Abstract

The goal of this paper is two-fold. It presents a cross-linguistic survey of non-local allomorphy and it develops a formal model that accounts for the observed patterns. The survey reveals that local and non-local allomorphy occupy two opposing sides of a markedness scale: local allomorphy is unmarked, while non-local allomorphy is marked and exceptional. Also, the distance between the trigger and target of allomorphy in non-local patterns is much more conservative than expected. We develop a model of allomorphy, where the markedness split in locality stems from an ECONOMY condition of the PF-interface.

1 The Problem of Non-Local Allomorphy

The nature of contextual allomorphy and its locality has recently been the subject of much attention in Distributed Morphology (Halle & Marantz 1993). There is consensus that syntactic phases represent locality boundaries for allomorphy (Embick 2010), but it is less clear how allomorphy is constrained within a phase. This is an important question, since a single phase may contain a great deal of complex morphology. The standard observation is that contextual allmorphy is triggered by purely local, i.e. adjacent heads (Embick 2010; Arregi & Nevins 2012), which was noted as a significant generalization in earlier work (Siegel 1978; Allen 1979; Simpson & Withgott 1986; Carstairs-McCarthy 1992; Stump 1996). Recent inquiry has, however, identified several patterns where the trigger is not adjacent and is hence non-local, for which different approaches have been proposed (Merchant 2015; Moskal 2015a; Moskal & Smith 2016).

Existing treatments of non-locally triggered allomorphy (Merchant 2015; Moskal 2015a; Moskal & Smith 2016) effectively treat *non-locality* as a general property of grammars, i.e. the only distinction between a local and a non-local pattern is the way the allomorphic rule is *encoded lexically*. This paper, on the other hand, argues for a formal distinction between *local* and *non-local* allomorphy. What is needed is a system that maintains strict locality, but allows it to be sometimes violated in a principled way. We propose that allomorphy is *by default* triggered by strictly local, adjacent heads, and that this represents the UNMARKED pattern. Non-local allomorphic patterns on the other hand, emerge as instances of systematic exceptions, representing the more MARKED pattern. We present several arguments for this, where the most glaring one involves the observation that *adjacency effects* can assert themselves in systems that exhibit non-local patterns. We devise a formally explicit model of Vocabulary Insertion, where this split in markedness stems from an economy condition of the PF-interface. This minimizes the amount of theoretical machinery needed to derive allomorphy and aims to maximize explanation of allomorphy in natural language.

Another novel contribution of the present paper is a cross-linguistic survey of non-local allomorphy, with particular emphasis on root suppletion. Based on the results of the survey, together with non-local patterns reported in the literature, we give an overview of non-local allomorphy. This overview offers a new generalization on distance in non-local patterns: non-local allomorphy appears to be much more conservative than initially expected, as non-local patterns involve 'skipping' at most one overt syntactic head context-wise.

The paper is organized as follows. In section 2, we discuss two key phenomena, viz. local blocking and fusion-allomorphy conspiracies. We show that they can only receive principled explanations if strict locality/adjacency is adhered by in the theory of allomorphy. Section 3 presents the results of a survey of non-local patterns cross-linguistically. Section 4 constructs a novel model of allomorphy by formalizing the core proposal of the paper. Section 5 discusses the alternative approaches to non-local allomorphy, specifically Merchant (2015) and Moskal (2015a), where we point out several conceptual and empirical problems that they struggle with. We demonstrate that the proposed model can, in turn, overcome such problems.

2 Locality bias in allomorphy

This sections presents two types of phenomena that can only receive principled explanations if some notion of strict locality/adjacency is encoded in our model of allomorphy. The first of these are instances of *local blocking effects*, discussed in section 2.1, and the second are *fusion-allomorphy conspiracies* discussed in section 2.2. Since these phenomena coexist with non-local allomorphic patterns *in the same grammar*, this paper proposes that they do not have equal status in the grammar. More concretely, we propose that local allomorphy is in a markedness relationship with non-local allomorphy:

(1) MARKEDNESS SPLIT IN ALLOMORPHY
Local allomorphy represents the unmarked pattern, while non-local allomorphy represents the marked pattern.

The central-most claim of this paper then is that non-local allomorphy is not a general property of grammars, but rather presents a much more marked and marginal pattern in natural language. We touch on this in the following two sections, but we also give further arguments from this in section 3. We derive the proposed split in markedness by arguing for the following 'general schema' that models of allomorphy should follow:

(2) PROPOSAL FOR A SET OF MODELS Λ
A Λ-model of allomorphy must have the following two properties:

a. **Locality Bias**: some preference for strict locality/adjacency needs to be encoded in the grammar

b. **Non-Locality**: the grammar needs to have some means of deriving non-locality, which is in conflict with strict locality

Models of allomorphy need to express some strict locality bias, and they need to have some means of deriving non-locality. This is a very general schema, here termed the set of models Λ (as in Λocal). In this way, this paper argues for a whole set of models, which are such that they have have Λ properties.

In section 5, however, we flesh out a specific member of this set, one that fits current empirical observations best. There, we tackle the question of how a locality bias and non-local allomorphy can co-exist in a single grammar. Specifically, we propose that the function that maps phonological exponents to syntactic heads, viz. Vocabulary Insertion, identifies the context of insertion through a *search procedure*, which we term SCAN. This procedure is subject to an ECONOMY CONDITION at the PF-interface that will require it to consider only minimal context – one head to the left and one to the right. Non-local patterns will be predicted to emerge as *violations* of this economy condition in a principled way. Such a system will be shown to have Λ-properties: it expresses a clear *locality bias*, while still being able to account for the exceptional non-local patterns in allomorphy.

2.1 Local blocking effects

In typical, local patterns of allomorphy,[1] the trigger of allomorphy is adjacent to the target. However, morphological configurations may arise where the trigger and target of allomorphy are separated by an intervening head, which blocks the allomorphic relation between the trigger and target. This is an instance of *local blocking*:

(3) Local blocking effect

$\surd\text{EXPONENT1} - \alpha^0$, $\boxed{\surd\text{EXPONENT2}} - \beta^0$ *trigger is local* $\rightsquigarrow \beta^0$

$\surd\text{EXPONENT1} - \gamma^0 - \alpha^0$, $\surd\underline{\text{EXPONENT1}} - \gamma^0 - \beta^0$ *trigger is non-local* $\rightsquigarrow \beta^0$

A syntactic head, here the root, suppletes in the context of β^0. When β^0 is also in the construction, but not immediately adjacent to the root, no suppletion occurs, i.e. the suppletion relation is *blocked* by γ^0. This is an instance of local blocking. These are instances where adjacency asserts itself a locality condition on allomorphy. It is particulary interesting that non-local patterns coexist with such local blocking in the same grammar. First, consider a non-local pattern from Slovenian, South Slavic:

[1] Two notes are in order. First, we assume with Harley (2014) that roots are subject to Late Insertion, rendering root suppletion a type of VI-derived allomorphy. Secondly, we say nothing about Readjustment Rules at this point, but later on we argue against them. See section 4.

(4) Slovenian verbs and participles (Božič 2016: 139)

ROOT	VERB:1P.SG	PARTICIPLE:F.SG	
√žanj-	žanj-e-m	ž-e-l̲-a	'reap'
√koln-	koln-e-m	kl-e-l̲-a	'swear'
√boj-	boj-i-m	b-a-l̲-a	'fear'

In Slovenian, the root undergoes suppletion in the context of the underlined participial suffix /-l/, and crucially the suppletion here is triggered across the overt suffix /-e/, as argued in Božič (2016). This means that the trigger of allomorphy is non-adjacent, and hence non-local. For more details on this pattern see section 6.3.

However, in Slovenian, root suppletion patterns also reveals instance of local blocking effects. Consider the instance of plural-triggered suppletion in the paradigm of 'man', yielding the alternation between √človek- and √ljudj-:

(5) Local blocking in Slovenian (√človek- ~ √ljudj-)

	SG	DU	PL	
N	√človek-∅	√človek-a	$\sqrt{\boxed{\text{ljudj}}}$-e	'man'
N+DIM	√človeč-ek-∅	√človeč-k-a	√človeč-k-i	
	√RT– (DIM^0–) #^{0}	√RT– (DIM^0–) #^{0}	√RT– (DIM^0–) #^{0}	

The adjacent plural feature triggers the insertion of the contextual item /ljudj-/. However, if the diminutive suffix /-(e)k/ is attached to the root, suppletion no longer occurs in the context of the plural feature because the attachment of the diminutive suffix renders it non-adjacent. This is an instance of local blocking, and it seems that it can appear in the same grammar as a non-local patterns, as shown above. The present paper interprets these data in the following way: the local blocking that occurs in the paradigm of 'man' represents the default state of affairs – *adjacency* asserts itself as a locality condition. The pattern of non-local suppletion in the partciples, however, is a marked phenomenon, effectively an exception to the general locality condition. This claim makes sense, given that the non-local pattern in Slovenian is very marginal, as only a handful of roots undergo it (Božič 2016). The alternative, as predicted by alternative accounts of non-local allomorphy (Merchant 2015; Moskal 2015a) is that non-locality is simply a general property of grammars. This, however, entails that both the local blocking and the non-local pattern are lexical accidents. Specifically, both the following rules have equal status:

(6) Rule 1: √MAN ↔ *ljudj-* /_____[PL]   *status of rule?* ⇝ **local**

Rule 2: √MAN ↔ *ljudj-* /_____Dim^0, [PL]   *status of rule?* ⇝ **non-local**

If non-locality is generally available in the grammar, then the blocking pattern in Slovenian is not a case of real blocking at all: it only occurs because Rule 1 is stored in the lexicon instead of Rule 2. Such an approach is hence unable to give principled explanations for any of the patterns that we have just observed. On the other hand, in the approach proposed in this paper, the blocking effect is predicted to be the *default* and *systematic property of grammar*, while only the non-locality is accidental. This means that the proposal advanced here retains *more* principled explanations than alternative approaches. See section 6 for a detailed comparison of the predictions that these alternative approaches make.[2]

It should be noted that the coexistence of local blocking and non-local patterns is not a property unique to Slovenian. Other languages exhibit similar patterns, which we only briefly note here. As we will see in section 3, Tamil exhibits non-local suppletion in pronouns (Moskal & Smith 2016), but instances of local blocking can be found in its affixal allomorphy (McFadden 2014: 15). Basque adjectival morphology also gives rise to non-local suppletion (Bobaljik 2012: 156-158), discussed in section 3, but instances of local blocking can again be found between case clitics adjoined to the complementizer (Arregi & Nevins 2012: 114-115).

Local blocking, in a Λ-model of allomorphy, thus follows from the active *locality bias* (for adjacency) that the grammar expresses. Non-locality, on the other hand, needs to be encoded as an exception in some way. Further arguments for this split are given in section 3, where we discuss the generalizations uncovered by the cross-linguistic survey of non-local allomorphy.

2.2 Fusion-allomorphy conspiracies

The next phenomenon in which adjacency must play an active role is *fusion-allomorphy conspiracies*. In these phenomena, a non-local trigger of allomorphy gets fused to the head that is local to the target of allomorphy. Through this, fusion renders a non-local trigger local and so facilitates allomorphy. Phenomena of this kind arise in systems that mix agglutinative and fusional morphology, and especially when these two morphology types are seen in the same paradigm. Consider the following schema, which represents a nominal or pronominal, paradigm where a subset of its cells is occupied by agglutinative morphology, and the other subset by fusional:

(7) Fusion-allomorphy conspiracy

	SG	PL
NOM	$\sqrt{\text{EXPONENT1}} - \{\#^0\} - \{K^0\}$	$\sqrt{\text{EXPONENT1}} - \{\#^0\} - \{K^0\}$
GEN	$\boxed{\sqrt{\text{EXPONENT2}}} - \{\#^0\text{+}K^0\}$	$\sqrt{\text{EXPONENT1}} - \{\#^0\} - \{K^0\}$
ACC	$\boxed{\sqrt{\text{EXPONENT2}}} - \{\#^0\text{+}K^0\}$	$\sqrt{\text{EXPONENT1}} - \{\#^0\} - \{K^0\}$

The fusion of $\#^0$ and K^0 facilitates K^0(ase)-triggered suppletion, which is absent in the agglutinative forms. In other words, fusion and suppletion conspire to create this pattern. This pattern can only receive a principled explanation if immediate adjacency plays some role in determining allomorphy.

We now examine data from Georgian along with two other Kartvelian languages. Kartvelian languages reveal the generalization that *fusional* morphology facilitates suppletion because all the functional material is fused into one root-adjacent suffix, while *agglutinative* morphology blocks suppletion. The pattern of interest is found in the pronominal system. Kartvelologists describe Georgian as containing two parallel systems of exponence in the nominal paradigms: an agglutinative system and a 'flexional' system (Tuite 1998: 50), where the agglutinative system is more common and productive, and hence also the default one. Nouns may belong to either

[2] Note that Merchant (2015) and Moskal (2015a) make other fine-grained predictions, which we do not discuss here, but rather examine them in some detail in section 6.

of the two systems, but in pronouns a split agglutinative-flexional system is used. Pronouns show an agglutinative form in the nominative, but flexional forms in the non-nominative cases. Consider demonstrative pronouns, based on Hewitt (1995: 77-78) and Brown et al. (2003), and the noun 'wife', from Tuite (1998: 50):

(8) Georgian 3P Georgian nouns 'woman'

	SG	PL	SG	PL
NOM	eg	ege-n-i / ege-eb-i	kal-i	kal-eb-i
DAT	maga-s(a)	maga-t(a)	kal-s	kal-eb-s
ERG	maga-n	maga-t(a)	kal-ma	kal-eb-ma
GEN	mag-is(a)	maga-t(a)	kal-is	kal-eb-is
INST	mag-it(a)	maga-t(a)	kal-it	kal-eb-it
ADVB	maga-d(a)	maga-t(a)	kal-ad	kal-eb-ad
	$\sqrt{\text{D}^0}$– #^{0}– K^0		$\sqrt{\text{RT}}$– #^{0}– K^0	

The noun 'woman' reveals the most typical inflection pattern for Georgian: the singular #^{0} is null, while the plural head is exponed with /-eb/. Case markers all have overt exponents. In the pronons, as noted above, the paradigm shows a 'flexional' system in the non-nominative cases. The singular part of the paradigm shows similar K^0-exponents to those found in the nouns (except for the ergative form). The plural part, on the other hand, reveals a single exponent throughout the non-nominative cases, viz. /-t(a)/.[3]

A plausible analysis is that the non-NOM forms in pronouns involve a fused [#^{0}+K^0]-head. This seems attractive to propose since /-t(a)/ is used in the plural even though a more default set of #^{0} and K^0-exponents exists in the language. Moskal (2015a: 89) offers some brief discussion on this paradigm: she reaches the same conclusion and notes that fusion is particularly attractive given that /-n/ is the spell-out of the ergative singular, even though /-ma/ is available to spell out ergative otherwise. As is evident from (8), suppletion is involved in the pronominal pattern. One could say that the NOM-forms are the suppletive ones, being sensitive to the non-local K$^0_{\text{NOM}}$. However, it is analytically preferable to posit an analysis that ties together suppletion and the fusion pattern, as these two patterns are perfectly correlated in the Georgian pronominal system. The fusion pattern is widespread in Georgian pronouns, as it can also be found in proximate and distal demonstratives, and in third person pronouns. In all these cases, suppletion correlates with fusion as in (8). This analysis involves the following set of VI-rules:[4]

(9) VI-rules for Georgian demonstratives
[K$^0_{\text{NOM}}$] $\longleftrightarrow$ $-i$
[PL, K^0] $\longleftrightarrow$ $-ta$
[D^0] $\longleftrightarrow$ $eg-$
[D^0] $\longleftrightarrow$ $maga-$ / _____[PL, K$^0_{\neg\,\text{NOM}}$]

[3] The /a/ segment appears to be optionally present, since it often undergoes truncation (Tuite 1998: 50). It is unclear if this /a/ is perhaps a separate suffix, but if so, it spells out a head that is positioned above K^0. This must be the case because the DAT, GEN, INST and ADVB suffixes in the singular are clearly the same suffixes that expone K^0 in the nouns, and they appear to the left of /a/. /-a/ is perhaps coding Person, but this can be set aside for our purposes.

[4] The '¬NOM' context is just convenient short-hand for encoding the fact that suppletion occurs in non-NOM cases. This could be formalized in various ways, e.g. by assuming a hierarchical approach to case (Caha 2009), where the reference to DAT-case would force reference to all other non-NOM forms.

An immediate explanation for this 'conspiracy' of fusion and suppletion can be given in a system that makes use of adjacency as a locality condition on allomorphy: fusion facilitates suppletion by making K^0 adjacent to D^0. However, in approaches where non-local allomorphy is freely available, as in Merchant (2015) and Moskal (2015a), no such prediction can be made: the fusion-suppletion correlation cannot be captured in such an approach, as it predicts that K^0-triggered suppletion in the NOM-form has equal status to suppletion in the non-NOM forms, which effectively reduces the correlation to a lexical accident.

It is interesting to observe what other Kartvelian languages reveal in their pronominal systems. First, let us mention that some Georgian dialects do away with suppletion and level the paradigm. Consider the 3P plural paradigm from Lower Imeretian, a dialect of Georgian:

(10) Lower Imeretian (Tuite 1998: 55)

	PL
NOM	mage-n-i
DAT	mage-n-ma
ERG	mage-n-s
GEN	mage-n-is

We find the same tendency in two other related Kartvelian languages, viz. Laz and Mingrelian, where the 3P pronouns reveal purely agglutinative inflection and no suppletion:

(11) Laz — Mingrelian (Tuite 1998: 55) z

	Laz		Mingrelian (Tuite 1998: 55) z	
	SG	PL	SG	PL
NOM	mu-k	mu-t-epe-∅	mu-∅	mu-n-epi-∅
DAT	mu-s	mu-t-epe-s	mu-s	mu-n-en-s
ERG	mu-k	mu-t-epe-k	mu-k	mu-n-en-k
GEN	mu-ši	mu-t-epe-ši	mu-ši	mu-n-ep-iši

The generalization seems to be that K^0-driven suppletion tends to correlate with fusion, but not with agglutination in Kartvelian languages. This confirms the previously stated idea that some sort of *locality bias*, which is beyond accident, is involved here. Allowing free non-locality, on the other hand, cannot capture the generalization in Kartvelian. This is an important argument for a Λ-type model of allomorphy, fleshed out at the start of this section, where adjacency is encoded as some type of locality bias that constrains allomorpy.

3 Crosslinguistic Overview of Non-Locality

In this section we provide a schematic overview of non-local cases of allomorphy cross-linguistically. The focus will here be on the distance between the target of insertion and context, showing that the distance is much more conservative than existing non-local approaches predict. The following general method was employed in identifying non-local patterns: in cases such $\sqrt{\boxed{X}}\text{–}Y\text{–}\underline{Z}$, where X was the target of suppletion and Z the trigger, the pattern was deemed non-local. In cases such as $\sqrt{\boxed{X}}\text{–}\underline{Y}\text{–}Z$, $\underline{Z}\text{–}\sqrt{\boxed{X}}\text{–}Y$ or $\underline{Z}\text{–}\sqrt{\boxed{X}}\text{–}\underline{Y}$, the pattern was deemed local. Only cases where the interveners were overt were considered, since null exponents may easily be re-analyzed some other way. More concretely, we assumed the common observation that null exponents are not interveners of allomorphic processes (Siddiqi 2006, 2009; Embick 2010; Arregi & Nevins 2012) and, therefore, did not treat patterns with null interveners as cases of non-local allomorphy. Only clear cases of suppletion were selected in this survey.

We first tackle *root suppletion*. The data presented here were collected by examining the Surrey Suppletion Database (Brown et al. 2003), which consists of 34 languages from distinct families. Below are the general results for suppletion:

(12) Surrey Suppletion Database results

Number of lang.	*Local Suppletion?*	*Non-Local Suppletion?*
34	31	4

31 languages in the database showed local suppletion, but only 4 languages showed a non-local suppletion pattern. This confirms the general observation that non-local suppletion is much rarer than local suppletion. Below, we give a list of the uncovered non-local patterns. In order to generalize about the distance involved in non-local suppletion, we also include 5 other patterns that have been reported in the literature. This amounts to 9 cases of non-local suppletion. Boxes represent the targets of suppletion, while underlining represents the triggering context. The third column represents the category of the word.

(13) Overview of Non-Local Suppletion

LANGUAGE	PATTERN	CAT	SOURCE
1. *Greek*	$\boxed{\sqrt{\text{RT}}}\text{-}\underline{\text{Voice}^0\text{-Asp}^0}$	V	Merchant (2015)
2. *Slovenian*	$\boxed{\sqrt{\text{RT}}}\text{-Asp}^0\text{-}\underline{\text{Ptc}^0}$	V	Božič (2016)
3. *Tamil*	$\boxed{\text{D}^0}\text{-\#}^0\text{-}\underline{\text{K}^0}$	D^0	Moskal & Smith (2016)
4. *Totonac*	$\boxed{\sqrt{\text{RT}}}\text{-Asp}^0\text{-}\underline{\text{Agr}^0_{2\text{P}}}$	V	Brown et al. (2003)
5. *Lak*	$\boxed{\sqrt{\text{RT}}}\text{-\#}^0\text{-}\underline{\text{K}^0}$	N	Radkevich (2014)
6. *Tariana*	$\boxed{\sqrt{\text{RT}}}\text{-Cl}^0\text{-}\underline{\text{\#}^0}$	A	Brown et al. (2003)
7. *Ket*	$\underline{\text{T}^0}\text{-AgrO}^0\text{-}\boxed{\sqrt{\text{RT}}}$	V	Brown et al. (2003)
8. *Basque*	$\boxed{\sqrt{\text{RT}}}\text{-Dim}^0\text{-}\underline{\text{Cmpr}^0}$	A	Bobaljik (2012)

These data are, of course, preliminary and each of the patterns deserves careful attention. But it is necessary to present them in a succint way here, first, for reasons of space, and secondly, to create and overview which permits us to observe their commonalities and differences, in order to construct a typology of locality. The relevant minimal pairs for each pattern are collected in Appendix I at the end of the paper.

Now, we turn to AFFIXAL ALLOMORPHY. Unfortunately, no database dedicated to affixal allomorphy exists, so the conclusions drawn for its patterning are by definition more tentative. Much fewer cases of non-local allomorphy have been reported in contrast to suppletion. Here are the two plausible cases of non-locality:[5]

[5] Another less plausible case exists in Itelmen (Kamchatkan), reported in Bobaljik (2000). In this language, the verbal complex consists of $\text{AgrS}^0\text{–}[\sqrt{\text{RT}}\text{–T}^0\text{–THM}^0\text{–AgrO}^0]$, where THM-allomorphy is

(14) Non-Local Affixal Allomorphy

LANGUAGE	PATTERN	CAT	SOURCE
1. *Kiowa*	$\sqrt{\text{RT}}$-$\underline{\text{v}^0}$–Neg^0-Dist^0-$\boxed{\text{Mod}^0}$	V	(Bonet & Harbour 2012)
2. *Bulgarian*	$\sqrt{\text{RT}}$-$\underline{\text{Thm}^0}$-$\underline{\text{T}^0}$-$\boxed{\text{Agr}^0}$	V	(Stump 1996; Scatton 1984)

One non-local pattern is found in Kiowa (Tanoan), and the other in Bulgarian (South Slavic). The Kiowa case involve significant distance between the trigger and target of allomorphy, while the distance is smaller in Bulgarian. The Kiowa pattern, actually, requires additional discussion: whether it is truly an instance of non-locality is more difficult to determine than is obvious at first glance. See section 6.5 for additional discussion. The relevant data are collected in Appendix II.

Two generalization can be made about the data that we have considered: one is about the status of locality in these patterns, and the other about the distance involved in non-locality. We start with the STATUS OF LOCALITY. It is accepted in much literature on allomorphy within Distributed Morphology that non-local patterns are much rarer than local patterns. This is directly supported by our findings, as already noted above. In just the results from searching the Surrey Suppletion Database, 31 languages showed local suppletion, while only 4 showed non-local suppletion. One might even go further and ask whether every language with non-local allomorphy shows *some* local allomorphy. This indeed turns out to be the case. We briefly indicate these cases in the following paragraph.

Apart from the non-local pattern given in (13), Greek does have local allomorphy, e.g. the sensitivity of Agr^0 to T^0 (Merchant 2015: 277).[6] Slovenian shows instances of local allomorphy, as well, as discussed in section 3.1, where we also showed the same for Tamil. In Totonac, verbs such as 'come' and 'go' undergo suppletion conditioned by person features with no evident intervener (Brown et al. 2003). In Tariana, several roots supplete in the context of Vocative features with no evident interveners (Brown et al. 2003; Aikhenvald 2003). In Ket, the root of the noun 'man' also suppletes in the context of an adjacent plural feature (Brown et al. 2003; Werner 1997). In Basque, we can find instances of local blocking effects, where the spell-out of T^0 is sensitive to an adjacent ergative clitic (Arregi & Nevins 2012). In Kiowa, the root of the verb 'drop', among others, suppletes in the context of adjacent [±AUGMENTED] features (Harbour 2008: 129). In Lak, the root of the noun 'horse', among others, suppletes in the context of the local #^{0} (Mel'čuk 2000: 516). In Bulgarian, among others, the possessor head undergoes allomorphy, conditioned by the preceding adjacent vowel /-a/ of the stem (Harizanov 2014: 248).

triggered jointly by the features of AgrO^0 and AgrS^0, according to Bobaljik. However, the pattern is surrounded by certain suspicious facts: analyses of Itelmen agreement (Bobaljik & Wurmbrand 2002; Keine 2010) endow each of the Agr-heads with a probe, where AgrO^0 needs to probe the subject if the object is 3P. All the cases of THM-allomorphy analyzed by Bobaljik occur when the object is 3P, which is a suspicious coincidence. Bonet & Harbour (2012: 232) offer possible suggestions for re-analysis, involving either purely syntactic agreement, or a combination of agreement and local allomorphy. We exclude this pattern from our dataset.

[6] Even in the patterns that involve non-local suppletion in Greek, discussed in section 3.1, only a subset of those forms actually involves overt interveners. This means that the remaning forms involved in the pattern actually involve local suppletion. That being said, Greek does contain a significant number of roots that seem to undergo some sort of non-local suppletion (Merchant 2015: 281). The point, however, is that even Greek is not entirely free of local patterns, which again makes the non-local ones look 'irregular'.

This indicates that, not only is non-local allomorphy rarer than local allomorphy, but also that it is somehow secondary to local allomorphy. For instance, in Slovenian we do find a small non-local pattern, but the rest of the patterns in the language seem to be fully local (Božič 2015, 2016). No language, to the best of our knowledge, has been uncovered where non-local allomorphy is the pervasive and primary pattern. This suggests a tentative universal generalization along the following lines:

(15) *Locality Implication*: *non-local* $\Longrightarrow$ *local*
If a language exhibits non-local contextual allomorphy, it also exhibits local contextual allomorphy.

This generalization need not mean that non-local allomorphy is actually derived from local allomorphy, but it encodes the observation that non-local allomorphy is more exceptional and not the default pattern that we expect to observe. In sum, it seems inadequate to award local and non-local allomorphy equal status in our theory of allomorphy. This generalization suggests that local allomorphy represents the default, *unmarked* pattern, and that non-local allomorphy is much more *marked*. Acknowledging a distinction in markedness does not in its own derive the Locality Implication, but it is useful to talk of unmarked vs. marked patterns, since that places them in a specific relationship.

We now turn to the DISTANCE IN NON-LOCALITY. If we consider our data in terms of the distance between triggers and targets of allomorphy/suppletion, we find a very strong trend to only be non-local for one extra head. In fact, all of the patterns of root suppletion are like this, as is also the Bulgarian case of affixal non-locality. The only seeming counter-example is Kiowa. However, the validity of this pattern is not completely clear, as discussed in section 6.5. For now, we take a more restrictive stance and generalize about the possible distance between trigger and target across all morphemes:

(16) *Distance in Non-Locality*
Non-local allomorphy/suppletion can only involve treating two heads as context and not more.

What is fascinating about the Distance in Non-Locality, in a general way, is how conservative non-local patterns appear to be in terms of distance. Among alternative approaches to non-local allomorphy, Merchant (2015) predicts that non-local suppletion can in principle involve all the heads between the root and the very top of the extended projection, which can result in very many overt intervening heads. This prediction is not borne out, to the best of our knowledge. This means that not only does our theory of allomorphy require some notion of *strict locality* combined with *non-locality*, as discussed in the previous section, but also a constraint that controls the distance in non-locality in some way.

4 Proposal: Re-Labelling and PF Economy

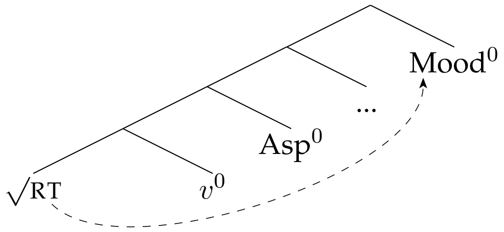
The preceding sections have determined that local and non-local allomorphy do not have equal status in the grammar, and that VI itself needs to be subject to locality constraints. The preceding sections have, hence, established what we set out to do in section 2, viz. that a theory of allomorphy needs to have Λ-model properties: some notion of strict locality, constrained VI and some way to derive non-local allomorphy. This section develops a member of the set of Λ models. The primary question is, how can local and non-local allomorphy coexist in the grammar? In addition, our model needs to be able to address Generalizations 1–2, uncovered in the previous section. We begin by proposing a formal definition of VI.

Before we proceed with formalizing VI, we need to discuss the role of Readjustment Rules and *null exponenece* in our model. Along with an increasing trend in Distributed Morphology, we assume that all allomorphy is derived by VI, rejecting Readjustment Rules (Siddiqi 2006, 2009; Bye & Svenonius 2012; Bermúdez-Otero 2012; Gribanova 2015) – see section 5.2 for some explicit criticism. We also observe the generalization that null heads do not count as interveners of allomorphy (Siddiqi 2009; Embick 2010; Arregi & Nevins 2012). This can be derived by subjecting all null heads between the target of insertion and the first overt head to a process of *generalized fusion* to the target, as proposed by Siddiqi (2006, 2009).[7] We do not expand on this point further here.[8] All the structures, given from now on, omit any kind of null intervening heads, unless stated otherwise.

4.1 Defining VI

We now turn to a formal definition of VI. Based on Trommer (1999), we assume that VI is defined as the *3-tuple* ⟨PHON, TARG, CTXT⟩, where 'CTXT' constitutes context and 'TARG' the target of insertion, and 'PHON' the phonological exponent that is inserted. We propose that CTXT is a defined as a *buffer* $\mathfrak{B}$, which consists of two slots $\{S_L, S_R\}$, each of them labelled with a 'left' or 'right' label, designated for storing left or right context. We also propose that VI performs a 'search' procedure that identifies the context of the target of insertion: this search procedure identifies a left and right head, and stores them in $\mathfrak{B}$. This is how VI formally accesses context. Before we proceed to define $\mathfrak{B}$, we need to determine what this search procedure is looking to identify. One possible option is that this procedure is very much like AGREE in syntax (Chomsky 2000), meaning that it searches for a particular *syntactic category*. Let us assume insertion at the root, and that VI performs a search for Mood^0:

(17) Searching for syntactic category: S_{MOOD}



[7] Siddiqi's system of fusion is driven by the PF economy constraint MINIMIZEEXPONENCE. In his system, only the null heads between the target of insertion and the first overt head can be fused (they are fused to the target of insertion): fusion applies freely up until it encounters the lexical entry for a non-null head, which blocks fusion. It is actually unclear at this point if fusion should apply even more generally, simply fusing all null heads to the overt ones across the entire word before VI takes place. This would certainly be favourable to MINIMIZEEXPONENCE, but it is unclear if this is needed.

[8] This is also important if our theory of allomorpy is to be compatible with some theories of narrow syntax, e.g. the Cartographic Approach (Cinque 1999), where numerous null heads are posited.

When inserting at the root, VI could search for a Mood^0-head, identify it, and then store it in $\mathfrak{B}$. However, notice that such a version of this search procedure can freely skip any head that intervenes between the target and the category it searches for. Under a this formulation, we would expect VI to be able to search all the way to the top of the extended projection without any kind of obstacles, expressing no effects of strict locality. Of course, we could add a stipulation and say that the search procedure is only allowed to look at immediately adjacent heads. However, this is undesirable since the notion of locality in allomorphy then remains underived, as the locality stipulation does not follow from anything already defined in the system.

This is an important point, and we take it to mean that VI *does not* search for syntactic categories directly in any way. Rather, we want to say that VI searches for a different kind of category, one connected to the *position* in the syntactic tree. We propose that the search operation is looking for *position information*: it needs to find one *left-adjacent* head, and one *right-adjacent* head, regardless of what category the relevant heads are. This makes the VI search operation unlike AGREE, in terms of what it is searching for. However, this is not unexpected: searching for position in the tree matters at the PF-interface, while in narrow syntax searching for categories is what matters. The different properties of these two operations immediately give rise to different locality constraints: AGREE can probe until it finds the first category it needs, while the VI search can 'probe' until it finds the first adjacent position it needs. We further discuss the notion of 'position' below.

We assume that it is the directionality labels L and R, $\mathfrak{B}$ = {S_L, S_R}, that trigger this search procedure. We assume terminology familiar from syntax to describe the operational aspects of this search procedure. We assume that {L, R} can either be *valued* or *unvalued*: they are unvalued if their S is empty, but if filled, they get *valued*. We represent *valuedness* with '+' and *unvaluedness* with '−', borrowing the notation from Rezač (2004). It is $-L$ and $-R$ that trigger the search procedure. We call the VI search procedure SCAN. We are now ready to define $\mathfrak{B}$:

(18) $\mathfrak{B}uffer$
CONTEXT is accessed through $\mathfrak{B}$, where $\mathfrak{B}$={S_α, S_α}. Let α be the set of directionality labels {L, R}, which can be either *valued* (+) or *unvalued* (−).

a. $|\mathfrak{B}| \leq 2$ and for every S there is precisely one α.

b. α = {$\pm L$, $\pm R$}: empty $S \rightsquigarrow -\alpha$, filled $S \rightsquigarrow +\alpha$.

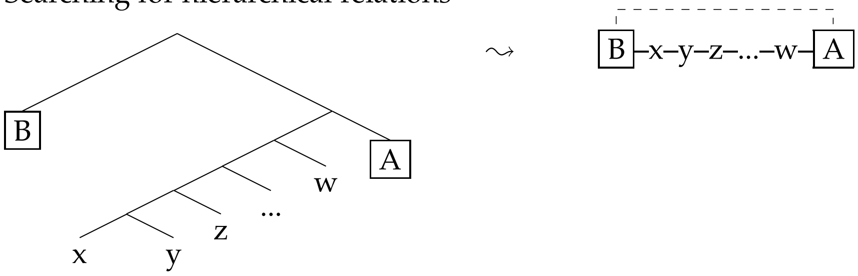
c. $-\alpha$ triggers SCAN to fill S and so receive a value.

To recap, each of the unvalued directionality labels trigger SCAN in order to fill their slot and so receive a value. Let us briefly touch on clause (18a) of the definition: it states that $\mathfrak{B}$ can be comprised of maximally two slots and each of the slots can have only one directionality label. This prevents the system from having several slots and so accessing more heads than is minimally required. This appears to be stipulation at this point, but we later show that it can be grounded in simple computational properties of the procedure. Now that we have provided a definition of $\mathfrak{B}$, which refers to the SCAN procedure, it itself also requires a definition, given below:

(19) SCAN
Search for a head H$_\alpha$ of category α, where α = {L, R}, s.t. $\mathfrak{B}$ = {S_α, S_α}. Upon finding the first H$_\alpha$, store H in $\mathfrak{B}$. Storing H in $\mathfrak{B}$ values $-\alpha$, making it $+\alpha$.

One aspect of the system developed here has not yet been defined, viz. the directionality labels $\{L, R\}$. Before we do so, it is important to explain that these labels actually search for *linear information* in the syntactic tree. Along with Arregi & Nevins (2012), we assume that syntactic structure undergoes a process that adds precedence information to the tree at PF, and that VI accesses this structure directly. This by definition means that linear information is retrievable from the tree. However, why could we not say that SCAN actually searches for information about the *hierarchical positions* in the tree? It would be possible to assume that SCAN simply looks at the structurally 'preceding' head (one head down) and structurally 'following' head (one head up). However, this seems to make predictions that are incompatible with a view that enforces linear adjacency. Consider the following:

(20) Searching for hierarchical relations



Given the structure above, assume that we are inserting at A. If SCAN were searching for hierarchical positions, then the two local heads for A would be w and B: w is 'one head down', and B is 'one head up'. Notice that no matter how many overt heads there are between x and w, A and B would always be in a *completely local relation*. This would then entail that completely local, and hence default, allomorphy could happen across vast numbers of overt heads in the linear string – this is shown in the right-hand side above. We do not find such effects in languages cross-linguistically. Specifically, the view where SCAN searches for hierarchical position makes the incorrect prediction that the relation between A and B is a local, and hence the expected, default allomorphic relation. We take this to mean that SCAN must be searching for *linear information* in the tree. We can now define *Set* α:

(21) *Set* α
α = $\{L, R\}$, where L denotes a relation of immediate *linear* L-adjacency to the target of insertion, and R denotes a relation of immediate *linear* R-adjacency to the target of insertion.

A similar conclusion about the role of linear information is reached by Ostrove (2016), who demonstrates that C^0 triggers root suppletion in Irish Gaelic; he argues that C^0 is structurally distant from the root, but linearly adjacent to it, which entails that linear position is important for VI. We have now defined *Set* α.[9]

4.2 Locality and Non-Locality: PF Economy

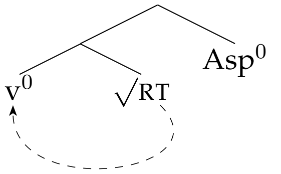
Let us now walk through an example of VI application. This will enable us to discuss the distinction between local and non-local allomorphy. In the example here, let us

[9] What now remains undefined in this model is the specific formal *algorithm* that is employed in identifying immediate adjacency in the tree. This is something that will remain undefined in this paper, and is best left for future research.

assume the structure [[v^0 [$\surd$RT]] Asp^0], where VI is inserting at the root:

(22) Inserting at $\surd$RT with default $\mathfrak{B}$ = {S_{-L}, S_{-R}}

a. First Scan increment

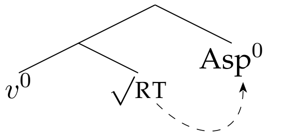


Scan increment #1: S_{L-}

⟶ SCAN(S_L) → discover v^0

⟶ $\mathfrak{B}$ = {[v^0]$_{+L}$, S_{-R}}

b. Second Scan increment



Scan increment #2: S_{-R}

⟶ SCAN(S_R) → discover v^0

⟶ $\mathfrak{B}$ = {[v^0]$_{+L}$, [Asp^0]$_{+R}$}

VI targets $\surd$RT, where $\mathfrak{B}$ has two slots, both of which carry unvalued directionality labels. Because there are two slots, and hence two labels, two increments of SCAN occur. In the first increment, $-L$ triggers SCAN, and finds v^0, storing it in $\mathfrak{B}$, valuing L. In the second increment, $-R$ triggers SCAN and finds Asp^0, storing it in $\mathfrak{B}$, which also values R. Now that both of the labels have been valued, the determination of context is complete. The exponent that matches the context is selected and inserted, in the standard way. VI can now move to the following head in the tree, which is v^0.

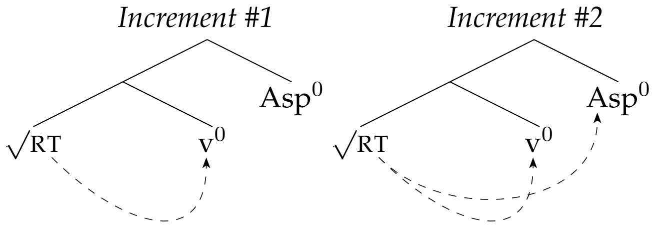
This kind of operation represents the local pattern of allomorphy, which we have called the default or *unmarked* in section 5. We propose that what defines the unmarkedness of this local operation is the following economy constraint:

(23) BUFFER ECONOMY
Access each $S \in \mathfrak{B}$ only once. (No unneeded tampering with $\mathfrak{B}$!)

This is an economy constraint that ensures that VI only scans for the minimal context, i.e. as little as it requires to determine its environment. More concretely, this predicts that VI will by default consider only minimal context, as it will be able to access a slot in $\mathfrak{B}$ only once: once when storing a head in S_L, and once when storing a head in S_R. This ensures a strictly local VI, as in (Embick 2010; Arregi & Nevins 2012). The novel contribution here is that this constraint is an economy condition of the PF-interface that is tied directly to the formal workings of VI.

It is important to observe that BUFFER ECONOMY is actually grounded in the computational properties of the system that has been defined so far. This is best observed by attempting to derive a non-local pattern of allomorphy. Let us consider [[[$\surd$RT] v^0] Asp^0], where VI is inserting at the root:

(24) Instance of non-local allomorphy



An instance of non-local allomorphy (suppletion, in this case) would arise when SCAN would need to access v^0, as well as Asp^0, when inserting at the root. This would force us to assume that $\mathfrak{B}$ in this case is, in fact, not {S_L, S_R}, but rather {S_{R1}, S_{R2}}, since the former could never scan anything more than v^0 to the right, but the latter could scan v^0 and Asp^0. However, what is crucial to observe about such a procedure is the following. $R1$ would trigger the first Scan increment: v^0 would be identified and sent to $\mathfrak{B}$. Then, $R2$ would trigger the second Scan increment, and this is where something additional will happen. In this second increment, v^0 would again be identified and stored in $\mathfrak{B}$. But since $\mathfrak{B}$ already contains an identical element, viz. the same v^0, a *clash* would occur and the second v^0 would be deleted. Only then would Scan identify Asp^0 and finally store it in $\mathfrak{B}$.

Notice that deriving such a non-local pattern is directly correlated with *increased computational complexity*. In other words, scanning for more than a single head in one direction is harder. This seems to be the property that we actually want non-local patterns to have: if increased computational complexity is correlated with the markedness of non-locality, then this derives the *split in markedness* in allomorphy, viz. that of local vs. non-local allomorphy. It should also be noted that this computational procedure involved in deriving non-locality is precisely the factor in which BUFFER ECONOMY is grounded. Notice that deriving non-local patterns *violates* BUFFER ECONOMY: the second slot in $\mathfrak{B}$, $R2$ is accessed twice: once when storing v^0, which is then removed from the buffer, and then when storing Asp^0. In this way, BUFFER ECONOMY is effectively a constraint that strives for minimal computation.

At this point, it is useful to return to the definition of $\mathfrak{B}$ in (18), where we specified the following clause, as a restriction on $\mathfrak{B}$, repeated below for convenience:

(25) $|\mathfrak{B}| \leq 2$ and for every S there is precisely one α.

This clause effectively means that there can be only two slots in $\mathfrak{B}$, and each can have only one label. We mentioned, in that section, that this restriction still needs to be explained. This restriction can also fall out of BUFFER ECONOMY, and by transitivity, out of principles of minimal computation. For instance, imagine that the restriction '$|\mathfrak{B}| \leq 2$' was lifted, which would mean that $\mathfrak{B}$ could have, for instance, three slots. Since the cardinality of the set of labels {L, R} logically cannot exceed *two*, this means that the extra slot in $\mathfrak{B}$ would either have a L or R-label. This, in turn, means that it would have the same label as one of the other two slots in $\mathfrak{B}$. Using such a buffer would violate BUFFER ECONOMY because the same procedure would result as in (24): the same head would be scanned at least twice. The same applies to having multiple labels on the slots. This is then telling us that the restriction repeated in (25) is simply a design property of the system, but one that is grounded in principles of Minimal Computation. This is a welcome result, since Minimal Computation plays a significant role in current Minimalist inquiry (Chomsky 2013: 41).

4.3 Re-Labelling Hypothesis

We have now formally derived the notion of strict locality and also the split in markedness that local vs. non-local patterns present. However, the system, as it stands, cannot actually derive non-local patterns because they violate BUFFER ECONOMY. We propose a simple solution to this issue, viz. we assume that BUFFER ECONOMY is in fact a violable constraint. It is conceivable that there a pressures against violating it, since it favours minimal computation, but nevertheless it can be violated. The violations, however, are constrained by the way $\mathfrak{B}$ is defined: recall that more than two slots cannot be specified in $\mathfrak{B}$ and that each slot can only have one label. However, we propose that the language learner may RE-LABEL the slots in order to encode a non-local pattern:

(26) RE-LABELLING HYPOTHESIS
A non-local pattern is encoded by using $\mathfrak{B}'$ in the VI *3-tuple*. The labelling on the S-slots in $\mathfrak{B}'$ is manipulated to encode the desired pattern.
Possible options $\longrightarrow$ $\{S_L, S_L\}$, $\{S_R, S_R\}$

For a specific syntactic head, be it a root or affix, $\mathfrak{B}$ may be re-labelled in the fashion described above. For instance, if a root, say $\sqrt{\text{EAT}}$, undergoes non-local suppletion, we propose that its buffer is re-labelled. This means that non-locality will only be tied to a construction associated with this root,[10] as shown below:

(27) DEFAULT CASE: $\langle$TARG, PHON, $\{S_L, S_R\}\rangle$
RE-LABELLING: $\langle\sqrt{\text{EAT}}$, PHON, $\{S_L, S_L\}\rangle$ or $\langle\sqrt{\text{EAT}}$, PHON, $\{S_R, S_R\}\rangle$[11]

This may, at first glance, appear to be a rather 'brute-force' solution to the problem of non-locality, but it actually adequately expresses the observations that we have made about non-locality: it is very marked and it is tied to specific constructions, never to entire grammars. And it is important to re-emphasize that suppletion, in general, is *not* a well-behaved, categorical property of the grammar: it emerges through historical accidents such as lexicalization of phonology and paradigm mixing (Ronneberger-Sibold 1988; Mel'čuk 2000: 520), and the same is true of affixal allomorphy (Paster 2006). In this sense, re-labelling emerges as a *last resort strategy* to capture a non-local pattern in a system that is designed to encode local patterns. This explains the marked and exceptional status of non-local allomorphy.

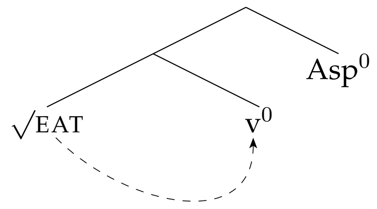
Let us now walk through the steps needed to derive a non-local pattern with re-labelling. Assume the structure [[[$\sqrt{\text{EAT}}$] v^0] Asp^0], where insertion occurs at the root. $\sqrt{\text{EAT}}$ undergoes non-local suppletion, by re-labelling $\mathfrak{B}$, such that VI is here defined as $\langle\sqrt{\text{EAT}}$, PHON, $\{S_R,S_R\}\rangle$. In the first Scan increment (28a), $-R$ triggers Scan and discovers v^0, which is then stored to $\mathfrak{B}$. This first increment is still local.

[10] The non-locallity may only happen in the verbal complex, for instance, but the root can still be used in nominal constructions, where no non-locality is apparent. This is not in itself a problem because nothing is forcing the root to actually undergo suppletion in nominal constructions.

[11] This kind of implementation makes a specific prediction: if $\sqrt{\text{EAT}}$ needs $\{S_R, S_R\}$ in verbal constructions, then it is associated with $\{S_R, S_R\}$ *throughout the grammar* and cannot show local allomorphy where it is conditioned by an L-adjacent head, say, in nouns. Whether this prediction is borne out is unclear at present. However, a different implementation can easily be proposed that fixes this. We could assume a more elaborate definition of VI, viz. $\langle$TARG, $\langle\langle$PHON, CTXT$_i^{lex}\rangle$, CTXT$_i\rangle\rangle$, where CTXT^{lex} is the context lexically specified by the VI-rule that inserts PHON, and CTXT is the actual buffer that performs the SCAN in the tree. CTXT^{lex} and CTXT are coindexed, implying that their contents must match for insertion to occur – this can force the buffer the have the kind of slot specification (either default, or re-labelled) that the VI-rule needs. This entails that as soon as a relevant syntactic head is selected for TARG, the list of available candidates for insertion would be retrieved, and each of the candidates would have a specified contextual restriction, viz. CTXT^{lex}, which the buffer (=CTXT) would need to match. This implies that a Scan would be performed for every VI-rule associated with a target, until the system found the one that fit the given context in the tree best. Such a system can derive different localities of the same root/head in different constructions, but it is unclear if this is needed. The system in (27) is simpler, so we adopt that one in this paper.

(28) Inserting at √EAT with re-labelled $\mathfrak{B} = \{S_{-R}, S_{-R}\}$

a. First Scan increment

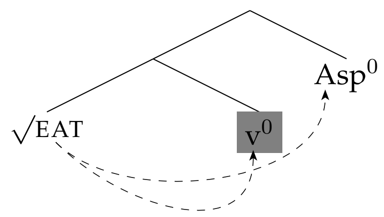


Scan increment #1: S_{-R}

$\longrightarrow$ SCAN(S_R) → discover v^0

$\longrightarrow \mathfrak{B} = \{[v^0]_{+R}, S_{-R}\}$

b. Second Scan increment



Scan increment #2: S_{R-}

$\longrightarrow$ SCAN(S_R) → discover v^0

$\longrightarrow \mathfrak{B} = \{[v^0]_{+R}, [\ v^0\ ]_{+R}\}$ ⇝ CLASH!

$\longrightarrow$ SCAN(S_R) → discover Asp^0

$\longrightarrow \mathfrak{B} = \{[v^0]_{+R}, [\text{Asp}^0]_{+R}\}$

In the second increment (28b), the second $-R$ triggers Scan, and v^0 is again identified and stored in $\mathfrak{B}$. However, having two identical elements in the buffer causes a 'clash', and the second v^0 is removed. Since this second $-R$ is still unvalued, it continues the Scan, which discovers Asp^0 and stores it in $\mathfrak{B}$, which values the label and the derivation converges. The appropriate exponent that matches this context is selected in the standard way and insertion at the root is complete.

Another technical detail remains to be cleared up. The buffer, as defined here, is an inherent part of VI in that it allows VI to access context in the syntactic structure. The buffer does not replace the lexical context specification for individual lexical items/exponents, i.e. the 'VI-rules'. Every exponent that is inserted still needs to be lexically specified for context it will occur in (unless it is the elsewhere item), in the traditional sense. The buffer only regulates how much syntactic structure VI can 'see'. In a typical insertion, a syntactic head is first targetted by VI, then a Scan derivation is performed, and after that the contents of the buffer are matched with the best fitting exponent (i.e. with the exponent's contextual restriction) in the lexicon. Thus in principle, the buffer need not directly constraint VI-lexical entries in any way – they could specify big amounts of non-local context – but there is no reason for such entries to occur in the lexicon, as the buffer cannot compute them.

4.4 Empirical Predictions

The presence of BUFFER ECONOMY successfully derives the Locality Implication from section 5, viz. that non-local allomorphy is more marked than local allomorphy. It also ensures that our system of allomorphy is strictly local by default. The Re-Labelling Hypothesis, on the other hand, provides a narrow window of non-locality. The combination of these properties make the proposed approach a Λ model of allomorphy, for which we have argued throughout this paper (see section 2). However, the Re-Labelling Hypothesis also derives the Distance in Non-Locality from section 5, viz. that non-local allomorphy is only non-local for one extra head.

We must now consider the predictions. The Re-Labelling Hypothesis predicts that two types of non-local patterns exist: {L, L} and {R, R}. As noted above, this derives the Distance in Non-Locality, which is concerned with the distance between the trigger and target. Looking back at the overview of non-local suppletion in (13) in section 5, we can see that all the predicted patterns are attested: most of them are of type $\{R, R\}$, and one is of type $\{L, L\}$ – this is the Yeniseian language, Ket.

(29)	*Local* →	*Attested?*	*Non-Local* →	*Attested?*	*Non-Local* →	*Attested?*
	$\{S_L, S_R\}$	✓	$\{S_R, S_R\}$	✓ $(Greek)$	$\{S_L, S_L\}$	✓ (Ket)

In terms of affixal non-locality, shown in (14) in section 5, Bulgarian is an instance of a $\{L, L\}$-type pattern, while Kiowa appears to be a $\{L, L, L\}$-type pattern. Kiowa is the only example that may falsify the Re-Labelling Hypothesis, or at least the version that we gave in the previous section. However, the Kiowa pattern has certain suspicious ponological properties which suggest it may not be a counter-example at all – see section 6.5 for discussion on this.

The Re-Labelling Hypothesis makes another prediction. In cases where $\mathfrak{B}$ is re-labelled to encode non-locality in one direction, for instance $\{R, R\}$, the proposed system predicts that the head which undergoes this re-labelling cannot be sensitive to a local head in the L-direction. This needs to be the case, because sensitivity for an extra head in the L-direction would imply $\mathfrak{B}$=$\{S_L, S_R, S_R\}$, but this system disallows having more than two slots or labels within the buffer, as we discussed in section 4.2. We are presently not aware of any such cases of allomorphy. If they are found, this would imply a weakening of the Re-Labelling Hypothesis.

4.5 Notes on Affixal Non-Locality: Kiowa

We now turn to the pattern of non-local affixal allomorphy found in Kiowa (Tanoan), first shown in section 5 in (14). As discussed there, Kiowa reveals a process of inward allomorphy that seems to be non-local for three heads to the left, which goes against the Re-Labelling Hypothesis, as stated in the previous section. However, due to certain suspicious phonological properties of this pattern, we will suggest that an alternative account be considered, one that complies with the Re-Labelling Hypothesis, and in fact interprets the Kiowa case as a local pattern. In the Kiowa verbal complex, the modality suffix, Mod^0, is sensitive to the transitivity features in v^0 (Watkins 1984: 170), but overt Neg^0 and Distr^0(ibutivity) can intervene between them, as is noted in Bonet & Harbour (2012), without blocking the allomorphic relation. Here is the relevant minimal pair:

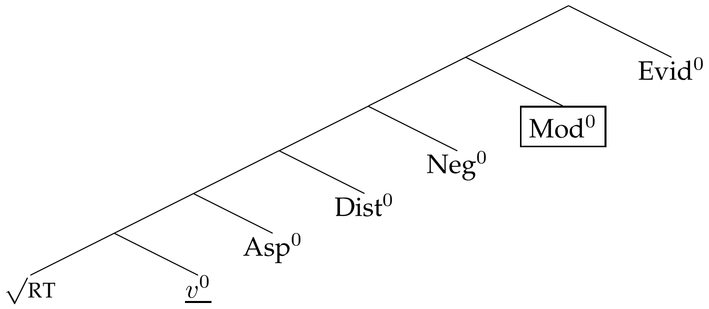
(30) Kiowa allomorphy (Bonet & Harbour 2012: 231)

a. héíb -e -gų̨ų̨ -mɔɔ -tɔɔ
enter TR DISTR NEG MOD
'will not bring in at different times/locations'

b. héíb -é -gų̨ų̨ -mɔɔ -t'ɔɔ
enter INTR DISTR NEG MOD
'will not com in at different times/locations'

Notice that the spell-out of Mod^0 alternates between /–tɔɔ/ and /-t'ɔɔ/, depending on the transitivity specification. This generalization was already pointed out in the standard grammar of Kiowa (Watkins 1984), while Bonet & Harbour (2012) tie its conditioning to v^0, in particular. The Mod^0-alternation is conditioned by *three heads* to the left. This is something that the Re-Labelling Hypothesis does not predict: Kiowa requires $\mathfrak{B}=\{S_L, S_L, S_L\}$, but the Re-Labelling Hypothesis allows at most $\mathfrak{B}=\{S_L, S_L\}$. This is a potential problem. To discuss more details of the pattern, first consider the Kiowa verbal complex, based on Adger et al. (2009):

(31) Kiowa verbal complex



It is to be noted that the alternation shown in (30) need not appear non-local all the time, i.e. Neg^0 and Dist^0 may not occur in the tree. Asp^0, on the other hand, is not an intervener, because in these constructions it is null. Agreement affixes are not shown in this tree, but they are generated as prefixes. The Mod^0-head typically expresses modality or also futurity (Watkins 1984: 147). The first question that this leads to is whether Mod^0 could perhaps be playing the role that T^0 plays in Indo-European languages, viz. be involved in agreement somehow, implying that it could possibly access information about argument structure. It is unclear that this is the case. In fact, Adger & Harbour (2007) argue that DP-arguments in Kiowa do not move out of vP at all, and that all agreement and Case checking is performed at AspP and below it. A possible avenue to explore is whether transitivity is encoded somewhere else in the Kiowa verbal complex, closer to Mod^0, but this is remains unclear for now.

However, the pattern of Mod^0-alternation in Kiowa has certain suspicious phonological properties. To begin with, it is very curious that the alternation between the two lexically fixed exponents of Mod^0 is so minimal: /-tɔɔ/ and /-t'ɔɔ/ only differ in the ejectivity of the onset consonant – the first is voiceless and the second is specified for [CG] ('constricted glottis'). Another phonological property of this pattern is that the ejectivity alternation seems to occur at the *right edge* of the prosodic word (PWd), i.e. Mod^0 seems to be the word-final morpheme most of the time.[12] Another correlating factor is that the suffixal domain in the Kiowa verbal complex reveals no underlyingly voiceless stops besides this modality suffix (Watkins 1984: 146-163). These morpho-phonological generalizations suggest that the following analysis should also be considered: some null head that is adjacent to v^0, spells out a *floating* [CG] feature (Wolf 2007; Akinlabi 2011) when v^0 is intransitive. This floating [CG] then docks on the modality exponent /-tɔɔ/, turning it to /-t'ɔɔ/:

[12] Other affixes can follow Mod^0, as in the tree in (31): these are either evidentials or different speech act operators, which are typically positioned very high in the tree. Miller (2015) discusses the formation of PWd's in Kiowa and notes that there seems to be no evidence for saying that the Evid^0-type affixes are in the same prosodic domain as the rest of the verbal suffixes or not. Interestingly, Watkins (1984: 236) notes that these Evid^0-type affixes are *clitics*. If these turn out to form their own spell-out/prosodic domains, this would not be surprising.

(32) Potential analysis of /-tɔɔ/ and /-t'ɔɔ/

√RT – v^0_{INTR} – ? – Dist^0 – Neg^0 – tɔɔ$_{\text{MOD}^0}$]$_{\text{PWD}}$

↓

[CG]

An analysis of this sort would avoid positing non-locality, and it would capture the morpho-phonological generalizations discussed above. The 'alternation in ejectivity' is automatically explained in this analysis, as is the right-edge position in the PWd: the docking of floating autosegments is typically driven by *alignment* with some edge of the PWd (Wolf 2007) – here, this is the right edge. In addition, most of the stops that occur in the suffixal domain are *voiced* stops (Watkins 1984: 146-183), which would make impossible landing sites for [CG], as the presence of voice is incompatible with ejectivity, making the /t/ in /-tɔɔ/ the only docking position.

A potential complication does occurs in this analysis. Kiowa contrasts voiceless, voiced, aspirated and ejective stops (Watkins 1984: 7). In coda position, all stops undergo devoicing (Watkins 1984: 51), in fact all laryngeal contrasts are neutralized there. Watkins (1984: 52-53) also notes that Kiowa exhibits a process of *progressive voice asismilation* in consonantal clusters. This effectively means that in [VC$_1$.C$_2$V], C$_2$ becomes devoiced because C$_1$ is voiceless in that position (codas are always voiceless), i.e. C$_1$ affects the voicing status of C$_2$. This causes a problem for the above analysis because in some cases, roots with stops in the final coda position, such as /√kóp-/ 'lie', can render the following voiced stop of a suffix voiceless: e.g. /√kóp-/ + /-gɔ/ (NEG) → [kópkɔ]. Since /√kóp-/ is an intransitive verb (Watkins 1984: 177), we would expect the [CG] floating feature to dock on the rightmost voiceless consonant, which here is the [k] in the negation [-kɔ], since Mod^0 need not have an overt spell-out here. We would thus predict *[kópk'ɔ]. In other words, voice assimilation would feed the docking of [CG]. This is a possible complication for the floating [CG] analysis.[13]

However, it should be noted that progressive assimilation has a precarious status in phonological theory. The expected natural direction of voice assimilation is typically *regressive*. It is conceivable to imagine that C$_2$ in [VC$_1$.C$_2$V] does not actually devoice but rather undergoes a type of *strengthening*, realized as aspiration, because it occurs adjacent to a *weak edge*, i.e. a devoiced consonant (p.c. Heather Goad). If the C$_2$ positions could be analyzed as aspirated, then it is possible that this 'strengthening' process would *not* feed [CG]-docking at all.[14] This is a plausible alternative to the non-local allomorphy account that must be considered.[15]

[13] Note that the voiceless consonants in roots never participate in this 'ejectivity alternation', but this is not unusual, since roots tend to require greater faithfulness than affixes, implying a high-ranked OT constraint along the lines of IDENT-ROOT[laryng] (Beckman 1998).

[14] This can be easily formalized in Optimality Theory (Prince & Smolensky 2004), by positing a high-ranked constraint that forces the 'strengthening' mentioned above, rejecting ejectivization. It does, however, need to be determined to what extent the phonetic facts in Kiowa align with this proposal.

[15] Voiced consonants may also undergo devoicing when they are preceded by a falling tone (Watkins 1984: 40). This can devoice certain verbal suffixes and render them voiceless (Watkins 1984: 154), which is another potential complication for the floating [CG]-analysis, but it is not clear that the tonal devoicing process would not perhaps interfere with the realization of ejectivity on the consonant that has undergone this devoicing process. If this is the case, then this is not an issue for the floating [CG]-analysis, since that would imply that tonal devoicing cannot feed ejectivization. One correlation which is striking, and may be pertinent, is that voiced and ejective stops have the same type of effect

The phonological facts of the Kiowa pattern raise some doubt as to whether non-local allomorphy is really at work here. In this section, we have suggested an alternative account that takes into consideration the phonological facts that surround this pattern. This alternative avoids the need for positing non-local allomorphy and so maintains the Re-Labelling Hypothesis. The pattern in Kiowa deserves the attention of an entire paper, but what we have here discussed illustrates that it need not be a counter-example to the Re-Labelling Hypothesis as stated in the previous sections.[16]

5 Competing Approaches to Non-Local Allomorphy

In section 2, we touched on how alternative approaches to non-local allomorphy (Merchant 2015; Moskal 2015a) compare to the one proposed here. In the present section, we give a detailed comparison between the predictions that these approaches makes and those that the proposal of this paper makes. We will point to several conceptual and empirical issues that these alternative approaches struggle with and demonstrate that the proposal advanced here is able to overcome them.

5.1 Span Hypothesis (Merchant 2015)

In this section, we discuss the Span Hypothesis (Merchant 2015) After describing the formal mechanisms involved, we also point out a prediction it makes that is not borne out empirically. To account for non-local patterns of allomorphy, Merchant (2015) proposes that the context of VI is not limited to single adjacent heads but can rather consist of whole 'spans' of heads. For instance, in cases of root suppletion, ⟨v^0, Voice^0, Asp^0⟩, ⟨v^0, Voice^0, Asp^0, T^0⟩, etc., are all possible spans and hence possible contexts for suppletion. The crucial data involve Greek suppletion conditioned by voice and aspect.

on the phonetic realization of tone in Kiowa, specifically F0, while voiceless stops pattern differently (Jurgensen 2011). This could mean that the tonal devoicing process would not feed ejectivization in Kiowa: this is because ejectivization would be banned in that position just like voicing is, as it has the same tone-related realization as ejectivization. But clearly more work is needed on this. It is, nevertheless, striking that all of the undesired potential 'candidates' for [CG] docking seem to arise through some independent devoicing process and are not underlyingly voiceless. Thus it is not at all clear that the floating [CG]-analysis is on the wrong track, lending credence to this view.

[16] If Kiowa were a clear-cut counter-example to the Re-Labelling Hypothesis, as discussed in sections 4.2–4.3, this may mean that we are dealing with a possible *inward-outward* asymmetry in allomorphy, which need not mean that the Re-Labelling Hypothesis is incorrect at all. To explain, this would mean that *inward affixal* allomorphy is perhaps less restricted than *outward* (root) allomorphy. The survey on root suppletion certainly suggests that outward allomorphy is quite restricted in terms of the trigger-target distance. It could then be the case that the 'extra' non-local cases of inward allomorphy do not follow from *re-labelling* at all: perhaps VI stores heads which it has previously processed in a 'computational stack', which may be accessed for purposes of allomorphic context – but a stack can only contain *inward* heads, by definition, so this would not affect the predictions for outward allomorphy. Since Kiowa seems to be only an apparent problem for a system with just Re-Labelling, we refrain from fleshing out this alternative.

(33) Active Imperfective | Non-Active Imperfective

√tro(ɣ) -o | √troɣ -omun
eat 1P.SG | eat 1P.SG

(34) Active Perfective | Non-Active Perfective

√fa -o | √faɣo -θ -ik -a
eat 1P.SG | eat NON-ACT PERF 1P.SG

The root exponent /tro(ɣ)-/ occurs across all imperfective paradigms, while two contextual suppletive forms occur in the perfective ones: /fa-/ with active voice and /faɣo-/ with non-active voice. This entials that reference to aspect is crucially needed, but Voice^0 intervenes between Asp^0 and the root. This means that VI must access the non-local Asp^0 head, as well as Voice^0, when inserting at the root. To capture this, Merchant (2015: 277-278) posits the following set of VI rules:

(35) $\sqrt{\text{EAT}} \longrightarrow$ fa / _____ $\langle \text{Voice}^0{}_{[+ACT]}, \text{Asp}^0{}_{[+PRF]} \rangle$
$\sqrt{\text{EAT}} \longrightarrow$ faɣo / _____ $\langle \text{Voice}^0{}_{[-ACT]}, \text{Asp}^0{}_{[+PRF]} \rangle$
$\sqrt{\text{EAT}} \longrightarrow$ tro(ɣ)

In other words, /fa-/ is inserted in active perfective paradigms, whereas /faɣo-/ is inserted in non-active perfective paradigms. Contextual reference to two heads is needed, which makes this a non-local pattern. Spans can consists of as many heads as we can find in an extended projection (Merchant 2015: 288).[17] This approach does constitute a non-local theory of VI, but to render it more constrained, Merchant (2015: 295) proposes that only heads that participate in the triggering of allomorphy can form a part of a span. This means that spans such as ⟨Voice^0, Asp$^0{}_{[+PRF]}$⟩, where the specification of Voice^0 does not matter, are impossible – Voice^0 *cannot play a vacuous role*. In other words, Voice^0 may only occur as 'Voice$^0{}_{[+\text{ACT}]}$' or 'Voice$^0{}_{[-\text{ACT}]}$' in a span, but never as just 'Voice^0'. This predicts that the following set of VI rules is impossible:

(36) $\sqrt{\text{EAT}} \longrightarrow X$ / _____ $\langle \text{Voice}^0, \text{Asp}^0{}_{[+PRF]} \rangle$
$\sqrt{\text{EAT}} \longrightarrow Y$

These two rules predict that X is inserted in the context of *all* perfective forms, be they active or non-active, and that Y is inserted in all other forms (elsewhere). This is a pattern that Merchant rules out, since the features of Voice^0, i.e. [+ACT] and [–ACT], do not participate in the Span. The role of Voice^0 is truly vacuous here.

This makes concrete predictions, but it raises some concern, as this constraint does not follow from any aspect of Merchant's formal system. It is a descriptive constraint that is not derived formally. However, if this constraint were not added, then the Span Hypothesis would allow any kind of non-local allomorphy within an extended projection, making few predictions about possible allomorphy patterns in natural language. This means that the adequacy of the Span Hypothesis as stated by Merchant (2015) is largely an empirical question.

[17] Phasal spell-out (Chomsky 2001) would help constrain allomorphic patterns somewhat, but this is beside the point since phases themselves may contain big amounts of complex morphology.

However, Merchant's constraint on vacuous spans can be falsified by certain non-local patterns, which suggests that there is a problem with models of allomorphy that solely rely on span-formation. One such a pattern is identified by Moskal & Smith (2016). In Tamil first and second person pronouns, K^{0}(ase) triggers suppletion across the #^{0}-head, realized as /-(n)ga(l)/ in the plural:

(37) Tamil 1P Tamil 2P (Moskal & Smith 2016: 306)

	SG	PL		SG	PL
NOM	naan-∅-∅	naan-ga-∅	NOM	nii-∅-∅	nii-nga-∅
OBL	en-∅-∅	en-ga-∅	OBL	on-∅-∅	on-ga-∅
DAT	en-∅-akku	en-gal-ukku	DAT	on-∅-akku	on-gal-ukku

Moskal & Smith (2016: 306) observe that the non-NOM features appear to trigger suppletion, which in the dative forms occurs across the overt number head /-gal/. The important observation is that the contextual span here needs to be ⟨#^{0}, K$^{0}_{\text{OBL/DAT}}$⟩, where #^{0} must be featurally underpecified because suppletion occurs in singular and plural forms. In other words, the featural make up of #^{0} does not participate in the triggering of allomorphy. One could assume that the nominative form is the suppletive one, but this would leave us with the same problem, as Moskal & Smith (2016) note, as VI would need to access the non-local null K^{0} across the #^{0} /-ga/. This appears to be a counter-example to Merchant's constraint on spans.

However, one must ask whether K$^{0}_{\text{NOM}}$ could not be fused with the #^{0} head. This would render the trigger fully local and would not violate Merchant's condition on span formation. But it does not seem that this is a strong case for a fused [#^{0}+K^{0}]-head. The K^{0}-head could very well be null, since the plural #^{0}-head has the same exponent across all the paradigms.[18] Furthermore, the exact same suffix expresses plurality in Tamil nouns (Schiffman 1999: 28). In other words, no clear independent evidence for such fusion is apparent here. These data render Merchant's key constraint questionable. If the constraint on vacuous heads cannot be maintained, then the Span hypothesis allows any kind of non-local allomorphy found within a single phase, making few predictions about allomorphic patterns in natural language.[19]

It must also be noted that Merchant's system of spans makes other concerning predictions. If all the heads within a span are involved in the allomorphy, it freely predicts suppletion to be sensitive to the very top of the extended projection, which can contain many overt heads. In section 3, we made the observation that the non-local patterns that have been uncovered are in fact much more conservative, imply-

[18] The plural exponent is completely regular across all the forms, as the alternations that can be observed in (9) are phonological in nature: Schiffman (1999: 28) explains that the underlying forms of the plural marker is /-ngal/. He notes that the /l/ deletes in word-final position but surfaces when it precedes a vowel. He also explains that /n/ deletes when preceded by a nasal.

[19] At least one more pattern is such that it exhibits interveners that do not seem to participate in the allomorphy at all. For instance, in Basque the comparative head Cmpr^{0} triggers root suppletion in adjectives, but a diminutive morpheme may intervene between them with only certain roots, as in $\sqrt{\text{RT}}$-(Dim^{0})-Cmpr^{0} (Bobaljik 2012: 156-158). Here, Dim^{0} plays a vacuous role. A possible pattern of the same type may be found in Kiowa (Tanoan). In Kiowa, the modality head Mod^{0} undergoes allomorphy triggered by the transitivity specification of v^{0}. However, a negation and distributivity suffix can optionally intervene, but do not block the process, as in $\sqrt{\text{RT}}$-v^{0}-(Neg^{0})-(Distr^{0})-Mod^{0} (Bonet & Harbour 2012: 231). Here Neg^{0} and Distr^{0} also seem to be playing purely vacuous roles. Though see section 4.5 on Kiowa. The pattern is not fully clear.

ing that contextual 'spans' are never that big. Specifically, we saw that non-locality only occurs for two heads either to the left or right, from the target of insertion.

Both of the problems with the Span Hypothesis can, however, be overcome by the model of allomorphy that we proposed in section 4. Let us first touch on the issue of *vacuous heads in spans*. In our model, vacuous heads can be easily specified as context in order to derive patterns such as suppletion in Tamil pronouns: the operation SCAN would need to identify $\#^0$ as well as K^0 and store both of them in the buffer, s.t. $\mathfrak{B}=\{\#^0, K^0\}$, but nothing at all is preventing a VI-rule specifying $\{\#^0, K^0{}_{DAT}\}$ as context, where $\#^0$ is featurally underspecified. The buffer, of course, contains the heads *and* their featural contents, but a VI-rule can then either specify just the heads without their features, or also with their features. This is so because the buffer does not regulate the lexical content of the rule, but only what part of the syntactic structure the rule can 'see'.

We now turn to the *conservative distance* involved in non-local allomorphy. The Span Hypothesis cannot predict that only two heads in one direction can be involved in non-local patterns. The model we advanced in section 4, on the other hand, directly predicts this, as expressed by the *Re-Labelling Hypothesis*: the buffer contains maximally two slots ($\mathfrak{B}=\{S_L\ S_R\}$), which may be re-labelled so that two heads can be accessed either in the left, or right direction.

Another issue with the Span Hypothesis is that it expresses *no formal difference* between *local* and *non-local patterns*. The distinction between these two pattern types needs to be expressed if a model of allomorphy is to capture their different status in the grammar (the markedness distinction), as discussed in sections 2 and 3. In section 4, we showed that our model can express this markedness scale through different degrees of *computational complexity*: computing non-local patterns is more complex than computing local patterns. Since the Span Hypothesis makes no distinction between local and non-local patterns it cannot derive *local blocking* effects, as briefly noted in section 2. Consider the Slovenian blocking repeated below:

(38) Local blocking in Slovenian (√človek- ~ √ljudj-)

	SG	DU	PL	
N	√človek-∅	√človek-a	√$\boxed{\text{ljudj}}$-e	'man'
N+DIM	√člověč-ek-∅	√člověč-k-a	√člověč-k-i	
	√RT– (DIM^0–) $\#^0$	√RT– (DIM^0–) $\#^0$	√RT– (DIM^0–) $\#^0$	

The adjacent plural-specified $\#^0$ triggers suppletion, but this is blocked if the diminutive head /-(e)k/ intervenes. The only way that the Span Hypothesis can derive this pattern is by positing a VI-rule, such as Rule 1 below:

(39) Rule 1: √MAN ↔ *ljudj-* /_____⟨ [PL] ⟩ *status of rule?* ⇝ **local**

Rule 2: √MAN ↔ *ljudj-* /_____⟨ Dim^0, [PL] ⟩ *status of rule?* ⇝ **non-local**

Rule 1 cannot apply if the context of the root includes more than just a plural-specified $\#^0$-head, which means that it will not occur when a diminutive suffix is also present. However, the fact that Rule 2 does not occur in Slovenian is just a lexical accident under this view. In this way, the Span Hypothesis cannot interpret the blocking pattern in Slovenian in a principled way. The model we proposed, on the other hand, handles the blocking pattern as the default reaction of the grammar: the default buffer ($\mathfrak{B}$={S_L, S_R}) can only identify one head in the right direction, which in the case of diminutive formations means it can only identify Dim^0. The consequence of this is that blocking will occur regardless of what the lexical content of the rule is, i.e. neither Rule 1 or Rule 2 could apply in diminutive formations, which makes for a principled account of blocking.

In sum, we have demonstrated that the Span Hypothesis faces two important challenges: it cannot derive the conservative natural of distance involved in non-local allomorphy and it expresses no distinction between local and non-local allomorphy. One of the consequences of the latter issue is that it fails to provide principled accounts of local blocking effects. The model proposed in this paper can, in turn, derive both of these generalizations.

5.2 Accessibility Domain (Moskal 2015a)

In this section, we discuss another approach to non-local allomorphy, developed by Moskal (2015a,b) and Moskal & Smith (2016), which crucially completely dispenses with *adjacency* as a locality condition on allomorphy. We show, in several different steps, that this leads to many problems for this approach, all of which can be solved if the proposal advanced in section 4 is adopted.

Moskal (2015a,b) and Moskal & Smith (2016) develop a conceptually different approach, where VI itself is not constrained in terms of locality, but the locality stems from the morphosyntactic configuration that VI accesses. Specifically, they argue that the locality on allomorphy stems from the 'Accessibility Domain' ($\mathcal{AD}$) of the root. The $\mathcal{AD}$ extends from the root up to the first functional head which is immediately above the categorial head: for instance, in $[\underline{[[[\sqrt{\text{RT}}\ ]\ x^0\ ]\ \text{A}^0\ ]}\ \text{B}^0\ ]$, where x^0 is a categorial head and A^0 and B^0 functional heads, x^0 and A^0 can trigger root suppletion, but B^0 cannot. The $\mathcal{AD}$ stems from the conception of Phase Theory, as in Bobaljik & Wurmbrand (2005), den Dikken (2007) and Bošković (2014), where the computational system needs to check the status of the next head in the hierarchical structure immediately above the categorial head. This makes a prediction that Moskal (2015a,b) and Moskal & Smith (2016) exploit: K^0(ase)-triggered root suppletion is possible in pronouns, but not in nouns:

(40) $\mathcal{AD}$ in nouns: $[\boxed{[[[\sqrt{\text{RT}}\ ]\ n^0\ ]\ \#^0\ ]}\ \text{K}^0\ ]$

No $\mathcal{AD}$ in pronouns: $[[[[\text{D}^0\ ]\ \#^0\ ]\ \text{K}^0\ ]\ ...\ ]$

K^0 is outside of the $\mathcal{AD}$ in nouns, but not in pronouns. Pronouns have no $\mathcal{AD}$ since they contain no categorial head. What must be noted is that within the $\mathcal{AD}$ itself, and outside of it, VI is subject to no locality condition, as Moskal (2015a) and Moskal & Smith (2016) explicitly argue against *adjacency* as a locality constraint on allomorphy. They propose a system where non-local allomorphy is freely available, as long as it occurs within or outside of the $\mathcal{AD}$.

There are several conceptually and empirically problematic aspects of the $\mathcal{AD}$ Hypothesis, which we discuss in the following paragraphs below. The $\mathcal{AD}$ Hypothesis runs into several empirical problems, which Moskal (2015a) solves by making use of two operations: *Readjustment Rules* and the *Pruning* of null heads. Below we discuss such examples and point out that they in fact pose a serious problem for the $\mathcal{AD}$ Hypothesis. We show that using Readjustment Rules and Pruning in these cases introduces too much power into the theory, preventing the $\mathcal{AD}$ from making falsifiable predictions about allomorphic patterns in natural language.

K^0-triggered root suppletion should not exist in nouns, according to the $\mathcal{AD}$ Hypothesis. However, the observation has been made that it indeed does exist, with examples stemming from Lezgian (Haspelmath 1993: 80), Latin and Scottish Gaelic (Moskal 2015a: 38-41) among others. (Moskal 2015a: 27-41) proposes to explain such cases in terms of Readjustment Rules and null-head deletion termed 'Pruning' (Embick 2010). First, we consider the Latin and Scottish Gaelic cases:

(41) Scottish Gaelic 'wife'

	SG	PL
NOM	bean	mnatha
GEN	mná	ban

Latin 'Jupiter' (Moskal 2015a: 38-41)

	SG	PL
NOM	Juppit-er	Jov-ēs
GEN	Jov-is	Jov-um

Moskal (2015a: 38-41) explains these cases of root suppletion as instances of Readjustment Rule application. A set of such rules is triggered by the roots in question in Scottish Gaelic and Latin. These rules appropriately modify the root's exponent. However, the use of Readjustment Rules is poorly motivated here. Firstly, a possible argument for readjustment are larger patterns in allomorphy, such as the English /ɪ/~/æ/ alternation (e.g. *drink~drank*, *swim~swam*, ...), where VI-induced suppletion could miss a generalization. But in Latin and Scottish Gaelic, the cases of suppletion are isolated cases, with no synchronic motivation for such readjustment in the rest of the systems. Moskal does note that the specific application of readjustment she posits is not available elsewhere in the respective sychronic systems, but this still renders the use of readjustment suspect here. Secondly, if Readjustment Rules perform allomorphy, then we expect them to be subject to the same or equivalent locality conditions. Notice that this is an issue here, as the Readjustment Rules violate the $\mathcal{AD}$ and no other locality condition is available within a single phase. This is an issue as, with the introduction of unconstrained Readjustment Rules, the theory ceases to make falsifiable predictions about allomorphy in natural language.

It should be noted that the status of Readjustment Rules is somewhat questionable. An increasing trend in Distributed Morphology rejects Readjustment Rules on conceptual grounds (Siddiqi 2006, 2009; Bye & Svenonius 2012; Gribanova 2015), which is also an assumption that our proposal will make (see section 6). Merchant (2015: 282) notes that Readjustment Rules have become a 'get-out-of-counterexample-free' card, while Bermúdez-Otero (2012: 80) determines that they 'utterly destroy the empirical content of morphological and phonological hypotheses'. What is important to acknowledge is that, by rejecting of Readjustment Rules, K^0-triggered suppletion in nouns can no longer be derived in Moskal's approach. The rejection of Readjustment Rules hence immediately falsifies the $\mathcal{AD}$.

Now we turn to the Lezgian pattern, where two roots, $\sqrt{\text{WATER}}$ and $\sqrt{\text{SON}}$, supplete for the singular oblique case. We give the paradigm of $\sqrt{\text{WATER}}$:

(42) Lezgian (Haspelmath 1993: 80), (Moskal 2015a: 28)

	SG	PL
ABS	jad-∅-∅	jat-ar-∅
OBL	c-∅-i	jat-ar-i

In this case, Moskal posits the null-head Pruning operation (Embick 2010): specifically, she proposes that the #^{0}-head is Pruned in the context of singular absolutive case, which brings K^0 into the $\mathcal{AD}$, enabling K^0 to condition suppletion. This maintains the predictions cast by the $\mathcal{AD}$. However, this analysis is also problematic because Pruning is not a generalized operation for Moskal. In fact, Pruning must be prohibited most of the time in order to maintain the $\mathcal{AD}$, by keeping K^0 unaccessible: in Moskal's approach, null heads are required to have the same status as overt heads, since adjacency plays no role there. In addition, there is no synchronic motivation for such Pruning in Lezgian, which renders the analysis stipulative.[20,21] It should also be noted that such use of Pruning involves derivational look-ahead: the null #^{0}-head needs to be deleted before VI has inserted the root, in order for K^0 to condition the insertion of the root. As defined in Embick (2010), Pruning is an operation that occurs after VI has attempted to insert at the head that is to be pruned, which means that the Pruning posited for Lezgian is not the typical Pruning operation often invoked in analyses of allomorphy.

In sum, it seems that, in a maximally restrictive theory that disallows the discussed use of Readjustment and Pruning, the $\mathcal{AD}$ is inevitably falsified. In the following section, 5.3, we demonstrate that the $\mathcal{AD}$ Hypothesis can also be violated in the verbal domain, while in sections 5.4 and 5.5 we also show that the unbounded nature of VI outside of the $\mathcal{AD}$, or in pronouns generally, misses generalizations on allomorphy.

5.3 A Counter-Example to the $\mathcal{AD}$ in the Verbal Domain

In this section, we discuss another counter-example to the $\mathcal{AD}$, but this time in the verbal domain. A case of non-local root allomorphy from Slovenian, South Slavic, seems to be triggered by a head that is located too far out of the $\mathcal{AD}$. We first present the basic pattern, taken from Božič (2016), where simple verbs show one type of root exponent, but participles formed with /-l/ show another. Theme vowels intervene between the root and /-l/. This pattern consists of a very small number of roots, of which we give three here:

(43) Slovenian verbs and participles (Božič 2016: 139)

ROOT	V:1P.SG	PTC:F.SG	PASS.PTC	INF	
√žanj-	žanj-e-m	ž-e-l-a	ž-e-t	ž-e-ti	'reap'
√koln-	koln-e-m	kl-e-l-a	kl-e-t	kl-e-ti	'swear'
√boj-	boj-i-m	b-a-l-a	b-a-n	b-a-ti	'fear'

[20] Moskal (2015a: 31) resorts to diahronic evidence to support this analysis. The two nouns were *pluralia tantum* historically, probably lacking a #^{0}-head. However, synchronically they are countable nouns, and it seems that suppletion was used to 'fill' the gaps in the paradigm, which is how suppletion very often emerges (Ronneberger-Sibold 1988; Mel'čuk 2000: 520). There is no synchronic reason for positing an exceptional Pruning rule here.

[21] Similar criticism has been expressed in a blog post by Haspelmath (2016).

The more 'impoverished' verbal exponents occur in the context of participles formed with /-l/, passive participles and infinitives. Bozic (2016: 140) notes that these might belong to the same head, call it Ptc^0, since they are in perfect complementary distribution and they all represent non-tensed formations. Note that T^0 spells out an auxiliary when participles are formed, but it is part of the complex head in simple verbs, which are tensed. The 'alternations' involved between the different root exponents cannot be phonologically conditioned in Slovenian (Božič 2016: 140) and must therefore be cases of contextual allomorphy. This raises the question of which of the exponents is the elsewhere item: for √REAP, is it /žanj-/ or /ž-/? It would seem that /žanj-/ is the elsewhere exponent, since it can occur in nominalizations with the suffix /-its/, which often forms nouns directly from roots in Slovenian (Marvin 2002: 117-119), i.e. /žanj-its-a/ 'reaper' (F.SG.) (Božič 2016: 139). This form also does not indicate that any verbal material is present between the root and the nominalizer. This suggests that the contexts of /žanj-/ do not form a natural class, but those of /ž-/ do: they are all non-tensed verbal formations, conditioned by the same head, viz. Ptc^0. This means that Ptc^0 must be the trigger. On the assumption that Readjustment Rules are not available, this allomorphy must be an instance of VI-allomorphy.

In what follows, we present novel data that reveal a refined layer of morphosyntactic heads between the root and Ptc^0 in Slovenian. Consider the following constructions given below:[22]

(44) Verbal morphology between √RT and Ptc^0

V_{IMPF}	V_{TEL}	$V_{SCD.IMPF}$	ADJ	
√zd-i-m	do-√zd-i-m	do-√zd-ev-a-m	do-√zd-ev-n-	'appear'
√grej-e-m	o-√grej-e-m	o-√gr-ev-a-m	o-√gr-ev-n-	'heat'
√zn-a-m	za-√zn-a-m	za-√zn-av-a-m	za-√zn-av-n-	'know'

These data show the different aspectual morphology that is located between the root and Ptc^0; a Ptc^0 suffix is not itself shown above, but it would simply follow the stems of the constructions in question, e.g. *do-√zd-ev-a-l-*, *o-√gr-ev-a-l-*, etc. The simplest (and default) constructions are imperfective. Telic verbs are formed with a prefix, and such constructions may then be given an iterative meaning by forming secondary imperfectives with /-ev/. These may, furthermore, form de-verbal adjectives. These data show that a theme vowel is found in the Slovenian verbal complex, as shown below. /-a/ always occurs with secondary imperfectives, but it may have other allomorphs in default aspect constructions.

(45)

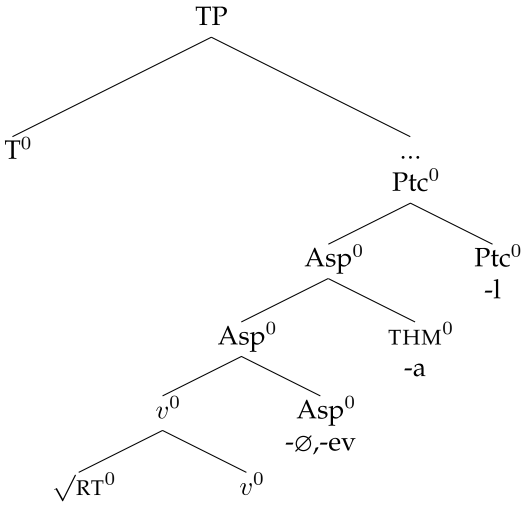
do-	√zd-	ev-	a-	vs.	za-	√zn-	av-	a-
PFX	appear	SCDN.IMP	THM		PFX	know	SCDN.IMP	THM

This is very much like the theme vowels in Russian, as discussed by Gribanova (2015) and Svenonius (2004a,b). The theme vowel is attached to Asp^0, exactly as Gribanova (2015) proposes for Russian.[23]

[22] Toporišič (2000) and Herrity (2000) can be consulted for the basic descriptions of these constructions. I would like to thank a native speaker of Slovenian for sharing her intuitions with me on the data that follow.

[23] It is tempting to decompose the secondary imperfective /-ev/~/-av/ into two separate suffixes. There is some plausibility to this since /-e/ and /-a/ also exist as theme vowels in Slovenian. How-

(46) The positions of THM^0, v^0 and Asp^0



Asp^0 is likely null in imperfectives and plain telic forms, but in the secondary imperfectives it spells out /-ev/. Ptc^0 spells out /-l/, and T^0 spells out an auxiliary verb in such cases. But in the case of simple verbs, T^0 is either null or fused with the agreement suffixes.

This fits the standard Slavic 'template' of verbal morphology. However, we have not yet presented a concrete diagnostic that shows that this layer of heads between the root and Ptc^0 is indeed located above v^0 in Slovenian. This is important to show because the position of v^0 directly delineates the $\mathcal{AD}$. We can determine this with the position of a verbalizing suffix, viz. /-ir/, which must be the spell-out of v^0:

(47) Formations with the verbalizer /-ir/

	NOUN	VERB	ADJ	
√doz-	doz-a	→ doz-ir-a-m	→ doz-ir-n-a	'dose'
√lak-	lak-∅	→ lak-ir-a-m	→ lak-ir-n-a	'polish'

/-ir/ is a productive verbalizer that forms verbs from noun stems, as illustrated above. Its semantics always express an action derived from the noun, e.g. *doz-ir-a-m* 'administer a dose' is derived from *doz-a* 'dose'. It is possible to form adjectives from verbs derived with /-ir/, as shown in the third column above. In such adjectival formations, the theme vowel does not appear. These forms illustrate that the /a/ in *doz-ir-a-m* is truly a separate theme suffix and not part of the verbalizer /-ir/. If the layer of aspectual morphology described in the previous paragraph can be shown to be located above /-ir/, this must entail that it occurs above v^0. This can indeed be demonstrated as some forms do inflect for this aspectual morphology. Consider the following:

(48) pre- √doz -ir -av -a → pre- √doz -ir -av -a -l
PFX dose V SCD.IMP THM PFX dose V SCD.IMP THM PTC

(49) s- √keš -ir -av -a → s- √keš -ir -av -a -l
PFX cash V SCD.IP THM PFX cash V SCD.IP THM PTC

ever, the secondary imperfective can sometimes also be spelt out as /-ov/, and /-o/ does not exist as a theme vowel, or any other independent suffix in the verbal domain, which is why we treat /-ev/ as a single morpheme, similar to Dickey (2003).

pre-√doz-ir-av-a- in (48) denotes 'to overdose', while *s-√keš-ir-av-a-* in (49) 'spend money'. Notice that both these forms can be turned into participles, by simply adding /-l/ to the stem.[24] This demonstrates that the layer of complex aspectual morphology does indeed occur above v^0.

Turning back to the non-local pattern in Slovenian in (43), we noted above that it is the Ptc^0 that triggers the root allomorphy in these cases. This directly violates the $\mathcal{AD}$, as proposed by Moskal (2015a) and Moskal & Smith (2016):

(50) $\mathcal{AD}$ and allomorphy in Slovenian

$$[[[\boxed{[[\sqrt{\text{RT}}\ ]\ v^0]\ \text{Asp}^0]}\ \text{THM}^0]\ \text{Ptc}^0]$$

Note that as much as two heads intervene between the categorial head v^0 and Ptc^0: Asp^0 and the theme vowel. The $\mathcal{AD}$ still predicts that root suppletion in this case can only be triggered by heads up to Asp^0, but here it is clear that an overt head *above Asp*^{0} triggers suppletion. Recall that the $\mathcal{AD}$ Hypothesis can involve no generalized notion of Pruning or Fusion of null heads, as then most of the generalizations it offers on root suppletion could be systematically rendered invalid. In sum, these Slovenian data falsify the $\mathcal{AD}$ in the verbal domain.

The $\mathcal{AD}$ can thus be falsified in several ways: by K^0-triggered suppletion in Latin, Scottish Gaelic and Lezgian (nominal domain), and also by Ptc^0-triggered suppletion in Slovenian (verbal domain). Since the $\mathcal{AD}$ Hypothesis makes incorrect predictions, we conclude the locality needs to stem from something other than the morphosyntactic configuration. We will take this as an indication that any locality constraint on allomorphy needs to stem from the computational properties of VI, as we argued in section 4.

5.4 $\mathcal{AD}$, Local Blocking and Fusion-Allomorphy Conspiracies

In terms of local blocking effects, the $\mathcal{AD}$ Hypothesis, as advocated in Moskal (2015a) and Moskal & Smith (2016), is able to make certain succesful predictions on local blocking effects. However, those predictions appear to be an accidental epiphenomenon of the $\mathcal{AD}$. Consider again the Slovenian pattern of blocking:

(51) Local blocking in Slovenian (√človek- ~ √ljudj-)

	SG	DU	PL	
N	√človek-∅	√človek-a	√$\boxed{\text{ljudj}}$-e	'man'
N+DIM	√človeč-ek-∅	√človeč-k-a	√človeč-k-i	
	√RT– (DIM^0–) #^{0}	√RT– (DIM^0–) #^{0}	√RT– (DIM^0–) #^{0}	

Moskal (2015a: 66-69) explains that the kinds of blocking seen in Slovenian and many other Slavic languages can be explained adequately by the $\mathcal{AD}$ Hypothesis: the plural-specified #^{0} head is within the $\mathcal{AD}$ (as it is the first node immediately above n^0), which is why it can trigger suppletion. However, as soon as Dim^0 is attached to n^0, #^{0} ceases being the immediate node above n^0 – in other words, the

[24] Some speakers may find the form in (49) odd, likely because the SCD.IMPF is not very productive in constructions with /-ir/.

the attachment of Dim^0 pushes #^{0} outside of the $\mathcal{AD}$. Since #^{0} is no longer within the $\mathcal{AD}$, it fails to trigger suppletion.

In this sense, blocking is predicted to occur, by the $\mathcal{AD}$ Hypothesis, only if the trigger of allomorphy is pushed outside of the $\mathcal{AD}$. This can be characterized in the following way:

(52) Attachment of a new head to $\boxed{\sqrt{\text{RT}}\text{–n}^0\text{–}\#^0}\text{–K}^0{}_1\text{–K}^0{}_2$

$\boxed{\sqrt{\text{RT}}\text{–n}^0\text{–}\#^0}\text{–}\mathbf{Dim}^0\text{–K}^0{}_1\text{–K}^0{}_2 \rightsquigarrow$ no blocking

$\boxed{\sqrt{\text{RT}}\text{–n}^0\text{–}\#^0}\text{–K}^0{}_1\text{–}\mathbf{Dim}^0\text{–K}^0{}_2 \rightsquigarrow$ no blocking

$\boxed{\sqrt{\text{RT}}\text{–n}^0\text{–}\mathbf{Dim}^0}\text{–}\#^0\text{–K}^0{}_1\text{–K}^0{}_2 \rightsquigarrow$ actual blocking

Attaching a new head outside of the $\mathcal{AD}$ cannot change the $\mathcal{AD}$ and, therefore, no blocking is predicted to occur in such cases. If you attach Dim^0 between K^0_1 and K^0_2, any allomorphic relation between K^0_1 and K^0_2 will not be subject to blocking. Only if you attach Dim^0 between n^0 and #^{0}, then allomorphic relations between them, or between $\sqrt{\text{RT}}$ and #^{0} will be blocked. In sum, actual blocking is only predicted to occur if it tampers with the $\mathcal{AD}$ *and* if the allomorphic relation is at the edge of the $\mathcal{AD}$. However, all other cases of local blocking need to be treated as lexical accidents, exactly as in the Span Hypothesis (Merchant 2015), as discussed in section 5.1.

A key prediction of the $\mathcal{AD}$ Hypothesis, then, is that there should be no locality effects at all in the absence of the $\mathcal{AD}$. In other words, in purely functional words, such as pronouns, which contain no categorizer, we expect to see no locality effects. However, this prediction also does not seem to be borne out. In section 2.2, we discussed a phenomenon termed a *Fusion-Allomorphy Conspiracy*, which occurs in the pronominal system of Georgian, and actually forms a generalization across Kartvelian languages. Consider the Georgian paradigm repeated from section 2.2:

(53) Georgian 3P

	SG	PL
NOM	eg	ege-n-i / ege-eb-i
DAT	maga-s(a)	maga-t(a)
ERG	maga-n	maga-t(a)
GEN	mag-is(a)	maga-t(a)
INST	mag-it(a)	maga-t(a)
ADVB	maga-d(a)	maga-t(a)
	$\sqrt{\text{D}^0}$– #^{0}– K^0	

(54) *Generalization in Kartvelian*
Suppletion in pronouns is correlated with fusing #^{0} and K^0.

As stated in section 2.2, the non-NOM cases undergo suppletion, which is correlated with the fusion of #^{0} and K^0 into a single adjacent head. This correlation seems to hold quite generally across Kartvelian languages, as stated in the generalization above. See section 2.2 for more in depth discussion of this pattern. In an approach that encodes some bias for adjacency, like the approach we proposed in section 4, the Generalization in Kartvelian can be easily captured: suppletion is analyzed as being triggered by the contents of K^0, but only when it is rendered adjacent by fusion. In our model, this naturally follows from the assumption that the *unmarked* and hence *default* specification of the contextual buffer contains one *slot* labelled for identifying left-adjacent context and the other *slot* labelled for identifying right-adjacent context, i.e. $\mathfrak{B}$={S_L, S_R}. Since by default, there is only one slot available for identifying right-adjacent context, only one immediately adjacent head in the *right* direction can be stored in $\mathfrak{B}$. In the instance of the nominative case, this would be only #^{0}, but *not* also K^0, and in the instance of the non-NOM case forms, this would be the fused [#^{0}+K^0]-head. If K^0 is triggering suppletion, then this analysis can readily capture the Generalization in Kartvelian: we predict to suppletion in the presence of fusion, but not in its absence.

The $\mathcal{AD}$ Hypothesis, as advanced by Moskal (2015a), on the other hand, cannot capture the Generalization in Kartvelian, crucially because it predicts that no *locality effects* in allomorphy will be found in structures without a categorizer, such as pronouns in Kartvelian languages.

(55) Rule 1: D^0 ↔ *maga-* /____⟨ [#^{0}+K^0] ⟩ $\overset{\textit{status of rule?}}{\rightsquigarrow}$ **local**

Rule 2: D^0 ↔ *maga-* /____⟨ [#^{0}], [K^0] ⟩ $\overset{\textit{status of rule?}}{\rightsquigarrow}$ **non-local**

To encode the suppletion pattern in Georgian, the $\mathcal{AD}$ Hypothesis needs to posit a rule such as Rule 1 above. However, it cannot express that Rule 2 is somehow less likely to occur, or that it is impossible as a default setting. Rule 2 is predicted to occur just as likely as Rule 1 under this view of allomorphic locality. This means that the $\mathcal{AD}$ Hypothesis misses the Generalization in Kartvelian, as fusion cannot be analyzed as a factor that facilitates suppletion in any way. In our proposal, advanced in section 4, on the other hand, it does not matter which of the rules above is specified in the lexicon, as Rule 2 can never apply in a default setting of the buffer, since the buffer can only consider purely adjacent context by default.

6 Conclusion

This paper has considered instances of non-local contextual allomorphy of different types. We have provided a cross-linguistic survey of locality in allomorphy, which uncovers new generalizations on allomorphic locality. The crucial generalization is that local and non-local allomorphy occupy two different ends of a *markedness* scale: while local allomorphy seems to represent the unmarked pattern, non-local allomorphy is much more marked and exceptional in nature. We have also observed that non-local allomorphy tends to be non-local for only two heads in one direction and no more. We gave two additional types of arguments for the claim that *local* allomorphy represents the *default* of any grammar: we discussed instances of *local blocking effects* and *fusion-allomorphy conspiracies*, which cannot be derived in a principled way if strict locality as adjacency does not exist in some way in the grammar. We showed that other approaches to non-local allomorphy, i.e. Merchant (2015) and Moskal (2015a), cannot offer principled solutions to such phenomena precisely because they make no distinction between local and non-local allomorphy.

Based on this series of generalizations, we argued for *new model of allomorphy*, one that crucially incorporates some notion of strict locality, but at the same time a mechanism that can derive non-local patterns as systematic instances of exceptions. The formal properties of this system involve a search procedure that identifies the context of insertion at the PF-interface. Crucially, this procedure is constrained by an ECONOMY CONDITION, which encodes the strict locality that any grammar needs to include. This economy condition can, however, be violated in a principled way, which gives rise to the non-local patterns that we have observed. This system successfully derives the markedness distinction in allomorphic locality. In addition, its formal aspects, discussed in sections 4.2–4.3, also capture the generalization that non-locality can never consider more than two heads in one direction.

Another positive aspect of the proposed system is that the economy condition it uses can be shown to follow from general principles of Minimal Computation (Chomsky 2013), as discussed in section 4.2. This allows hypothesizing about the general properties of the PF-interface: the principles proposed in this paper are in broad agreement with the basic premises of the Minimalist Program, viz. that interface operations can be reduced to general principles of computation. This implies that the behaviour of contextual allomorphy in natural language stems from 'third factor' design properties (Chomsky 2005), which is an interesting result.

References

Adger, D. & Harbour, D. (2007). Syntax and syncretisms of the Person Case Constraint. *Syntax*, *10*, 2–37.

Adger, D., Harbour, D., & Watkins, L. J. (2009). *Mirrors and Microparameters: Phrase structure beyond free word order*. Cambridge: Cambridge University Press.

Aikhenvald, A. Y. (2003). *A grammar of Tariana, from northwest Amazonia*. Cambridge: Cambridge University Press.

Akinlabi, A. (2011). Featural affixes. In Marc van Oostendorp, Colin J. Ewen, E. V. H. (Ed.), *The Blackwell Companion to Phonology*, (pp. 1945–1971). John Wiley and Sons.

Allen, M. (1979). *Morphological investigations*. University of Connecticut.

Arregi, K. & Nevins, A. (2012). *Morphotactics: Basque Auxiliaries and the Structure of Spellout*. Studies in Natrual Language and Linguistic Theory 86. New York: Springer.

Beckman, J. N. (1998). *Positional Faithfulness*. PhD thesis, University of Massachusetts Amherst, Amherst, MA.

Bermúdez-Otero, R. (2012). The architecture of grammar and the division of labour in exponence. In Trommer, J. (Ed.), *The morphology and phonology of exponence*, (pp. 8–83)., Oxford. Oxford University Press.

Bobaljik, J. (2000). The ins and outs of contextual allomorphy. In Grohman, K. K. & Struijke, C. (Eds.), *University of Maryland Working Papers in Linguistics 10*, (pp. 35–71).

Bobaljik, J. (2012). *Universals in Comparative Morphology: Suppletion, superlatives and the structure of words*. Cambridge, MA: MIT Press.

Bobaljik, J. & Wurmbrand, S. (2002). Notes on Itelmen Agreement. *Linguistic Discovery*, (1.1).

Bobaljik, J. & Wurmbrand, S. (2005). The domain of agreement. *Natural Language and Linguistic Theory*, *23*(4), 809–865.

Bonet, E. & Harbour, D. (2012). Contextual allomorphy. In Trommer, J. (Ed.), *The morphology and phonology of exponence*, (pp. 195–235)., Oxford. Oxford University Press.

Bošković, Z. (2014). Now I'm a phase, now I'm not a phase: On the variability of phase with extraction and ellipsis. *Linguistic Inquiry*, *45*(1), 27–89.

Božič, J. (2015). Spell-Out of Phonological Domains: The Case of Slovenian. Master's thesis, University of British Columbia, Vancouver, BC.

Božič, J. (2016). Locality of exponence in Distributed Morphology: Root Suppletion in Slovenian. In Hammerly, C. & Prickett, B. (Eds.), *Proceedings of NELS 46*, (pp. 137–146)., Amherst, MA. GLSA.

Brown, D., Chumakina, M., Corbett, G. G., & Hippisley, A. (2003). Surrey Suppletion Database. *University of Surrey*. http://dx.doi.org/10.15126/smg.12/1.

Bye, P. & Svenonius, P. (2012). Non-concatenative morphology as epiphenomenon. In Trommer, J. (Ed.), *The morphology and phonology of exponence*, (pp. 427–495)., Oxford. Oxford University Press.

Caha, P. (2009). *The nanosyntax of case*. PhD thesis, University of Tromso, Tromso.

Carstairs-McCarthy, A. (1992). *Current Morphology*. Routledge.

Chomsky, N. (2000). Minimalist inquiries: The framework. In Martin, R., Michaels, D., & Uriagereka, J. (Eds.), *Step by Step: Minimalist Essays in Honor of Howard Lasnik*, (pp. 89–155)., Cambridge, MA. MIT Press.

Chomsky, N. (2001). Derivation by Phase. In Kenstowicz, M. (Ed.), *Ken Hale. A Life in Language*, (pp. 1–52)., Cambridge, MA. MIT Press.

Chomsky, N. (2005). Three factors in language design. *Linguistic Inquiry*, *36*(1), 1–22.

Chomsky, N. (2013). Problems of projections. *Lingua*, (130), 33–49.

Cinque, G. (1999). *Adverbs and Functional Heads – A Crosslinguistic Perspective*. Oxford: Oxford University Press.

den Dikken, M. (2007). *Relators and Linkers*. Cambridge, MA: MIT Press.

Dickey, S. M. (2003). Verbal aspect in slovene. In Stolz, T. (Ed.), *STUF - Language Typology and Universals*, volume 56, (pp. 182–207).

Embick, D. (2010). *Localism versus globalism in morphology and phonology*. Cambridge, MA: MIT Press.

Gribanova, V. (2015). Exponence and morphosyntactically triggered phonological processes in the Russian verbal complex. *Journal of Linguistics*, 519–561.

Halle, M. & Marantz, A. (1993). Distributed Morphology and the Pieces of Inflection. In *The View from Building 20*, MIT Working Papers in Linguistics, (pp. 111–176)., Cambridge, MA. MIT Press.

Harbour, D. (2008). Discontinuous agreement and the syntax-morphology interface. In Harbour, D., Adger, D., & Béjar, S. (Eds.), *Phi Theory: Phi-features across modues and interfaces*, (pp. 184–220)., Oxford. Oxford University Press.

Harizanov, B. (2014). *On the Mapping From Syntax to Morphophonology*. PhD thesis, University of California Santa Cruz, Santa Cruz, CA.

Harley, H. (2014). On the identity of roots. *Theoretical Linguistics*, *40*(225-276), 225–276.

Harley, H. & Ritter, E. (2002). Person and number pronouns: A feature-geometric analysis. *Language*, *78*(3), 482–526.

Haspelmath, M. (1993). *A Grammar of Lezgian*. Berlin: Mouton de Gruyter.

Haspelmath, M. (2016). Number suppletion vs. case suppletion: Does "locality" provide an explanation? *Diversity Linguistics Comment (blog). http://dlc.hypotheses.org/902*.

Herrity, P. (2000). *Slovene: A Comprehensive Grammar*. London, New York: Routledge.

Hewitt, B. G. (1995). *Georgian: A Structural Reference Grammar*. Amsterdam: John Benjamins.

Jurgensen, A. (2011). Consonant types and F0 in Kiowa. *UC Berkeley Phonology Lab Annual Report*, 63–76.

Keine, S. (2010). *Case and Agreement from Fringe to Core: A Minimalist Approach*. Berlin: De Gruyter.

Marvin, T. (2002). *Topics in the Stress and Syntax of Words*. PhD thesis, MIT, Cambridge, MA.

McFadden, T. (2014). Why nominative is special: Stem allomorphy and case structures. *Handout from GLOW 37, 4th April, Brussels*.

Mel'čuk, I. (2000). Suppletion. In Booij, G., Lehmann, C., Mugdan, J., Kesselheim, W., & Skopeteas, S. (Eds.), *Morphology 1: An International Handbook on Inflection and Word-Formation*, (pp. 510–522)., Berlin. De Gruyter.

Merchant, J. (2015). How much context is enough? Two cases of span-conditioned stem allomorphy. *Linguistic Inquiry*, *46*(2), 273–303.

Miller, T. L. (2015). A prosodic analysis of the word in Kiowa. *University of Pennsylvania Working Papers in Linguistics*, *21*(1), 1–10.

Moskal, B. (2015a). *Domains on the Border: Between Morphology and Phonology*. PhD thesis, University of Connecticut, Storrs, CT.

Moskal, B. (2015b). Limits on allomorphy: A case study in nominal suppletion. *Linguistic Inquiry*, *46*(2), 363–375.

Moskal, B. & Smith, P. W. (2016). Towards a theory without adjacency: Hypercontextual VI-rules. *Morphology*, (26), 295–312.

Nevins, A. (2007). The representation of third person and its consequences for person-case effects. *Natural Language and Linguistic Theory*, *25*(2), 273–313.

Ostrove, J. (2016). Portmanteaux and locality in the irish verbal complex. *Handout from LSA 2016, 7th January. Washington DC*.

Paster, M. E. (2006). *Phonological Conditions on Affixation*. PhD thesis, University of California Berkeley, Berkeley, CA.

Prince, A. & Smolensky, P. (1993/2004). *Optimality Theory: Constraint Interaction in Generative Grammar*. Oxford: Blackwell Publishing.

Radkevich, N. (2014). Nominal allomorphy in Lak. *Poster presented at NELS 45*.

Ronneberger-Sibold, E. (1988). Enstehung von suppletion un natürliche morphologie. *Zeitschrift für Phonetik, Sprachwissenschaft und Komunikationsforschung*, (41), 453–462.

Scatton, E. A. (1984). *A Reference Grammar of Modern Bulgarian*. Columbus, OH: Slavica Publishers, Inc.

Schiffman, H. (1999). *A Reference Grammar of Spoken Tamil*. Cambridge: Cambridge University Press.

Siddiqi, D. (2006). *Minimize exponence: economy effects on the morphosyntactic component of the grammar*. PhD thesis, University of Arizona, Tucson, AZ.

Siddiqi, D. (2009). *Syntax within the word: Economy, allomorphy and argument seletion in Distributed Morphology*. John Benjamins.

Siegel, D. (1978). The adjacency constraint and the theory of morphology. In Stein,

M. (Ed.), *Proceedings of the Annual Meeting of the Northeast Linguistic Society*, volume 8, (pp. 189–197)., Amherst, MA. GLSA.

Simpson, J. & Withgott, M. (1986). Pronominal clitic clusters and templates. In Borer, H. (Ed.), *Syntax and Semantics, Volume 19: The Syntax of Pronominal Clitics*, (pp. 149–74)., Orlando, FL. Academic Press.

Stump, G. (1996). Template morphology and inflectional morphology. In Booij, G. & van Marle, J. (Eds.), *Yearbook of Morphology 1996*, (pp. 217–241). Kluwer Academic Publishers.

Svenonius, P. (2004a). Slavic prefixes and morphology: An introduction to the *Nordlyd* volume. *Nordlyd*, *32*(2), 177–204.

Svenonius, P. (2004b). Slavic prefixes inside and outside VP. *Nordlyd*, *32*(2), 205–253.

Toporišič, J. (2000). *Slovenska slovnica*. Maribor: Obzorja.

Trommer, J. (1999). Morphology consuming syntax's resources: Generation and morphology consuming syntax's resources: Generation and parsing in a minimalist version of distributed morphology. *MS. Universität Potsdam, Potsdam.*

Tuite, K. (1998). *Kartvelian Morphosyntax: Number Agreement and Morphosyntactic Orientation in the South Caucasian Languages*. München: LINCOM Europe.

Vajda, E. J. (2003). Ket verb structure in typological perspective. *Sprachtypol. Univ. Forsch. (STUF)*, *56*(1/2), 55–92.

Watkins, L. J. (1984). *A Grammar of Kiowa*. Lincoln: University of Nebraska Press.

Werner, H. (1997). *Die ketische Sprache*. Wiesbaden: Harrassowitz.

Wolf, M. (2007). For an Autosegmental Theory of Mutation. In Bateman, L., O'Keefe, M., Reilly, E., & Werle, A. (Eds.), *University of Massachusetts Occasional Papers in Linguistics 32: Papers in Optimality Theory III*, (pp. 315–404)., Amherst, MA. GLSA.

7 Appendix I: Cases of Non-Local Patterns of Suppletion

(56) **GREEK**
Voice^0-Asp^0-triggered suppletion in V (Merchant 2015)

$\boxed{\sqrt{\text{tro}}}$	-∅	-∅	-o
eat	ACT	IMPF	1P.SG

$\boxed{\sqrt{\text{troɣ}}}$	-∅	-∅	-omun
eat	NON-ACT	IMPF	1P.SG

$\boxed{\sqrt{\text{fa}}}$	-∅	-∅	-o
eat	ACT	PRF	1P.SG

$\boxed{\sqrt{\text{faɣo}}}$	-θ	-ik	-a
eat	NON-ACT	PRF	1P.SG

Comments: according to Merchant (2015: 277), [tro-] and [troɣ-] are the same exponent, viz. /tro(ɣ)-/, as the /ɣ/ is sometimes dropped for independent reasons. /tro(ɣ)-/ must be the elsewhere exponent here, as it cuts across voice specification, as assumed by Merchant.

(57) **SLOVENIAN** (South Slavic)
Ptc^0-triggered suppletion in V (Božič 2016)

a.

$\boxed{\sqrt{\text{žanj}}}$	-e	-∅	-m
reap	ASP/THM	PRES.TNS	2P.SG

b.

$\boxed{\sqrt{\text{ž}}}$	-e	-l	-a
reap	ASP/THM	PTC	F.SG

Comments: Ptc^0 is not present in the first example, which is tensed. T^0, on the other hand, is not present in the participle, and spells out an auxiliary in such cases (not shown here). This is why Agr^0 codes person and number in the first example (this is what T^0 probes for in Slovenian), and gender and number in the second example (this is what Ptc^0 probes for). Also note that /žanj-/ needs to be the elsewhere item here, as argued in Bozic (2016) and discussed in section 3.3 of this paper.

(58) **TAMIL** (Dravidian)
K^0-triggered suppletion of D^0 in PRONOUN (Moskal & Smith 2016: 306)

a.

$\boxed{\text{naan}}$	-gal	-∅
1P.PRON	PL	NOM

b.

$\boxed{\text{en}}$	-gal	-ukku
1P.PRON	PL	DAT

(59) **TOTONAC** (Totozoquean)
PERSON-triggered suppletion in V (Brown et al. 2003)[25]

a.

$\boxed{\sqrt{\text{ma:}}}$	-ná	-∅
lie	IMPF	1P.PL

b.

$\boxed{\sqrt{\text{pa:}'}}$	-nán	-tit
lie	IMPF	2P.PL

[25] Brown et al. (2003) specify personal communication from Paulette Levy as the source for this pattern.

c.

ta-	$\boxed{\sqrt{\text{má:}}}$	-na
3P.PL	lie	IMPF

Comments: the elsewhere exponent here must be /ma:-/. We need to assume that this is the case since 1P and 3P cannot form a natural class to the exclusion of 2P under any treatment of person ϕ-features, be it geometric (Harley & Ritter 2002) or binary (Nevins 2007). For instance, if 1P is [+speaker,+participant] and 3P [–speaker,–participant], these two categories will not be able to form a natural class to the exclusion of 2P [–speaker,+participant].

(60) **LAK** (Northeast Caucasian)
K^0-triggered suppletion in N (Radkevich 2014), (Moskal 2015a: 35)

a.

$\boxed{\sqrt{\text{barz}}}$	-ru	-∅
moon	PL	NOM

b.

$\boxed{\sqrt{\text{zur}}}$	-dald	-il
moon	PL	ERG

(61) **TARIANA** (Arawakan)
NUMBER-triggered suppletion in A (Brown et al. 2003), (Aikhenvald 2003: 173)

a.

$\boxed{\sqrt{\text{hanu}}}$	-pua	-∅
big	CLASS	SG

b.

$\boxed{\sqrt{\text{male}}}$	-pua	-pe
big	CLASS	PL

(62) **KET** (Yeniseian)
$T^0_{\text{PRES/PAST}}$-triggered suppletion in V (Brown et al. 2003), (Werner 1997: 284)[26]

a.

ku-	∅-	ɣu-	$\boxed{\sqrt{\text{tus'}}}$
2P.SUBJ	PRES.TNS	2P.OBJ	intend

b.

∅-	il'-	gu-	$\boxed{\sqrt{\text{dɛn}}}$
2P.SUBJ	PAST.TNS	2P.OBJ	intend

(63) **BASQUE**
CMPR^0-triggered suppletion in A (Bobaljik 2012: 156-158)

a.

$\boxed{\sqrt{\text{asko}}}$			
much			[positive degree]

b.

$\boxed{\sqrt{\text{gehi}}}$	-ago		
much	CMPR		[comparative degree]

$\boxed{\sqrt{\text{gehi}}}$	-xe	-ago	
much	DIM	CMPR	[comparative degree]

[26] Glosses based on Vajda (2003).

Comments: the exponent in the positive degree, viz. /asko-/, needs to be the elsewhere item. According to Bobaljik (2012), the positive degree has no head encoding its 'positive' status, but the comparative degree does (viz. Cmpr^0). Hence, the positive degree is always a proper subset of the comparative degree, and a contextual rule can only operate on the presence of additional heads and not their absence.

8 Appendix II: Non-Local Affixal Allomorphy

(64) **BULGARIAN** (South Slavic)
T$^0_{[\text{IMPRF/AOR}]}$-THM-triggered allomorphy (Scatton 1984: 223-228)

a. √krad -$\boxed{\text{ɛ}}$ -ʃ -ɛ
steal THM$_{\text{CL1}}$ IMPERF.TNS 2P.SG

b. √krad^j -a -x -tɛ
steal THM$_{\text{CL1}}$ IMPERF.TNS 2P.PL

c. √krad -$\boxed{\text{ɛ}}$ -∅ -∅
steal THM$_{\text{CL1}}$ AOR.TNS 2P.SG

d. √krad -o -x -tɛ
steal THM$_{\text{CL1}}$ AOR.TNS 2P.PL

Comments: In Bulgarian, the allomorph of the theme vowel (THM) has a contextual spell-out in the context of 2/3P.SG forms in the imperfect and aorist tenses, but this only happens with a fixed class of verbs, labelled as CL1 in the data above. Only the 2P forms are shown here. Stump (1996) and Bobaljik (2000) suggest that THM is outwardly sensitive to the features of T^0 and Agr^0.

(65) **KIOWA** (Tanoan)
$v^0_{\text{TRANS/INTRANS}}$-triggered allomorphy of Mod^0 (Bonet & Harbour 2012: 231)

a. héíb -e -gųų -mɔɔ $\boxed{\text{-tɔɔ}}$
enter TR DISTR NEG MOD
'will not bring in at different times/locations'

b. héíb -é -gųų -mɔɔ $\boxed{\text{-t'ɔɔ}}$
enter INTR DISTR NEG MOD
'will not com in at different times/locations'

See section 6.5 on this pattern. This may not be an instance of non-local allomorphy.