Phonology (2026), **Vol. 43**:e4 1–33
doi:10.1017/S0952675726100311

PHONOLOGY

ARTICLE

Superadditive cumulativity in categorical prosodic patterns: prosodic minimality in Bosnian/Croatian/Montenegrin/Serbian

Aljoša Milenković

Linguistics, Harvard University, USA
Email: aljosamilenkovic@g.harvard.edu

Received: 23 September 2023; **Revised:** 9 June 2024; **Accepted:** 5 January 2025

Keywords: tone; stress; prosodic minimality; cumulativity; superadditivity; local constraint conjunction; Optimality Theory; Harmonic Grammar

Abstract

This article presents a case of superadditive ganging-up cumulativity in the metrical phonology of Bosnian/Croatian/Montenegrin/Serbian (BCMS). BCMS individually permits monomoraic feet and feet with a toneless head mora, but prohibits toneless monomoraic feet. Across BCMS dialects, several prosodic processes conspire against this doubly marked structure. Because of the superadditive character of this interaction, both Optimality Theory and, importantly, Harmonic Grammar require local constraint conjunction to capture the ban on toneless monomoraic feet in BCMS. This demonstration constitutes evidence for conjoined constraints in weighted constraint grammar. The study contributes to the typology of cumulativity effects by documenting superadditive ganging up in a categorical prosodic pattern, whereas virtually all previously reported cases of superadditivity have been observed in variable phonological patterns.

Contents

1. Introduction

Constraint-based theories of phonology employ two strategies to assess the severity of constraint violation: strict ranking and numeric weights. In strict-ranking Optimality Theory (OT; Prince & Smolensky [1993] 2004), the winner is the candidate that fares best on the highest-ranking constraint(s), irrespective of the overall violation profile. In Harmonic Grammar (HG; Legendre *et al.* 1990), well-formedness is computed by the weighted sum of all constraint violations.

HG predicts that the combined effect of multiple lower-weighted penalties can surpass the effect of violating a single higher-weighted constraint. Constraint cumulativity effects are of two types: GANGING-UP cumulativity and COUNTING cumulativity (Jäger & Rosenbach 2006). Ganging up, exemplified in the HG tableau in (1), arises when violating constraints $\mathbb{B}$ and $\mathbb{C}$ individually is less severe than violating $\mathbb{A}$, but violating $\mathbb{B}$ and $\mathbb{C}$ together surpasses violating $\mathbb{A}$ alone (Farris-Trimble 2008; Kenstowicz 2009; Pater 2009b; Albright 2012; McPherson 2016; Ryan 2017; Breiss 2020; Breiss & Albright 2022). In counting cumulativity, as in (2), multiple violations of a weaker constraint $\mathbb{B}$ incur a more severe penalty than a single violation of higher-weighted $\mathbb{A}$ (Breiss 2020; Kawahara 2020; Kawahara & Breiss 2021; Kawahara & Kumagai 2021; Kim 2022).

(1) *Ganging-up cumulativity*

	$\mathscr{H}$	$\mathbb{A}$ 2	$\mathbb{B}$ 1.5	$\mathbb{C}$ 1.5
☞ a. cand1	−2	−1		
b. cand2	−3		−1	−1

(2) *Counting cumulativity*

	$\mathscr{H}$	$\mathbb{A}$ 2	$\mathbb{B}$ 1.5
☞ a. cand1	−2	−1	
b. cand2	−3		−2

Constraint cumulativity has figured prominently in the debate between OT and HG. OT uses LOCAL CONSTRAINT CONJUNCTION (LCC; Smolensky 1993, 2006) to capture ganging-up cumulativity. HG accommodates constraint cumulativity without LCC (as illustrated in (1) and (2); see also Farris-Trimble 2008; Pater 2009b, 2016; Potts *et al.* 2010). This has been viewed as a major argument in favour of HG over OT (Pater 2009a, b, 2016), in light of conjoined constraints' propensity for overgeneration (see Kirchner 1997; McCarthy 2003, among others).

However, without non-default mechanisms such as LCC, HG successfully models only a subset of cumulativity effects. Recent work (e.g., Smith & Pater 2020; Breiss & Albright 2022) distinguishes three types of cumulative effects: LINEAR (i.e., ADDITIVE), SUPERLINEAR (i.e., SUPERADDITIVE) and SUBLINEAR (i.e., SUBADDITIVE). Linear effects, as defined in (3a), occur when the effect on acceptability/probability of simultaneously violating constraints $\mathbb{A}$ and $\mathbb{B}$ equals the combination of independent effects of $\mathbb{A}$ and $\mathbb{B}$. The combined effect of $\mathbb{A}$ and $\mathbb{B}$ is superlinear if the effect of coincident violation of $\mathbb{A}$ and $\mathbb{B}$ exceeds the joint effect of their independent violations, as in (3b). In sublinear cumulativity, defined in (3c), the effect of coincident violation of $\mathbb{A}$ and $\mathbb{B}$ results in a smaller decline in acceptability/probability than the combination of their independent effects.

(3) a. *Linearity/additivity*
$e(\mathbb{A}) + e(\mathbb{B}) = e(\mathbb{A}, \mathbb{B})$

b. *Superlinearity/superadditivity*
$e(\mathbb{A}) + e(\mathbb{B}) < e(\mathbb{A}, \mathbb{B})$

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

c. *Sublinearity/subadditivity*
$e(\mathbb{A}) + e(\mathbb{B}) > e(\mathbb{A}, \mathbb{B})$

where $e(\mathbb{C})$ is the effect of violating constraint $\mathbb{C}$ and $(\mathbb{C}_1, \mathbb{C}_2)$ is the coincident violation of $\mathbb{C}_1$ and $\mathbb{C}_2$

HG captures additive cumulativity with no appeal to nondefault devices, as in (4), but not superadditive ganging-up cumulativity (Albright 2009; Green & Davis 2014; Shih 2017; Smith 2022). This is illustrated in (5), where candidate (5a), the intended winner, loses out to candidate (5b). (Suppose (i) that there is independent evidence that $\mathbb{A}$ individually outweighs both $\mathbb{B}$ and $\mathbb{C}$, and (ii) that the joint violation of $\mathbb{B}$ and $\mathbb{C}$ incurs a greater penalty than the joint violation of $\mathbb{A}$ and $\mathbb{B}$.)

(4) *Additivity*

	$\mathscr{H}$	$\mathbb{A}$ 2	$\mathbb{B}$ 1.5	$\mathbb{C}$ 1.5
☞ a. cand1	−2	−1		
b. cand2	−3		−1	−1

(5) *Superadditivity*

	$\mathscr{H}$	$\mathbb{A}$ 2	$\mathbb{B}$ 1.5	$\mathbb{C}$ 1.5
☹ a. cand1	−3.5	−1	−1	
💣 b. cand2	−3.0		−1	−1

To model superadditive constraint cumulativity, HG adopts weighted constraint conjunction (Albright 2009; Green & Davis 2014; Shih 2017). In (6), the local conjunction of constraints $\mathbb{B}$ and $\mathbb{C}$ adds weight to their simultaneous violation, enabling candidate (6a) to prevail over candidate (6b) despite the shared violation of $\mathbb{B}$. Unlike in OT (Itô & Mester 1998; Baković 2000), the conjoined constraint need not dominate individual conjuncts in HG. This follows from the cumulativity of constraint violations.

(6) *Modelling superadditivity with local conjunction*

	$\mathscr{H}$	$\mathbb{A}$ 2	$\mathbb{B}$ 1.5	$\mathbb{C}$ 1.5	$\mathbb{B}\&\mathbb{C}$ 1
☞ a. cand1	−3.5	−1	−1		
b. cand2	−4.0		−1	−1	−1

Thus, HG does not entirely dispense with LCC. However, the two frameworks make different use of LCC: OT employs LCC to model both additive and superadditive ganging-up, while HG uses it only to model superadditivity.

This article provides further evidence for the necessity of conjoined constraints in weighted constraint grammar through the examination of a complex prosodic minimality effect in Bosnian/Croatian/Montenegrin/Serbian (BCMS). BCMS tolerates both monomoraic feet and feet headed by a toneless mora alone, but not their combination – monomoraic feet with a toneless head mora. The prohibition on toneless monomoraic feet manifests itself in several processes, including tone-sensitive monosyllabic lengthening (ML). Lengthening applies to toneless monomoraic forms, as in (7a), but not to their high-toned counterparts, as in (7b).

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

(7) a. /lɛd/ [ˈ(lɛɛd)] ‘ice.NOM.SG’
b. /brát/ [ˈ(brát)] ‘brother.NOM.SG’ not *[ˈ(bráat)]

This pattern illustrates the joint effect of two independently active constraints in the language: HEAD-H (Yip 2001), which requires that foot-heading moras be high-toned, and FOOTBINARITY (Prince & Smolensky [1993] 2004), which dictates that feet be binary at some level. I show that the coincident violation of HEAD-H and FOOTBINARITY results in a greater decline in metrical well-formedness than expected from their independent effects, indicative of superlinearity.

This article makes two contributions to the study of constraint cumulativity. First, contrary to the common position in the HG literature (Potts *et al.* 2010; Pater 2009b), I show that additive cumulativity of simplex weighted constraints does not supplant the full range of effects of LCC (Green & Davis 2014; Shih 2017; Smith 2022). This adds to the growing evidence for conjoined constraints in weighted constraint grammar (Albright 2009; Hayes *et al.* 2012; Green & Davis 2014; Shih 2017). Second, the article identifies superadditive cumulativity in a categorical prosodic pattern. Previous studies have mainly examined superadditivity in variable patterns (Shih 2017; Smith & Pater 2020; Breiss & Albright 2022; Kim 2022).

Additionally, the article identifies an interplay between prosodic minimality and tone in BCMS, contributing to prosodic typology. Tone is known to interact with stress (de Lacy 2002; Gordon 2023), and the BCMS patterns discussed herein indicate that tone also interacts with prosodic size constraints.

The article is organised as follows. In §2, I present a perplexing tone-sensitive prosodic minimality effect in BCMS and provide new dialectal data that substantiate this effect. In §3, I propose a unified analysis of the patterns outlined in §2 and discuss their theoretical implications. §4 compares the present analysis with the alternative account proposed by Zec (1999), showing that only the present analysis captures all prosodic minimality phenomena in BCMS. §5 concludes.

2. Prosodic minimality in BCMS

2.1. Background

All BCMS dialects examined in this article have lexical tone. Morphological structure and syllable weight are the primary factors in lexical tone assignment (Zec 1999). Animacy predicts tonal class affiliation in some inflectional classes (Martinović 2012). The crucial properties of the lexical tone system of BCMS are the following:

1. *Mora as the tone-bearing unit (TBU)*: Following Inkelas & Zec (1988), it is the mora rather than the syllable that functions as the TBU in BCMS (*contra* Langston 1997; see Zsiga & Zec 2013: 101 for arguments).
2. *Privativity and underspecification:* Underlyingly, moras either are high-toned or lack tone. Only high tone is specified in the lexicon; all moras that do not bear a H tone underlyingly are toneless and surface with a low pitch by default (Inkelas & Zec 1988; Zec 1999).
3. *Culminativity*: BCMS displays tonal culminativity, since the language allows for at most one singly-linked high tone per word. BCMS therefore falls into the category of restricted tone systems (Hyman 2006).
4. *No rising contours*: Rising contour tones are strictly prohibited in Štokavian BCMS: highs must be attached to the first mora of a heavy syllable (Ivić 1958; Inkelas & Zec 1988; Bethin 1998).

In BCMS, lexical highs determine the locus of stress (Inkelas & Zec 1988; Zec 1999; de Lacy 2002; Zec & Zsiga 2022). The two major dialect groups, Neoštokavian (NS) and Old Štokavian (OS), share the same tonal system but differ in stress placement relative to high tone. This section provides a brief overview of the stress pattern of NS varieties of BCMS, and a more detailed overview of OS word prosody.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

The word prosody of NS, the standard variety of BCMS, has been a topic of considerable debate. Extensive research has been conducted on the acoustics of NS stress (Lehiste & Ivić 1986; Zsiga & Zec 2013; Batas 2014), as well as its phonological representation (Inkelas & Zec 1988; Bethin 1998; Zec & Zsiga 2010). OS prosody has received less attention in the generative literature than the prosody of NS (a notable exception being Zec & Zsiga 2022). There are two primary reasons for this disparity. First, OS exhibits a high degree of prosodic variation across individual dialects (Ivić 1958), which makes it difficult to fit into a coherent theory. Second, most existing descriptive studies of OS are exclusively available in BCMS, limiting their accessibility to a broader audience.

2.1.1. Neoštokavian. In NS, lexical highs determine stress position. Stress falls on the syllable immediately preceding the word's only high, as in (8), or on the high-toned syllable if no syllable precedes, as in (9). In underlyingly toneless inputs, NS displays initial stress with high tone insertion, as in (10).

(8)	a.	/u.sta.nɔ.ʋí.ti/	[u.sta.ˈnɔ.ʋí.ti]	'establish.INF'
	b.	/u.ráa.diim/	[ˈu.ráa.diim]	'do.PRS.1SG'
	c.	/dɔ.pri.nɔ́.sii.mɔ/	[dɔ.ˈpri.nɔ́.sii.mɔ]	'contribute.PRS.1PL'
(9)	a.	/ráa.dii.mɔ/	[ˈráa.dii.mɔ]	'work.PRS.1PL'
	b.	/nɔ́.sii.mɔ/	[ˈnɔ́.sii.mɔ]	'carry.PRS.1PL'
	c.	/dá.dɔ.ʃɛ/	[ˈdá.dɔ.ʃɛ]	'give.AOR.3PL'
(10)	a.	/prɔ.da.dɛ/	[ˈprɔ́.da.dɛ]	'sell.AOR.2/3SG'
	b.	/u.kra.dɛ/	[ˈú.kra.dɛ]	'steal.AOR.2/3SG'
	c.	/prɔ.pa.dɛ/	[ˈprɔ́.pa.dɛ]	'fail.AOR.2/3SG'

Forms in (9) and (10) are indistinguishable in isolation; the tonal patterning of prefixes indicates whether the initial syllable is underlyingly high-toned (as in (9)) or has an inserted high (as in (10)). Forms with an initial lexical high retain their tone with a prefix, as in (11), while underlyingly toneless forms have a high inserted on the prefix, as in (12).

(11) *Stem-initial lexical high*

a.	/nɛ-ráa.dii.mɔ/	[ˈnɛ.ráa.dii.mɔ]	'not-work.PRS.1PL'
b.	/nɛ-nɔ́.sii.mɔ/	[ˈnɛ.nɔ́.sii.mɔ]	'not-carry.PRS.1PL'
c.	/nɛ-dá.dɔ.ʃɛ/	[ˈnɛ.dá.dɔ.ʃɛ]	'not-give.AOR.3PL'

(12) *No lexical high*

a.	/nɛ-prɔ.da.dɛ/	[ˈné.prɔ.da.dɛ]	'not-sell.AOR.2/3SG'
b.	/nɛ-u.kra.dɛ/	[ˈné.u.kra.dɛ]	'not-steal.AOR.2/3SG'
c.	/nɛ-prɔ.pa.dɛ/	[ˈné.prɔ.pa.de]	'not-fail.AOR.2/3SG'

Bethin (1994, 1998) analyses NS as a SYLLABIC TROCHEE system.[1] This analysis sheds light on the stress–tone adjacency in NS seen in (8): the stressed and high-toned syllable must be contained within a disyllabic trochee, as shown in (13).

(13) *NS syllabic trochee*
/u.ráa.diim/ → [(ˈu.ráa).diim] 'do.PRS.1SG'

Thus, NS stress assignment is QUANTITY-INSENSITIVE: stress falls on the syllable preceding a lexical H regardless of syllable weight, as in (8).

[1] Alternatively, Zec & Zsiga's (2010) analysis of the NS word prosodic system is foot-free.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

2.1.2. Old Štokavian. As a general tendency, stress in OS typically falls on the high-toned syllable (Zec & Zsiga 2022), as shown in (14).

(14)

a.	/sɛ.kí.ra/	[sɛ.ˈkí.ra]	‘axe.NOM.SG’
b.	/ka.púut/	[ka.ˈpúut]	‘coat.NOM.SG’
c.	/u.ráa.dii.mɔ/	[u.ˈráa.di.mɔ]	‘do.PRS.1SG’
d.	/prɔ.dáaʊ.ni.ʦa/	[prɔ.ˈdáaʊ.ni.ʦa]	‘store.NOM.SG’

The preference for stress on high-toned syllables arises from the HEAD-H constraint (Yip 2001), defined in (15).

(15) HEAD-H: Every foot-heading mora must bear a high tone.

Despite the general preference for stressed high-toned syllables in OS, stress does not invariably fall on the high-toned syllable in all prosodic environments in all OS dialects (Zec & Zsiga 2022). There are two cases in which some OS dialects avoid stressing the high-toned syllable. The first point of variation across individual OS dialects stems from the prosodic patterning of final light syllables. When a lexical H is attached to a final light syllable, stress falls on the final light in only a handful of OS dialects, as in (16), and on the toneless penult in most OS dialects, as in (17):

(16) *Final stress permitted*

a.	/ʊɔ.dá/	[ʊɔ.ˈdá]	‘water.NOM.SG’
b.	/pɔ.ʧók/	[pɔ.ˈʧók]	‘creek.NOM.SG’
c.	/ʊru.ʨi.ná/	[ʊru.ʨi.ˈná]	‘heat.NOM.SG’
d.	/za.saa.dí/	[za.saa.ˈdí]	‘plant.IMP.2SG’

(17) *Final stress prohibited*

a.	/ʊɔ.dá/	[ˈʊɔ.dá]	‘water.NOM.SG’
b.	/pɔ.ʧók/	[ˈpɔ.ʧók]	‘creek.NOM.SG’
c.	/ʊru.ʨi.ná/	[ʊru.ˈʨi.ná]	‘heat.NOM.SG’
d.	/za.saa.dí/	[za.ˈsaa.dí]	‘plant.IMP.2SG’

In OS dialects that prohibit final stress, as in (17), only *light* final syllables are barred from bearing stress, as in (18), while final heavies regularly receive stress if high-toned underlyingly, as in (19).[2] Recall that a H tone must be attached to the first mora of a heavy syllable.

(18) *Final high-toned light syllables*

a.	/ʊɔ.dá/	[ˈʊɔ.dá]	‘water.NOM.SG’
b.	/pɔ.ʧók/	[ˈpɔ.ʧók]	‘creek.NOM.SG’
c.	/ruu.ká/	[ˈruu.ká]	‘arm.NOM.SG’
d.	/naa.rɔ́d/	[ˈnaa.rɔ́d]	‘people.NOM.SG’

(19) *Final high-toned heavy syllables*

a.	/ʊɔ.dɛ́ɛ/	[ʊɔ.ˈdɛ́ɛ]	‘water.GEN.SG’
b.	/ʊɔ.dɔ́ɔm/	[ʊɔ.ˈdɔ́ɔm]	‘water.INS.SG’
c.	/ruu.kɛ́ɛ/	[ruu.ˈkɛ́ɛ]	‘arm.GEN.SG’
d.	/ruu.kɔ́ɔm/	[ruu.ˈkɔ́ɔm]	‘arm.INS.SG’

[2]For the time being, I ignore the patterning of tone and weight in toneless penultimate lights (see §2.4).

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

Word-final high-toned lights regularly receive stress in enclisis, which results in productive stress alternations, as in (20).

(20)

a.	[ˈʊɔ.dá]	‘water’	[ʊɔ.ˈdá=jɛ]	‘water is’
b.	[ˈjɛ.zík]	‘tongue’	[jɛ.ˈzík=mi]	‘my tongue’
c.	[ˈruu.ká]	‘arm’	[ruu.ˈká=mi]	‘my arm’

The illicitness of stressed final lights but not of stressed final heavies in OS falls out from the moraic version of NONFINALITY.[3] Mora-oriented NONFINALITY penalises word-level prominence on the rightmost mora of the prosodic word (ω) (Hyde 2007). Stressed final heavies do not violate the moraic version of NONFINALITY because they form a bimoraic foot headed by the syllable’s first mora, which is the penultimate mora of a word. Thus, when stress is on a final heavy, the prominence peak does not coincide with the word’s rightmost mora.

(21) NONFINALITY(μ,ω,Ft): No mora which is final in the prosodic word heads a foot.

The second point of variation in OS concerns the metrical treatment of toneless heavy syllables. Individual OS dialects differ with respect to whether they prefer stressing high-toned lights over toneless heavies, as in (22), or *vice versa,* as in (23).

(22) *High-toned light prevails*

a.	/u.raa.dí.ʃɛ/	[u.raa.ˈdí.ʃɛ]	‘do.AOR.3PL’
b.	/naa.rɔ́.di.ma/	[naa.ˈrɔ́.di.ma]	‘people.DAT.PL’

(23) *Toneless heavy prevails*

a.	/u.raa.dí.ʃɛ/	[u.ˈraa.dí.ʃɛ]	‘do.AOR.3PL’
b.	/naa.rɔ́.di.ma/	[ˈnaa.rɔ́.di.ma]	‘people.DAT.PL’

The pattern in (23) arises in response to the WEIGHT-TO-STRESS PRINCIPLE (WSP; Prince 1990), defined in (24). For (23) to arise, WSP must dominate HEAD-H.

(24) WEIGHT-TO-STRESS PRINCIPLE: Assess a violation for every unstressed heavy syllable.

The combinations of the metrification strategies in (16)–(17) and (22)–(23) give rise to four basic prosodic types in OS. In Type 1 (East Montenegro; Stevanović 1933), stress invariably falls on the high-toned syllable, as in (25a). In Type 2 (Southwest, Central and East Serbia, parts of Montenegro and Kosovo; Jović 1968; Simić 1972), stress generally falls on the high-toned syllable, except when the high-toned syllable is light and final, as in (25b). In Type 3 (Lepetane in the Kotor Bay, Montenegro; Tomanović 1935) high-toned syllables are stressed unless the immediately preceding syllable is heavy, in which case stress falls on the toneless heavy, as in (25c). Type 4 (Smederevo–Vršac dialect cluster; Ivić 1958) avoids stressing final high-toned lights and prefers stressing toneless heavies over high-toned lights, as in (25d). In all environments not listed in (25), stress invariably falls on the high-toned syllable in all four types of OS dialects.

(25)

a.	*Type 1*		
	/ʊɔ.dá/	[ʊɔ.ˈdá]	
	/u.raa.dí.ʃɛ/	[u.raa.ˈdí.ʃɛ]	
b.	*Type 2*		
	/ʊɔ.dá/	[ˈʊɔ.dá]	
	/u.raa.dí.ʃɛ/	[u.raa.ˈdí.ʃɛ]	

[3] The present approach to nonfinality effects in OS differs slightly from Zec & Zsiga (2022), but the differences between the two approaches are beyond the scope of this article.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

c. *Type 3*
/ʊɔ.dá/ [ʊɔ.ˈdá]
/u.raa.dí.ʃɛ/ [u.ˈraa.dí.ʃɛ]

d. *Type 4*
/ʊɔ.dá/ [ˈʊɔ.dá]
/u.raa.dí.ʃɛ/ [u.ˈraa.dí.ʃɛ]

The prosodic variation observed across individual OS dialects can be attributed to the interplay of three constraints in OS: HEAD-H, NONFINALITY and WSP. In Type 1 dialects, HEAD-H dominates both NONFINALITY and WSP. In Type 2 dialects, HEAD-H takes precedence over WSP, but ranks below NONFINALITY. Type 3 is the mirror image of Type 2: HEAD-H outranks NONFINALITY, but ranks below WSP. In Type 4, HEAD-H is dominated by both NONFINALITY and WSP. The present study is primarily concerned with Type 2 and 4 OS dialects.

Putting the pieces together, OS stress is tone-driven and quantity-sensitive. The sensitivity to tone is evident from the preference for stressing syllables with a lexical H. Quantity-sensitivity manifests itself in several interactions. In Type 3 and 4 OS dialects, toneless heavy syllables attract stress over high-toned lights. The ban on final prominence in OS likewise makes reference to weight. There is an asymmetry between final high-toned lights, which are unstressable in Type 2 and 4 OS dialects, and final high-toned heavies, which regularly receive stress in all OS dialects. All descriptive facts presented in this section can be unified by assuming that OS forms the MORAIC TROCHEE defined in (26), the head mora of which is preferably high-toned.

(26) *OS moraic trochee*
($\acute{\mu}_s\mu_w$), where:
μ_s = strong, foot-heading mora; μ_w = weak, non-head mora
$\acute{\mu}$ = high-toned mora; μ = toneless mora

This footing preference explains the weight–stress and tone–stress interactions, along with nonfinality effects in OS, in a unified fashion. The moraic trochee analysis sheds light on another central property of OS word prosody: the ban on rising contours. In OS, highs must be attached to the first mora of a heavy syllable. Assuming the metrical structure in (26) for OS, stressed heavy syllables form a bimoraic trochee in this dialect group. Accordingly, the second mora of a stressed heavy constitutes the non-head position of a bimoraic trochee. Thus, in rising contours, a H tone would be linked to the non-head mora. Cross-linguistically, there is a dispreference for high-toned foot non-heads. This is formally captured by the constraint *NONHEAD-H (de Lacy 2002):

(27) *NONHEAD-H: Assess a violation for every instance of a high-toned mora in the non-head position of a foot.

In OS, *NONHEAD-H penalises stressed heavies with a rising contour tone because they feature a H tone on the non-head mora of a bimoraic trochee (see §3.2.2 for further discussion).

2.1.3. Interim summary. OS and NS have virtually identical underlying representations. The key distinction between the two dialect groups is that OS preferably stresses the high-toned syllable, while NS stresses the immediately preceding syllable:

(28)

	UR	OS	NS	Gloss
a.	/sɛ.kí.ra/	[sɛ.ˈkí.ra]	[ˈsɛ.kí.ra]	‘axe.NOM.SG’
b.	/u.pɔ.réɛ.diiʃ/	[u.pɔ.ˈréɛ.diʃ]	[u.ˈpɔ.réɛ.diiʃ]	‘compare.PRS.2SG’
c.	/ʊɔ.dɔ́ɔm/	[ʊɔ.ˈdɔ́ɔm]	[ˈʊɔ.dɔ́ɔm]	‘water.INS.SG’
d.	/daa.náa/	[daa.ˈnáa]	[ˈdaa.náa]	‘day.GEN.PL’

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

OS is a moraic trochee dialect, while NS exhibits the syllabic trochee. The footing differences between the two dialect groups are outlined in (29).

(29)	UR	OS	NS	Gloss
a.	/sɛ.kí.ra/	[sɛ.(ˈkí.ra)]	[(ˈsɛ.kí).ra]	‘axe.NOM.SG’
b.	/u.pɔ.réɛ.diiʃ/	[u.pɔ.ˈ(réɛ).diʃ]	[u.(ˈpɔ.réɛ).diiʃ]	‘compare.PRS.2SG’
c.	/ʊɔ.dɔ́ɔm/	[ʊɔ.ˈ(dɔ́ɔm)]	[(ˈʊɔ.dɔ́ɔm)]	‘water.INS.SG’
d.	/daa.náa/	[daa.ˈ(náa)]	[(ˈdaa.náa)]	‘day.GEN.PL’

OS stress pattern is quantity-sensitive, with feet containing at most two moras. NS stress pattern is quantity-insensitive: feet comprise two syllables regardless of their weight.

2.2. The minimal foot in BCMS

BCMS tolerates both degenerate (i.e., monomoraic) feet and feet headed by a toneless mora. Monomoraic feet with a toneless head mora, the combination of these two individually tolerable marked structures, is categorically prohibited. This gives rise to a typologically rare gap in the foot inventory of BCMS, outlined in (30) (see also Zec 1999: 237).

(30) *Moraic foot inventory of BCMS*

a.	Monomoraic feet:	(μ́)	*(μ)
b.	Bimoraic feet:	(μ́μ)	(μμ)

Effectively, bimoraic feet can incorporate two toneless moras or a high-toned mora followed by a toneless mora. However, the only mora of a degenerate foot must be high-toned. Across BCMS dialects, there is a widespread conspiracy against degenerate feet with a toneless head mora. Two strategies are employed to eliminate this illicit structure. First, BCMS displays a vowel lengthening process that targets toneless, but not high-toned, stressed lights (Zec 1999) (§2.3). Second, there is a process of tonal flop in some OS dialects, which shifts highs from unstressed syllables to underlyingly toneless stressed lights, but not to stressed toneless heavies (Ivić 1958). Further evidence for the conspiracy against toneless degenerate feet is provided by the fact that many OS dialects that do not allow tonal flop regularly exhibit vowel lengthening in stressed toneless lights. Thus, I argue that tonal flop and vowel lengthening are competing strategies employed to eliminate toneless degenerate feet across different OS dialects of BCMS (§2.4).

2.3. Monosyllabic lengthening

Previous work on BCMS tone has identified two tonal classes in nominals and verbs, the high-toned class and the toneless class (Inkelas & Zec 1988; Zec 1999; Martinović 2012). The two tonal classes observe different size restrictions: words that belong to the toneless class are minimally bimoraic, while high-toned words are allowed to be monomoraic. Different minimality conditions in the two classes fall out from a productive ML process which targets underlyingly toneless monomoraic content words, as in (31), but not their high-toned counterparts, as in (32) (cf. Zec 1999).[4] The lengthening pattern in (31) is observed in virtually all Štokavian dialects with contrastive vowel quantity, including both NS and OS dialects (Kapović 2015: 620).

[4] All toneless monosyllables with an underlyingly short vowel are treated as monomoraic and hence are subject to ML, which shows that final consonants do not contribute weight in BCMS.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

(31) *ML in toneless monosyllables*

a.	[ˈlɛɛd]	‘ice.NOM.SG’	[ˈlɛ.da]	‘ice.GEN.SG’
b.	[ˈktɕɛɛr]	‘daughter.ACC.SG’	[ˈktɕɛ.ri]	‘daughter.GEN.SG’
c.	[ˈkɔɔst]	‘bone.NOM.SG’	[ˈkɔ.sti]	‘bone.GEN.SG’
d.	[ˈbɔɔs]	‘barefoot.NOM.SG.M’	[ˈbɔ.si]	‘barefoot.NOM.PL.M’

(32) *No lengthening in high-toned monosyllables*

a.	[ˈbrát]	‘brother.NOM.SG’	[ˈbrá.ta]	‘brother.GEN.SG’
b.	[ˈdɛ́d]	‘grandfather.NOM.SG’	[ˈdɛ́.da]	‘grandfather.GEN.SG’
c.	[ˈzdrávʊ]	‘healthy.NOM.SG.M’	[ˈzdrá.ʊi]	‘healthy.NOM.PL.M’
d.	[ˈlíx]	‘pour.AOR.1SG’	[ˈlí.smɔ]	‘pour.AOR.1PL’

The alternation in (31) results from vowel lengthening in monosyllabic forms rather than vowel shortening in the initial syllable of polysyllabic forms. The data in (33) indicate that there is no vowel shortening in the initial syllable of BCMS polysyllables, refuting the shortening analysis of (31).

(33) *Long stem-vowel: no shortening*

a.	[ˈʊuuk]	‘wolf.NOM.SG’	[ˈʊuu.ka]	‘wolf.GEN.SG’
b.	[ˈgraad]	‘town.NOM.SG’	[ˈgraa.da]	‘town.GEN.SG’
c.	[ˈrɛɛʧ]	‘word.NOM.SG’	[ˈrɛɛ.ʧi]	‘word.NOM.PL’

Cross-linguistically, ML processes are commonly driven by the requirement that the smallest permissible foot or prosodic word comprise at least two moras, stated as a violable constraint in (34).

(34) FootBinarity: All feet are binary on the moraic or syllabic level.

However, if BCMS ML applied in response to FtBin alone, both (31) and (32) would be expected to undergo lengthening. Puzzlingly, ML makes crucial reference to tone: it is only toneless and not high-toned monomoraic forms that undergo this process in BCMS.[5]

The dual behaviour of BCMS monosyllables illustrated in (31) and (32) can be analysed in two ways. On one account, vowel lengthening in monomoraic forms is the default. Lengthening is blocked in high-toned monosyllables because the resulting structure is marked. This is the gist of the analysis proposed by Zec (1999). Another possibility, which I argue for here, is that the absence of lengthening in high-toned monosyllables is the default. Toneless monomoraic feet constitute a uniquely marked structure, which triggers the otherwise inactive lengthening process.

2.4. Tonal flop and penultimate lengthening

In addition to ML, there are other processes that conspire against toneless monomoraic feet across BCMS dialects. Some OS dialects exhibit foot optimisation strategies that remove toneless monomoraic

[5]It should be noted that the context of ML becomes non-surface-apparent as a result of a high tone insertion process. In BCMS, all underlyingly toneless prosodic words receive a high tone on their initial mora by a default insertion rule (recall (10)). The interaction between ML and tone insertion is illustrated in (i).

(i)

UR	(32a) /brát/	(31a) /lɛd/	(10a) /prɔpadɛ/
Stress assignment	ˈbrát	ˈlɛd	ˈprɔpadɛ
ML	—	ˈlɛɛd	—
Tone insertion	—	ˈlɛ́ɛd	ˈprɔ́padɛ
Surface form	[ˈbrát]	[ˈlɛ́ɛd]	[ˈprɔ́padɛ]

Tone insertion counterbleeds ML: were tone insertion to apply first, toneless monomoraic forms would be rendered indistinguishable from high-toned monosyllables, which are exempt from lengthening. This interaction may be ascribed to contrast preservation (Łubowicz 2003), whereby lengthening in toneless monosyllables overapplies to preserve the contrast between the two tonal classes in monomoraic stems. The overapplication of ML is beyond the scope of the present article.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

feet in polysyllables. These involve tonal flop (TF; Ivić 1958; Simić 1972; Remetić 1985) and penultimate lengthening (PL; Ivić 1958; Remetić 1985). In this section, I provide a unified account of these two previously unaddressed metrical processes in OS.

In OS, stress generally falls on the word's only high-toned syllable. However, in Type 2 and Type 4 OS dialects (§2.1.2), stress falls on the toneless penult if the high-toned syllable is light and final. Curiously, in OS dialects that prohibit final prominence, toneless penultimate lights behave differently under stress from toneless penultimate heavies. This is reminiscent of the dual patterning of monosyllables seen in (31) and (32). Stressed toneless heavies consistently surface faithfully in all Type 2 and Type 4 OS dialects (Jović 1968: 26; Simić 1972: 51–54; Remetić 1985: 40), as (35) shows. (Feet are delimited by parentheses from here on.)

(35)	a.	/ruu.ká/	[ˈ(ruu).ká]	'arm.NOM.SG'
	b.	/nii.sám/	[ˈ(nii).sám]	'not.am'
	c.	/za.saa.dí/	[za.ˈ(saa).dí]	'plant.IMP.2SG'
	d.	/pɔ.mɛɛ.ní/	[pɔ.ˈ(mɛɛ).ní]	'mention.IMP.2SG'
	e.	/u.zɛɛ.ʃɛ́/	[u.ˈ(zɛɛ).ʃɛ́]	'take.AOR.3PL'
(36)	a.	/ruu.ká/	[ruu.(ˈká=mi)]	'arm=my'
	b.	/nii.sám/	[nii.(ˈsám=sɛ)]	'not.am=REFL'
	c.	/za.saa.dí/	[za.saa.(ˈdí=ga)]	'plant=it'
	d.	/pɔ.mɛɛ.ní/	[pɔ.mɛɛ.(ˈní=me)]	'mention=me'
	e.	/u.zɛɛ.ʃɛ́/	[u.zɛɛ.(ˈʃɛ́=ni)]	'took=us'

Unlike penultimate heavies, stressed penultimate lights undergo TF or PL in various OS dialects. TF shifts high tone from unstressed final syllables to the preceding light syllable, as in (37).[6] Enclisis reveals the original position of tone, as in (38).

(37)	a.	/vɔ.dá/	[(ˈvɔ́.da)]	'water.NOM.SG'
	b.	/dɔ.ʃlí/	[(ˈdɔ́.ʃli)]	'came.M.PL'
	c.	/ɔ.táʦ/	[(ˈɔ́.taʦ)]	'father.NOM.SG'
	d.	/grɛ.ɔ.tá/	[grɛ.(ˈɔ́.ta)]	'shame.NOM.SG'
	e.	/ɔ.ʒɛ.ní/	[ɔ.(ˈʒɛ́.ni)]	'marry.IMP.2SG'
	f.	/sa.kri.vɛ́n/	[sa.(ˈkrí.vɛn)]	'hidden.M.SG'
(38)	a.	/vɔ.dá/	[vɔ.(ˈdá=jɛ)]	'water=is'
	b.	/dɔ.ʃlí/	[dɔ.(ˈʃlí=su)]	'come=are.3PL'
	c.	/ɔ.táʦ/	[ɔ.(ˈtáʦ=mi)]	'father=my'
	d.	/grɛ.ɔ.tá/	[grɛ.ɔ.(ˈtá=jɛ)]	'it is shame'
	e.	/ɔ.ʒɛ.ní/	[ɔ.ʒɛ.(ˈní=sɛ)]	'get married'
	f.	/sa.kri.vɛ́n/	[sa.kri.(ˈvɛ́n=jɛ)]	'hidden=is'

The data in (37) and (38) exhibit a stress-dependent tonal alternation. In forms that display penultimate stress in response to NONFINALITY, as in (37), high tone shifts to the stressed penult. However, the penultimate syllable remains toneless in enclisis, where stress falls on the originally high-toned syllable, as in (38).

In a subset of OS dialects, PL is employed instead of TF. The counterparts of the forms in (37) exhibit no tone shift, as the lexical high tone remains on the light final syllable on the surface. Instead, the stressed toneless penult becomes heavy by PL, as in (39).[7]

[6] The TF data in (37) are from Simić (1972: 47–48) and Remetić (1985: 40–41).

[7] The PL data in 39 are from Remetić (1985: 40–44).

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

(39)	a.	/vɔ.dá/	[ˈ(vɔɔ).dá]	‘water.NOM.SG’
	b.	/dɔ.ʃlí/	[ˈ(dɔɔ).ʃlí]	‘came.M.PL’
	c.	/ɔ.táʦ/	[ˈ(ɔɔ).táʦ]	‘father.NOM.SG’
	d.	/grɛ.ɔ.tá/	[grɛ.ˈ(ɔɔ).tá]	‘shame.NOM.SG’
	e.	/ɔ.ʒɛ.ní/	[ɔ.ˈ(ʒɛɛ).ní]	‘marry.IMP.2SG’
	f.	/sa.kri.vén/	[sa.ˈ(krii).vén]	‘hidden.M.SG’

In enclisis, illustrated in (40), the light ultima receives stress, while the penult, which is unstressed in this context, displays a short vowel.

(40)	a.	/vɔ.dá/	[vɔ.(ˈdá=jɛ)]	‘water=is’
	b.	/dɔ.ʃlí/	[dɔ.(ˈʃlí=su)]	‘come=are.3PL’
	c.	/ɔ.táʦ/	[ɔ.(ˈtáʦ=mi)]	‘father=my’
	d.	/grɛ.ɔ.tá/	[grɛ.ɔ.(ˈtá=jɛ)]	‘it is shame’
	e.	/ɔ.ʒɛ.ní/	[ɔ.ʒɛ.(ˈní=sɛ)]	‘get married’
	f.	/sa.kri.vén/	[sa.kri.(ˈvén=jɛ)]	‘hidden=is’

This alternation in vowel quantity must be attributable to lengthening in the stressed position rather than shortening in the pretonic position. The latter option is inconsistent with the treatment of long vowels in pretonic syllables in OS. Underlyingly long vowels remain long in pretonic syllables in OS dialects, as indicated by forms like [ruu.(ˈká=mi)] ‘my arm’ in (36a).

Like ML, TF and PL do not apply across the board, but are rather restricted to stressed toneless lights. TF and PL therefore constitute additional evidence for the prohibition on toneless degenerate feet in BCMS.

2.5. Conspiracy

TF and PL have received considerable attention in BCMS dialectology, primarily as classification criteria for OS dialects (Ivić 1958; Remetić 1985; Ligorio 2016).[8] However, no explanatory account of these processes has been offered to date.

TF and PL can profitably be conceived of as foot-optimisation strategies. TF makes the foot-heading mora high-toned, bringing the foot in compliance with HEAD-H in (15). Similarly, PL makes the foot bimoraic, as dictated by FTBIN in (34). However, the purpose of TF and PL goes beyond the removal of a singly marked structure in much the same way as ML.

TF shifts high tone from unstressed syllables to stressed toneless lights, as in (37). By contrast, highs do not shift to stressed toneless heavies, as can be seen in (35).

In the same vein, PL targets toneless stressed lights, as in (39). However, no vowel lengthening is observed in high-toned stressed lights. Consider the data from Gallipoli Serbian (Ivić 1957), a Type 2 OS dialect that exhibits PL, in (41).

(41)	a.	/kɔ́.maa.ta/	ˈ(kɔ́).maa.ta	‘piece.GEN.SG’
	b.	/pɔ́.jaa.sɛ/	ˈ(pɔ́).jaa.sɛ	‘belt.ACC.PL’
	c.	/í.ʤaa.ʃɛ/	ˈ(í).ʤaa.ʃɛ	‘go.IMPF.2/3SG’
	d.	/mɔ́.laa.ʃɛ/	ˈ(mɔ́).laa.ʃɛ	‘beg.IMPF.2/3SG’

In (41), stress falls on a high-toned light syllable, which forms a monomoraic foot.[9] These stressed lights do not undergo vowel lengthening, unlike their counterparts in (39). The difference between (39)

[8]The traditional terms for TF and PL in BCMS dialectology are METATAXIS (BCMS *metataksa*) and KANOVIAN LENGTHENING (BCMS *kanovačko duženje*), respectively (see Ligorio 2016 for terminology and an accessible overview of the processes).

[9]On the following assumptions: (i) in OS, feet are maximally bimoraic, which prevents the peninitial heavy from being included in the foot and (ii) OS respects syllable integrity, which ensures that the first mora of the peninital syllable cannot be included in the foot to the exclusion of the syllable’s second mora.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

and (41) is that the lengthened stressed lights in (39) are toneless, while the stressed lights in (41), which are not subject to lengthening, are high-toned.

TF and PL thus bear a striking resemblance to ML: TF and PL do not apply across the board, but are restricted to a specific, doubly marked environment – stressed toneless lights. TF and PL do not eliminate toneless feet or degenerate feet in isolation, but specifically target toneless degenerate feet.

I propose that the purpose of TF and PL is to eliminate doubly marked toneless degenerate feet, analogous to the ML process discussed in §2.3. Therefore, TF, PL and ML all take part in a CONSPIRACY (Kisseberth 1970) against toneless degenerate feet in NS and OS. The present analysis correlates three superficially distinct metrical processes in BCMS, highlighting their functional unity.

ML always applies to toneless monomoraic forms in all OS dialects, as elsewhere in BCMS. In a subset of OS dialects, either TF or PL applies to toneless monomoraic feet in polysyllabic forms. TF and PL are in complementary distribution: they never cooccur within the same OS dialect. Therefore, ML eliminates toneless monomoraic feet in monosyllables, while either TF or PL eliminates this impermissible structure in polysyllables.

Crucially, OS and NS dialect groups share in the ban on toneless monomoraic feet despite the differences in the types of footing they employ (recall §2.1.3). OS is a quantity-sensitive moraic trochee dialect. By contrast, NS is a quantity-insensitive syllabic trochee dialect. NS consistently forms disyllabic feet in polysyllabic forms. Underlyingly monomoraic content words therefore constitute the only case in which monomoraic feet can arise and thus the only context in which FTBIN can be violated in NS. This explains why ML is the only process employed in NS to eliminate toneless degenerate feet: NS never forms degenerate feet outside monomoraic forms, consistent with the syllabic trochee analysis. Feet are maximally bimoraic in OS, resulting in cases where monomoraic feet can arise even in polysyllabic forms (see §3.2). Toneless degenerate feet occur in both monosyllables and polysyllables in OS, and thus all three processes argued here to militate against toneless degenerate feet (ML, TF and PL) are employed across OS dialects.

The significance of the OS data adduced in (37)–(39) is twofold. First, these data indicate that the prohibition against toneless degenerate feet in BCMS goes beyond the process of ML. Second, these data provide a straightforward answer to one of the outstanding questions regarding prosodic minimality in BCMS: is there a difference between the minimal prosodic word (ω_{min}) and the minimal foot (Ft_{min}) of the language?

Following McCarthy & Prince (1986), it has been hypothesised that the minimal prosodic word is universally coextensive with the minimal foot. However, in many languages, there is a mismatch between the minimal prosodic word and the minimal foot (Downing 1998; Garrett 1999). Based solely on the ML data exemplified in (31), it is impossible to determine whether ML arises due to foot minimality or a separate word minimality constraint. The OS data presented in (37)–(39) show unequivocally that the complex lengthening pattern in (31) is not restricted to monosyllables, and thus cannot instantiate an independent word minimality effect.

In BCMS, there is a significant relationship between tone and prosodic minimality requirements. Lexical tone not only interacts with stress, but also takes part in defining the minimal acceptable foot in the language. Although OS and NS exhibit different types of footing, they share the conspiracy against toneless monomoraic feet.

3. Proposal

3.1. Joint and independent activity of HEAD-H and FTBIN in BCMS

The upshot of §2 is that both monomoraic feet and feet with a toneless head mora are individually permitted, but toneless monomoraic feet are strictly prohibited in BCMS. Various dialects of BCMS employ a series of foot-optimisation strategies to eliminate toneless monomoraic feet. This prosodic minimality effect arises via a complex interplay between constraints on the permissible size of metrical constituents and those on tone–stress interaction.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

Table 1. *Moraic trochees and their violation profiles (✓ = satisfies the constraint in question; ✗ = violates the constraint in question)*

	Head-H	FtBin	Permitted?	Example	
($\acute{\mu}_s\mu_w$)	✓	✓	Yes	(74)	[ˈ(kráaʌ́)]	‘king’
($\mu_s\mu_w$)	✗	✓	Yes	(35a)	[ˈ(ruu).ká]	‘arm’
($\acute{\mu}_s$)	✓	✗	Yes	(32a)	[ˈ(brát)]	‘brother’
(μ_s)	✗	✗	No		—	

Since BCMS is a trochaic language with tone-driven stress (Zec 1999, among others), I customise Prince’s (1990) Trochaic Rhythmic Harmony Scale by incorporating into it tonal restrictions on foot-heading moras:

(42) *Foot harmony scale for BCMS*
($\acute{\mu}_s\mu_w$) ≻ {($\mu_s\mu_w$), ($\acute{\mu}_s$)} ≻ *(μ_s)

Table 1 provides a summary of the foregoing discussion, listing different types of moraic trochees, their violation profiles and attestedness in BCMS.

Individual Head-H violations are tolerable across BCMS dialects, given numerous instances of feet headed by a toneless mora as in (39). Likewise, BCMS allows monomoraic feet with a high-toned head mora (see (32) and (41)). Accordingly, FtBin is independently violable in the language.

Although the independent violations of Head-H and FtBin are allowed in BCMS, coincident violation of these two markedness constraints is strictly prohibited. This explains the absence of toneless monomoraic feet from the language’s foot inventory (shown in (30)). To avoid doubly marked toneless monomoraic feet, BCMS subjects underlyingly toneless monomoraic feet to vowel lengthening (ML or PL) or TF. Both processes are otherwise hindered by faithfulness in BCMS.

I propose that the ban on doubly marked toneless monomoraic feet in BCMS instantiates a gang effect (Jäger & Rosenbach 2006; Pater 2009b, 2016; Breiss 2020; Breiss & Albright 2022), whereby the combination of two individually tolerable marked structures is strictly disallowed. The purpose of ML, TF and PL is to remove the coincident violation of Head-H and FtBin incurred by a toneless degenerate foot. ML, TF and PL do not aim to eliminate degenerate feet across the board, but target those degenerate feet which contain a toneless mora. This explains why these processes are restricted to stressed toneless light syllables and fail to apply in other marked environments. The relevant markedness constraints (Head-H and FtBin) are individually overridden by faithfulness. However, in environments that violate both, these two constraints gang up to overcome a stronger faithfulness constraint.

The minimal foot in BCMS is shaped by the interaction of two markedness constraints, Head-H and FtBin. Both constraints involved in this cumulative effect are independently active in BCMS phonology.

Recall from §2 that Head-H plays a pivotal role in OS, as high tone generally attracts stress in this dialect group. By contrast, in NS dialects, Head-H ranks relatively low, because stress does not coincide with high tone except in initial syllables, as in (9) and (10). However, even in NS, Head-H is active in some corners of the grammar. Toneless input forms display initial stress and receive a H tone in the stressed syllable, as in (10). This indicates that although Head-H ranks too low in NS to consistently enforce tone–stress alignment, it lures inserted highs towards foot-heading syllables by virtue of the emergence of the unmarked (McCarthy & Prince 1994).

The independent effect of FtBin in BCMS is somewhat difficult to evaluate outside the context of prosodic minimality. However, Werle (2009: 74–77; 211) demonstrates that FtBin is active in two other areas of BCMS prosodic phonology: pitch accent assignment and the prosodic patterning of prepositions. In addition, stress–tone adjacency in NS provides indirect evidence for FtBin. As argued

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

in §2.1.1, stress–tone adjacency in NS is the result of the dialect's preference for disyllabic trochees (Bethin 1998), which are formed in response to FtBin.[10]

In sum, the minimal foot requirement in BCMS arises through the interaction of two markedness constraints: Head-H and FtBin. Both constraints are independently motivated in the language.

3.2. OT analysis

3.2.1. Monosyllabic lengthening through local constraint conjunction. BCMS tolerates feet with a toneless head mora, as in (33) and (39). In bimoraic feet, toneless head moras cannot acquire a high by tone insertion or TF. The faithful realisation of toneless foot heads in this context suggests that Head-H is individually overridden by the relevant faithfulness constraints, including Dep-H, which prohibits the insertion of a H tone, and NoFlop-tone (Alderete 2001), which militates against TF. The rankings are shown in the tableaux in (43) and (44).

(43)

/ʊuuk/	Dep-H	Head-H
☞ a. ˈ(ʊuuk)		*
b. ˈ(ʊúuk)	*!	

(44)

/ruu.ká/	NoFlop-H	Dep-H	Head-H
a. ruu.ˈ(ká)		*!	
b. ˈ(rúu).ka	*!		
☞ c. ˈ(ruu).ká			*

BCMS likewise permits monomoraic feet, as in (32). This fact suggests that FtBin is dominated by the anti-lengthening faithfulness constraint Dep-μ, as demonstrated in (45).

(45)

/brát/	Dep-μ	FtBin
☞ a. ˈ(brát)		*
b. ˈ(bráat)	*!	

When violated simultaneously, Head-H and FtBin gang up and override a higher-ranking faithfulness constraint: Dep-μ in the case of ML and PL, and NoFlop-H in the case of TF. Classical OT is incapable of deriving gang effects without nonstandard devices (Pater 2009a, b, 2016). The failed OT derivation of ML is provided in (46).

(46) *Failed OT analysis of ML*

/lɛd/	Dep-H	Dep-μ	Head-H	FtBin
💣 a. ˈ(lɛd)			*	*
b. ˈ(lɛ́d)	*!			*
c. ˈ(lɛ́ɛd)	*!	*		
☹ d. ˈ(lɛɛd)		*!	*	

OT accommodates ganging-up cumulativity using LCC (see §1), which incurs a violation whenever all conjuncts are violated within a designated local domain (syllable, foot or the like). To capture the observed gang effect in BCMS, I introduce the local conjunction of Head-H and FtBin at the foot level. The LCC analysis of vowel lengthening in toneless monosyllables is outlined in (47).

[10]See Milenković (2023) for an OT account of tone–stress adjacency in NS.

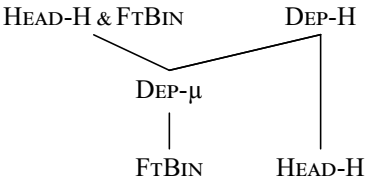
https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

(47)

/lɛd/	HEAD-H & FTBIN	DEP-H	DEP-μ	HEAD-H	FTBIN
a. ˈ(lɛd)	∗!			∗	∗
b. ˈ(lɛ́d)		∗!			∗
c. ˈ(lɛ́ɛd)		∗!	∗		
☞ d. ˈ(lɛɛd)			∗	∗	

The candidate that obeys DEP-μ wins out when FTBIN is violated on its own. This is the case in (45), where the faithful candidate (45a) [ˈbrát] violates FTBIN but obeys HEAD-H. Faithful realisation is preferred because there is no simultaneous violation of HEAD-H and FTBIN. However, when HEAD-H and FTBIN are jointly violated, they gang up against DEP-μ as in (47). There are two ways to avoid the coincident violation of HEAD-H and FTBIN: vowel lengthening (candidate (47d)) or high insertion (candidate (47b)). Vowel lengthening is chosen over tone insertion because DEP-H outranks DEP-μ (see also Zec 1999: 245, especially tableau (57)). The constraint grammar that derives ML is diagrammed in (48).

(48)



HEAD-H alone cannot decide between the winner (47d) [ˈlɛɛd] and the faithful loser (47a) *[ˈlɛd], since both candidates violate it. However, the conjoined constraint HEAD-H & FTBIN favours unfaithful mapping over faithful realisation. The winning candidate repairs only one of the two markedness violations incurred by its faithful competitor. Nevertheless, the partial repair offered by the lengthened form is sufficient to satisfy HEAD-H & FTBIN. This renders vowel lengthening optimal for toneless monomoraic inputs. Thus, the ML process is driven jointly by HEAD-H and FTBIN. In strict-ranking OT, this cumulative constraint interaction is captured by LCC, as (47) demonstrates.

3.2.2. Tonal flop and penultimate lengthening: constraint-based analysis. The analysis proposed in §3.2.1 readily extends to the processes of TF and PL observed in some OS dialects of BCMS. These are argued in §2.5 to have the same purpose as ML: the elimination of toneless degenerate feet.

TF targets stressed toneless *lights* as in (37), but not stressed toneless heavies (35). The absence of TF in cases like /ruu.ká/ → [ˈ(ruu).ká] (not *[ˈ(rúu).ka]) ‘arm.NOM.SG’ in (35a) indicates that HEAD-H cannot be the sole driving force behind TF, else it would apply across the board. Judging from the absence of TF in (35), HEAD-H must be individually overridden by NOFLOP-TONE, as in (49).

(49) NOFLOP ≫ HEAD-H

/ruu.ká/	NONFIN	NOFLOP	HEAD-H
☞ a. ˈ(ruu).ká			∗
b. ruu.ˈ(ká)	∗!		
c. ˈ(rúu).ka		∗!	

The constraint grammar in (49) predicts that TF will be inhibited in all environments in OS. This is problematic because stressed toneless lights attract tone from a final light syllable in some OS dialects, as (37) shows.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

In (50), I outline a failed attempt to derive the TF pattern in (37) under the constraint rankings from (49). This grammar erroneously favours the faithful candidate with penultimate stress over the intended winner that undergoes TF. Note that I ignore the issue of foot structure in the candidate forms in (50).

(50)

/vɔ.dá/	NONFIN	NOFLOP	HEAD-H
a. vɔ.ˈdá	*!		
💣 b. ˈvɔ.dá			*
☹ c. ˈvɔ́.da		*!	

Consequently, the mappings in (49) and (50) appear to be inconsistent with each other. Example (49) points to NOFLOP ≫ HEAD-H. However, this ranking fails to derive the correct winner in (50).

Looping back to the form [ˈvɔ.dá] in (50), there are two structural analyses consistent with it. One possibility is outlined in (51a): the form displays a bimoraic trochee with a toneless head mora and a high-toned non-head mora. Alternatively, in (51b), the final high-toned light is unfooted and the penult forms a toneless degenerate foot.

(51) *Structural analyses of* [ˈvɔ.dá]

a. [(ˈvɔ.dá)]

b. [ˈ(vɔ).dá]

The structure in (51a) satisfies FTBIN at the expense of *NONHEAD-H, which requires that the non-head mora of a foot not be high-toned. As argued in §2.1.2, high-toned foot non-heads are prohibited in OS. This prohibition manifests itself in the dialect's ban on rising contours, but also in the illicitness of (51a). Accordingly, it is better for high-toned moras in OS to be left unfooted than to be included in the non-head position of a foot.[11] This preference in OS is consistent with de Lacy's (2002) theory of tone–stress interaction, where *NONHEAD-H penalises high-toned foot non-heads and no constraint punishes unfooted high-toned moras or syllables.

In (51b), adherence to *NONHEAD-H results in a toneless monomoraic foot. This structure is penalised by the conjoined constraint HEAD-H&FTBIN. Therefore, both faithful candidates that exhibit penultimate stress violate some top-ranking constraint in OS. The bimoraic foot in (51a) is ruled out by *NONHEAD-H. The form in (51b) fatally violates HEAD-H & FTBIN. TF offers a way to satisfy both of these constraints. This process makes the stressed penult high-toned, satisfying HEAD-H, and renders the final light syllable toneless. This enables the inclusion of the final syllable in the non-head position of a bimoraic trochee, which benefits FTBIN without offending *NONHEAD-H. By so doing, the TF candidate [(ˈvɔ́.da)] outperforms all faithful competitors, as shown in (52). TF is preferred over vowel lengthening because DEP-μ outranks NOFLOP-TONE.

(52) *TF*

/vɔ.dá/	HEAD-H & FTBIN	*NONHEAD-H	NONFIN	DEP-μ	NOFLOP	HEAD-H	FTBIN
a. vɔ.ˈ(dá)			*!				*
b. (ˈvɔ.dá)		*!				*	
c. ˈ(vɔ).dá	*!					*	*
d. ˈ(vɔɔ).dá				*!		*	
☞ e. (ˈvɔ́.da)					*		

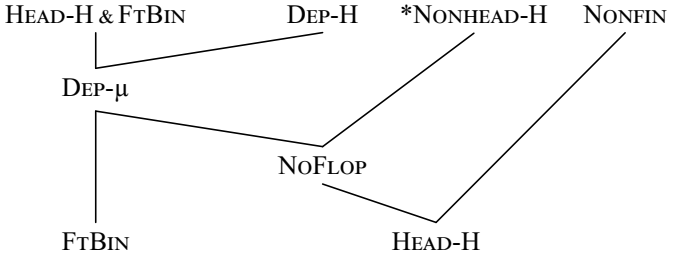
[11] The behaviour of high-toned moras in OS parallels the pattern observed in Ayutla Mixtec (de Lacy 2002: 11), where high-toned syllables are left unfooted to avoid high tone on the non-head syllable of a disyllabic foot.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

In sum, NoFlop cannot be violated in OS to comply with Head-H alone, as seen in (49). However, this constraint gives way to the local conjunction of Head-H and FtBin in (52).

OS dialects that exhibit TF likewise display vowel lengthening in toneless monomoraic words, as in /lɛd/ → [ˈ(lɛɛd)] 'ice.NOM.SG'. The full constraint hierarchy for the subgroup of OS dialects that display TF is diagrammed in (53).

(53) *Constraint hierarchy for OS dialects with TF*



The constraint hierarchy in (53) correctly derives both ML and TF. Since Dep-μ dominates NoFlop, TF is preferred to vowel lengthening in polysyllabic forms, as in (52). Vowel lengthening is the optimal strategy to eliminate toneless degenerate feet in monomoraic forms because Dep-H outranks Dep-μ. Importantly, all crucial rankings relevant for ML in (48) are included in (53).

The LCC analysis devised in (52) carries over to PL. This lengthening process makes the foot bimoraic, satisfying FtBin. However, FtBin cannot be the sole PL-inducing constraint. OS dialects tolerate monomoraic feet, as long as they are high-toned, as in (54). This indicates that in OS, FtBin ranks below Dep-μ.

(54) *Monomoraic feet tolerable in OS:* Dep-μ ≫ FtBin

a.	/brát/	'brother.NOM.SG'	[ˈ(brát)]	not *[ˈ(bráat)]
b.	/kɔ́.maa.ta/	'piece.GEN.SG'	[ˈ(kɔ́).maa.ta]	not *[ˈ(kɔ́ɔ).maa.ta]

Thus, PL targets toneless, but not high-toned, stressed lights. The process does not militate against all monomoraic feet, but rather *toneless* monomoraic feet.

Importantly, all OS dialects display the metrical requirement that no foot be toneless and monomoraic simultaneously. All OS dialects employ ML in toneless monomoraic forms to avoid this doubly marked structure. However, individual OS dialects employ different strategies to achieve this goal in disyllabic and polysyllabic forms. In some OS dialects, this is done by means of TF. These dialects rank Dep-μ above NoFlop, as shown in (52). Conversely, PL arises in those OS dialects that rank NoFlop above Dep-μ. The analysis is provided in (55).

(55) *PL*

/ʊɔ.dá/	Head-H & FtBin	*Nonhead-H	Nonfin	NoFlop	Dep-μ	Head-H	FtBin
a. ʊɔ.ˈ(dá)			*!				*
b. (ˈʊɔ.dá)		*!				*	
c. ˈ(ʊɔ).dá	*!					*	*
d. (ˈʊɔ́.da)				*!			
☞ e. ˈ(ʊɔɔ).dá					*	*	

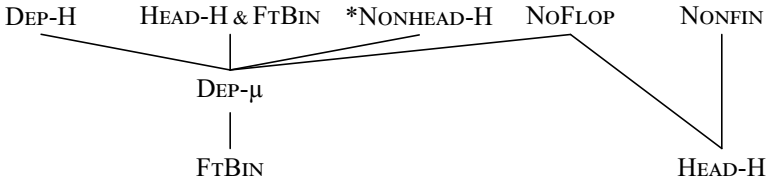
As in (52), final stress (candidate (55a)) is ruled out by Nonfinality. Candidate (55b) is knocked out by *Nonhead-H. Candidate (55c) fatally violates Head-H & FtBin. The TF candidate, (55d), violates

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

NoFlop, which is undominated in this subgroup of OS dialects. The winning candidate, (55e), satisfies all top-ranking constraints at the expense of Dep-μ.

The sole difference between OS dialects that display PL and those that display TF lies in the ranking of Dep-μ relative to NoFlop. The constraint hierarchy for the OS dialects that exhibit PL is given in (56).

(56) *Constraint hierarchy for OS dialects with PL*



In (56), NoFlop ≫ Dep-μ ensures that PL is preferred to TF in polysyllables. Similarly, Dep-H ≫ Dep-μ ensures that ML is preferred to tone insertion in toneless monomoraic words. The constraint grammar in (56) retains all the constraint relations from (48) needed to derive ML.

Taking stock, I identified two subgroups of OS dialects. In one subgroup, which has the grammar in (52), there are two competing strategies to avoid toneless degenerate feet. In monosyllables, vowel lengthening is the optimal strategy to do away with a toneless degenerate foot. In polysyllables, TF takes precedence over vowel lengthening, ensured by the constraint hierarchy in (53). The other subgroup, exemplified by (55), consistently employs vowel lengthening to eliminate toneless degenerate feet, both in monosyllables and in polysyllabic forms. This preference falls out from the constraint hierarchy in (56).

I also provided a unified account of three superficially distinct metrical processes in BCMS: ML, TF and PL. These have not been considered functionally related in previous work. These processes take part in a conspiracy against toneless degenerate feet in the language. To capture the prohibition on toneless degenerate feet in BCMS, I introduced the local conjunction of Head-H and FtBin.

3.3. HG analysis

This section focuses on how the gang effect observed in this article is modelled in HG. Unlike OT, HG is capable of modelling gang effects without locally conjoined constraints (Pater 2009a, b, 2016; Farris-Trimble 2008; Potts *et al.* 2010). However, the picture is more nuanced than this. Both OT and HG need conjoined constraints to model superadditive ganging-up, whereby the effect of simultaneously violating two constraints exceeds the combination of effects of violating them individually (Albright 2009; Green & Davis 2014; Shih 2017). This article lends further support to this view. I show that the illicitness of doubly marked toneless degenerate feet instantiates a superadditive gang effect. In line with this demonstration, the HG analysis of prosodic minimality in BCMS does not dispense with the local conjunction of Head-H and FtBin introduced earlier to model the effect of their coincident violation in OT.

3.3.1. Prosodic minimality in BCMS in standard HG. In the cases of cumulative markedness interaction typically discussed in the HG literature (Pater 2009b, 2016), the repair strategy employed to remove coincident marked configurations eliminates *all* marked structures that are barred from cooccurring within a single domain. At the level of HG analysis, such cumulativity effects are modelled by asymmetric (i.e., one-to-many) trade-offs between constraint weights (Pater 2009b, 2016; Smith 2022). Asymmetric trade-off between faithfulness ($\mathbb{F}$) and markedness constraints $\mathbb{M}_1$ and $\mathbb{M}_2$ arises when the candidate that satisfies *either* $\mathbb{M}_1$ *or* $\mathbb{M}_2$ at the expense of $\mathbb{F}$ loses out to the faithful competitor, while the candidate that satisfies *both* $\mathbb{M}_1$ *and* $\mathbb{M}_2$ at the expense of $\mathbb{F}$ prevails against the faithful competitor.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

This is how standard HG models the restrictions on voiced obstruents in Japanese loanword phonology, an oft-quoted example of ganging-up cumulativity (Nishimura 2003; Kawahara 2006; Pater 2009b, 2016). In Japanese, multiple voiced obstruents are tolerated within a morpheme in loanwords like those in (57), unlike in the native vocabulary (Lyman 1894). Likewise permitted in loanword phonology are otherwise intolerable voiced geminates, as in (58). However, when a voiced geminate co-occurs with a voiced singleton within a morpheme, the geminate optionally devoices, as in (59).

(57) a. [gibu] 'give'
b. [bagi] 'buggy'

(58) a. [webbu] 'web'
b. [heddu] 'head'

(59) a. [gutto] 'good'
b. [dokku] 'dog'

Since multiple voiced singletons are allowed within a morpheme (as in (57)), IDENT-VOICE, which protects the underlying voicing of an obstruent, outweighs OCP-VOICE, which bars multiple voiced obstruents within a morpheme. Similarly, IDENT-VOICE outweighs NOVOICEDGEMINATE, which discriminates against voiced geminates, as evident from (58). Geminate devoicing in (59) is possible because OCP-VOICE and NOVOICEDGEMINATE gang up to overcome higher-weighted IDENT-VOICE.

Crucially, the repair employed to eliminate coincident marked structures in (59) removes *both* relevant marked structures. This explains why it is the voiced geminate rather than the voiced singleton that undergoes devoicing. Singleton devoicing, as in *[toggu], satisfies only OCP-VOICE, but not NOVOICEDGEMINATE. Thus, *[toggu] shares a NOVOICEDGEMINATE violation with the faithful contender [doggu] and violates the individually strongest constraint (IDENT-VOI), as shown in (60).

(60) *Asymmetric trade-off between constraint weights*

/doggu/	$\mathscr{H}$	ID-VOICE 2	OCP-VOICE 1.5	NOVOICEDGEMINATE 1.5
a. doggu	−3.0		−1	−1
☞ b. dokku	−2.0	−1		
c. toggu	−3.5	−1		−1

Notably, in geminate devoicing in Japanese, a single unfaithful mapping satisfies *both* markedness constraints at the expense of a higher-weighted faithfulness constraint. This contrasts with the cumulativity effect in toneless monomoraic feet in BCMS, where no single repair strategy removes both marked structures. Both available repairs (vowel lengthening and tone insertion) remove only one of the coincident marked structures, while the other one persists.

Tone insertion (/lɛd/ → [ˈ(lɛ́d)]) makes the foot-heading mora high-toned, thereby satisfying HEAD-H. However, the foot remains monomoraic, incurring a FTBIN penalty. Vowel lengthening (/lɛd/ → [ˈ(lɛɛd)]) makes the foot bimoraic in response to FTBIN. The head mora remains toneless, violating HEAD-H. The only way to remove both of the coincident marked structures and satisfy both ganging constraints is to subject a toneless monomoraic foot to *both* vowel lengthening and tone insertion: [ˈ(lɛ́ɛd)]. This option violates *two* higher-weighted faithfulness constraints: DEP-H and DEP-μ.

Contrary to the Japanese geminate devoicing in (60), there is a SYMMETRIC (i.e., ONE-TO-ONE) trade-off between constraint weights in the HG analysis of ML in BCMS (61). For the intended winner [ˈ(lɛɛd)] to prevail against the faithful contender [ˈ(lɛd)], a higher-weighted DEP-μ penalty must be traded against a *single* lower-weighted penalty (FTBIN) in HEAD-H-violating contexts. An HG grammar with no means to amplify the severity of coincident constraint violations erroneously selects as optimal the faithful candidate for toneless monomoraic inputs. Due to the shared violation of HEAD-H, faithful realisation is favoured over lengthening regardless of the weight assigned to HEAD-H, as (61) shows.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

(61)

/lɛd/	$\mathscr{H}$	Dep-μ 2	Head-H *n*	FtBin 1.5
a. ˈ(lɛd)	−(1.5+*n*)		−1	−1
☹ b. ˈ(lɛɛd)	−(2.0+*n*)	−1	−1	

In (61), the intended winner (candidate (61b)) satisfies only one of the markedness constraints, namely, FtBin, and still violates Head-H. Since the intended winner and the faithful candidate (61a) share a Head-H violation, vowel lengthening incurs a greater cumulative penalty than no lengthening. Having no means to add weight to the coincident violation of Head-H and FtBin, the standard HG analysis cannot generate a gang effect under the weighting conditions in (61). These weighting conditions are independently motivated by the lack of lengthening in high-toned monosyllables in BCMS (as seen in (32)).

A similar result is observed in the PL process in OS. Under the weighting conditions independently established for OS, vowel lengthening in stressed toneless lights incurs a greater cumulative penalty than a doubly marked toneless monomoraic foot. Because Dep-μ is individually stronger than FtBin, the weight incurred by coincident violation of Dep-μ and Head-H is greater than that incurred by simultaneously violating FtBin and Head-H. Under the standard HG analysis of PL, shown in (62), the intended winner (62e) loses to the faithful candidate (62c), which exhibits a toneless degenerate foot.

(62)

/vɔ.dá/	$\mathscr{H}$	NoFlop 5	Dep-H 4	Nonfin 3	*Nhd-H 3	Head-H 2	Dep-μ 2	FtBin 1.5
a. vɔ.ˈ(dá)	−4.5			−1				−1
b. (ˈvɔ.dá)	−5.0				−1	−1		
c. ˈ(vɔ).dá	−3.5					−1		−1
d. (ˈvɔ́.da)	−5.0	−1						
☹ e. ˈ(vɔɔ).dá	−4.0					−1	−1	

Unlike the lengthening processes in (61) and (62), standard HG is capable of deriving TF. The HG analysis of TF is outlined in the tableau in (63).

(63)

/vɔ.dá/	$\mathscr{H}$	Dep-μ 5	Dep-H 4	Nonfin 3	NoFlop 3	*Nhd-H 2	Head-H 2	FtBin 1.5
a. vɔ.ˈ(dá)	−4.5			−1				−1
b. (ˈvɔ.dá)	−4.0					−1	−1	
c. ˈ(vɔ).dá	−3.5						−1	−1
d. ˈ(vɔɔ).dá	−7.0	−1					−1	
☞ e. (ˈvɔ́.da)	−3.0				−1			

The weighting conditions in (63) are similar to those in (62), the main difference being the fact that Dep-μ outweighs NoFlop in (63). As a result, the TF candidate (63e) emerges as optimal. TF satisfies both Head-H and FtBin in one go, producing an asymmetric trade-off between NoFlop and two lower-weighted markedness constraints: Head-H and FtBin. The HG analysis of TF accounts for this gang effect without any supplementary devices.

To recapitulate: standard HG is incapable of deriving tone-sensitive vowel lengthening in BCMS, wherein toneless monomoraic feet, but not their high-toned counterparts, become bimoraic to comply with the language's foot minimality condition. Under the independently motivated weighting conditions

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

in BCMS, the HG analysis of the relevant lengthening processes fails to replicate the effect of constraint conjunction in OT; see (47), (52) and (55). This suggests that HG's built-in additive cumulativity of constraint violations does not supplant all effects of conjoined constraints, which calls for the enrichment of standard HG's apparatus.

3.3.2. Weighted constraint conjunction in HG. The effect of violating DEP-μ in BCMS is more severe than the effect of violating FTBIN individually. This is apparent from cases like (32), where high-toned monomoraic feet are unable to undergo vowel lengthening. The effect of coincident violation of DEP-μ and HEAD-H is expected to exceed the effect of violating FTBIN and HEAD-H simultaneously, as the HG analysis in (61) and (62) encapsulates. The reasoning behind this expectation is schematised in (64).

(64) a. e(DEP-μ) > e(FTBIN) (Observed)
b. e(DEP-μ, HEAD-H) > e(FTBIN, HEAD-H) (Expected)

However, what transpires in BCMS is inverse to (64b): the joint violation of HEAD-H and FTBIN has a more detrimental effect on metrical well-formedness than the coincident violation of DEP-μ and HEAD-H, as (65) shows:

(65) e(FTBIN, HEAD-H) > e(DEP-μ, HEAD-H) (Observed)

For (64a) and (65) both to be true, it must be the case that in doubly violating contexts, HEAD-H and FTBIN add up to more than the combination of their independent contributions. In other words, the effect of the joint violation of HEAD-H and FTBIN is greater than the combination of the effects of these constraints' independent violations. Thus, the cumulative effect of HEAD-H and FTBIN violations in BCMS is superadditive rather than additive, based on the criteria defined in (3):

(66) *Superadditivity of coincident FTBIN and HEAD-H violation*
e(FTBIN, HEAD-H) > e(FTBIN) + e(HEAD-H)

Following Green & Davis (2014) and Shih (2017), the HG analysis of prosodic minimality in BCMS must adopt the weighted local conjunction of HEAD-H and FTBIN to accommodate the exacerbated severity of their joint violation, as in (66). The presence of this conjoined constraint in the HG grammar of BCMS adds weight to the coincident violation of HEAD-H and FTBIN. The HG account of ML is provided in (67). The extra weight attributed to the doubly marked candidate (67a) enables the lengthening candidate (67b) to prevail against the faithful competitor.[12]

(67) *Monosyllabic lengthening in HG*

/lɛd/	$\mathscr{H}$	DEP-μ 2	HEAD-H n	HEAD-H & FTBIN 2	FTBIN 1.5
a. ˈ(lɛd)	−(3.5+n)		−1	−1	−1
☞ b. ˈ(lɛɛd)	−(2.0+n)	−1	−1		

The weighted LCC analysis in (67) readily extends to PL in OS dialects of BCMS, shown in (62) to pose a challenge to standard HG. With the added weight incurred by the joint violation of HEAD-H and

[12] As the Associate Editor suggests, although both OT and HG need the local conjunction of HEAD-H and FTBIN to capture their interaction in BCMS, the way this conjoined constraint achieves superadditivity is different across the two frameworks. For superadditivity to arise in strict-ranking OT, the conjunction must outrank both conjuncts, as well as the constraint that the conjuncts gang against. This is the case in (47), where HEAD-H & FTBIN outranks not only HEAD-H and FTBIN, but also DEP-μ. In HG, superadditivity arises by the very presence of HEAD-H & FTBIN in the grammar, as long as this conjoined constraint is assigned non-zero weight. This follows from the cumulativity of constraint violations: the penalty incurred by the individual conjuncts counts towards harmony, and the conjoined constraint makes this cumulative penalty more severe.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

FtBin (candidate (68c)), subjecting the toneless light penult to vowel lengthening (candidate (68e)) emerges as the least costly metrical strategy under the weighting conditions shown:

(68)

/vɔ.dá/	$\mathscr{H}$	NoFlop 5	Dep-H 4	Nonfin 3	*Nhd-H 3	Hd-H 2	Head-H & FtBin 2	Dep-μ 2	FtBin 1.5
a. vɔ.ˈ(dá)	−4.5			−1					−1
b. (ˈvɔ.dá)	−5.0				−1	−1			
c. ˈ(vɔ).dá	−5.5					−1	−1		−1
d. (ˈvɔ́.da)	−5.0	−1							
☞ e. ˈ(vɔɔ).dá	−4.0					−1		−1	

While HG can derive TF in isolation without the need for LCC (as shown in (63)), this does not imply that OS dialects displaying TF do not require LCC for other relevant processes. Notably, all OS dialects with TF also exhibit ML in toneless monomoraic forms. Given that standard HG fails to account for ML (see (61)), the local conjunction of Head-H and FtBin is necessary even in OS dialects with TF to capture the full range of foot minimality effects.

Therefore, the formal account of TF is fully compatible with that of ML: both processes can be captured under a single constraint grammar enriched with Head-H&FtBin, under the weighting conditions presented in (69). This is shown for TF in (69a). The same grammar generates ML in toneless monomoraic forms, as in (69b), and no lengthening in their high-toned counterparts, as in (69c).

(69) *OS dialects with TF*

a.

/vɔ.dá/	$\mathscr{H}$	Dep-μ 5	Dep-H 4	Nonfin 3	NoFlop 3	*Nhd-H 2	Head-H 2	Head-H & FtBin 2	FtBin 1.5
a. vɔ.ˈ(dá)	−4.5			−1					−1
b. (ˈvɔ.dá)	−4.0					−1	−1		
c. ˈ(vɔ).dá	−5.5						−1	−1	−1
d. ˈ(vɔɔ).dá	−7.0	−1					−1		
☞ e. (ˈvɔ́.da)	−3.0				−1				

b.

/lɛd/	$\mathscr{H}$	Dep-μ 5	Dep-H 4	Nonfin 3	NoFlop 3	*Nhd-H 2	Head-H 2	Head-H & FtBin 2	FtBin 1.5
a. ˈ(lɛd)	−8.5			−1			−1	−1	−1
☞ b. ˈ(lɛɛd)	−7.0	−1					−1		
c. ˈ(lɛ́d)	−8.5		−1	−1					−1
d. ˈ(lɛ́ɛd)	−9.0	−1	−1						

c.

/brát/	$\mathscr{H}$	Dep-μ 5	Dep-H 4	Nonfin 3	NoFlop 3	*Nhd-H 2	Head-H 2	Head-H & FtBin 2	FtBin 1.5
☞ a. ˈ(brát)	−4.5			−1					−1
b. ˈ(bráat)	−5.0	−1							

In conclusion, the prosodic minimality effects in BCMS arise as a result of the interaction of Head-H and FtBin. The cumulative effect of Head-H and FtBin violations was shown to be superadditive. That being the case, the HG analysis of the processes that conspire against toneless degenerate feet

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

necessitates the local conjunction of HEAD-H and FTBIN – just like the equivalent OT analysis advanced in §3.2.

OT and HG thus end up requiring the same machinery to model the interaction of HEAD-H and FTBIN in BCMS. This demonstration contributes to the OT–HG debate by providing further evidence for conjoined constraints in weighted constraint grammar (Albright 2009; Hayes *et al.* 2012; Green & Davis 2014; Shih 2017), contrary to the widespread assumption that HG obviates the need for conjoined constraints (Farris-Trimble 2008; Pater 2009a, b, 2016; Potts *et al.* 2010).

However, this study, considered in isolation, is indeterminate between OT and HG as a theory of ganging-up cumulativity, though other work has provided more conclusive evidence relevant to the OT–HG debate (see Zuraw & Hayes 2017; Breiss 2020; Smith & Pater 2020, among others, for the current state of this debate). Importantly, the two frameworks invoke conjoined constraints for different purposes. OT avails itself of conjoined constraints to model all cases of ganging-up cumulativity, treating additive and superadditive ganging as equally complex in terms of the formal machinery they necessitate. By contrast, HG needs this supplementary device only for modelling superadditive ganging-up. HG thus treats superadditive ganging as more complex than additive ganging-up.

3.4. No adverse effects of constraint conjunction

The main contribution of HEAD-H & FTBIN is that it captures prosodic minimality in BCMS, which cannot be modelled with non-conjoined constraints only (§§3.2 and 3.3). This section investigates whether this conjoined constraint has any adverse effect on BCMS prosody beyond those corners of BCMS prosodic grammar that it was introduced to model. I algorithmically checked whether HEAD-H & FTBIN is consistent with the rest of the prosodic grammar of BCMS. The survey focused on OS, given that this dialect group displays more prosodic diversity than NS and employs all three processes discussed in this article: ML, TF and PL.

The simulations were performed using the OT-Help software (Staubs *et al.* 2010), which uses the Recursive Constraint Demotion (RCD) algorithm (Tesar & Smolensky 2000) to determine possible optima and calculate typology. To assess how much extra power HEAD-H & FTBIN adds to the phonological grammar of BCMS, I compared two constraint models: a baseline model, which included eight independently motivated constraints (listed in Table 1 in the Supplementary Material), and the LCC model, which included the eight baseline constraints and the local conjunction of HEAD-H and FTBIN.[13]

The list of prosodic grammars generated by the OT-Help software was exported and subjected to further scrutiny using the R statistical programming environment (R Core Team 2021). An R script was created to aid in the comparison of the two constraint sets' predictions, and in the identification of input–output mappings inconsistent with the prosodic grammars of OS dialects.

The results showed that the baseline model could not generate the ML pattern under any ranking or weighting of constraints. By contrast, the model enriched with HEAD-H & FTBIN generated 13 grammars that display this pattern. These grammars are listed in Table 3 in the Supplementary Material.

Out of these 13 grammars, eight matched known OS dialects. Notably, the remaining five grammars were not brought about by HEAD-H & FTBIN. For example, 4 of these 5 unattested grammars showed tonal flop to stressed heavy syllables: /ruu.ká/ → ['(rúu).ka]. This pattern is not observed in OS dialects and can also be derived in the baseline model with the ranking HEAD-H ≫ NOFLOP. Similarly, the remaining unattested pattern (row 9 in Table 3 in the Supplementary Material) represents an accidental gap: this is a theoretically possible pattern that happens not to occur in the observed OS data. This pattern cannot be attributed to HEAD-H & FTBIN.

In sum, the introduction of HEAD-H & FTBIN achieves its intended purpose without introducing any effects that would not otherwise arise under the non-conjoined constraints independently motivated in BCMS prosody.

[13] The input forms and candidates submitted to the OT-Help software are listed in Table 2 in the Supplementary Material.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

4. Alternatives

4.1. No uneven trochees: Zec (1999)

On Zec's (1999) account, vowel lengthening in toneless monosyllables in BCMS is driven by FOOT-BINARITY, which is assumed to outrank DEP-μ[14]:

(70)

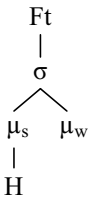
/lɛd/	FTBIN	DEP-μ
a. ˈ(lɛd)	∗!	
☞ b. ˈ(lɛɛd)		∗

To restrict ML to toneless monosyllables (31), Zec adopts Prince's (1990) trochaic harmony scale, which asserts that symmetric trochees are preferred over uneven ones:

(71) *Trochaic harmony* (Prince 1990)
In a rhythmic unit (S W), |S| = |W|,
where |*x*| is the relative prosodic size of *x*.

Bimoraic feet formed by a single heavy syllable are left-headed (Prince 1983; Kager 1993).[15] Accordingly, the illicit lengthened form *[bráat] 'brother' constitutes a moraic trochee, the structure of which is represented in (72).

(72)



Per Zec (1999: 243–247), the structure in (72) is ruled out because it displays an uneven bimoraic foot, since its strong mora bears a H tone and its weak mora is toneless. The form thus fatally violates the size constraint dubbed TROCHAICQUANTITY (TROCHQUANT), which requires equal metrical strength between the head and non-head of a trochee. TROCHQUANT penalises ($\acute{\mu}_s\mu_w$) trochees for the same reason heavy–light trochees are considered disharmonic on Prince's (1990) account. In BCMS, TROCHQUANT is satisfied at the expense of FTBIN, ensuring the faithful realisation of high-toned monomoraic inputs, as in (73).

(73)

/brát/	TROCHQUANT	FTBIN	DEP-μ
☞ a. ˈ(brát)		∗	
b. ˈ(bráat)	∗!		∗

Moraic trochees with a high-toned head position are not categorically prohibited in BCMS, as can be seen in /kráaʎ/ → [ˈ(kráaʎ)] 'king' in (74). This follows from the fact that TROCHQUANT ranks below MAX-H, which protects underlying highs, and below MAX-μ, which militates against vowel shortening.

[14]Notably, ML is not the central concern of Zec (1999). The process, which is discussed in passing, is primarily intended to furnish additional evidence for the foot inventory proposed in the article.

[15]Modulo marginal cases involving diphthongs with ascending sonority, which are considered to form moraic iambs (Kager 1993).

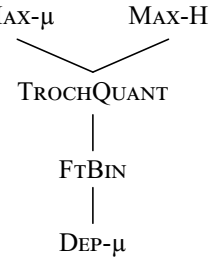
https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

(74)

/kráaʎ/	MAX-H	MAX-μ	TROCHQUANT
☞ a. ˈ(kráaʎ)			*
b. ˈ(kráʎ)		*!	
c. ˈ(kraaʎ)	*!		

Thus, Zec (1999) captures the full range of relevant facts in BCMS monosyllables with the following constraint hierarchy:

(75)



To summarise: Zec's (1999) account ascribes the absence of ML in high-toned monosyllables to the avoidance of the putatively marked ($\acute{\mu}_s\mu_w$) trochees as in (73). Despite being disfavoured by TROCHQUANT, ($\acute{\mu}_s\mu_w$) trochees are found elsewhere in the language, because TROCHQUANT is dominated by faithfulness, as in (74).

The chief difference between Zec's (1999) account and the analysis proposed in this article lies in what is considered the default state in BCMS monosyllables: vowel lengthening or the absence of lengthening. I argue that the absence of ML (as in (32)) is the default. Lengthening is generally inhibited except to eliminate toneless monomoraic feet. Therefore, the dual treatment of toneless and high-toned monosyllables in BCMS instantiates a TRIGGERING (i.e., DO-SOMETHING-ONLY-WHEN) effect (in the sense of Prince & Smolensky [1993] 2004: §3 and §4). Conversely, Zec (1999) holds that vowel lengthening in BCMS monosyllables (as in (31)) is the default. It is only when ML results in a presumedly illicit ($\acute{\mu}_s\mu_w$) trochee that the process is blocked. This instantiates a BLOCKING (i.e., DO-SOMETHING-EXCEPT-WHEN) effect.

4.2. Theoretical and empirical issues with TROCHAICQUANTITY

Zec's (1999) account of the lengthening process in BCMS monosyllables remains uncontested to date in the South Slavic and prosodic literature. The account generates the lengthening pattern in BCMS monosyllables, but at a cost. In what follows, I discuss the theoretical implications and predictions of Zec's account of ML, ultimately rejecting it. §4.2.1 takes up some questions surrounding the theoretical underpinning of this account. §4.2.2 and §4.2.3 show that it suffers from both overgeneration and undergeneration.

4.2.1. TROCHAICQUANTITY and pitch-based rhythmic grouping. Central to Zec's (1999) explanation of why ML fails to apply in high-toned monosyllables in BCMS is the preference for equal grouping in trochaic systems, ensuring that the foot-heading position is not heavier than the weak position (McCarthy & Prince 1986; Prince 1990; Hayes 1995). Zec (1999) extends this requirement to pitch-based grouping. The TROCHQUANT constraint penalises the strong–weak contrast based on relative pitch, ensuring that the strong syllable/mora of a trochee must not be higher-pitched than the weak syllable/mora.

It has been observed that rhythmic groupings with varying intensity are preferably perceived as left-headed, while groupings with a contrast in duration are perceived as right-headed. The rationale behind the TROCHQUANT constraint is that grouping preferences based on pitch parallel those based on

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

duration. There has been plentiful experimental work on rhythmic grouping effects (see Crowhurst 2020 for an overview). Many experimental studies on pitch-based grouping effects report a trochaic bias in alternating sequences with varying pitch (Bion *et al.* 2011; Bhatara *et al.* 2013; de la Mora *et al.* 2013). These results suggest that pitch-based grouping effects differ from duration-based grouping effects, given that higher-pitched elements are preferably associated with group onsets.

Therefore, pitch patterns with intensity rather than duration with respect to rhythmic grouping. The reported trochaic bias in pitch-based grouping poses a challenge for Zec's (1999) account of BCMS ML, which holds that trochees with a higher-pitched first element are disfavoured in much the same way as those with a contrast in duration. Thus, one of the central theoretical underpinnings of Zec's account – the parallelism between quantity-based and tone-based grouping preferences – is contradicted by available experimental evidence. This challenges Zec's assertion that pitch behaves like duration in rhythmic grouping.

4.2.2. TrochaicQuantity penalises doubly unmarked structure. TrochQuant treats bimoraic trochees headed by a high-toned mora as marked. This contradicts two prevalent cross-linguistic tendencies, one being the attraction of high tone to prosodically prominent positions (Goldsmith 1987; Hayes 1995; Yip 2001; de Lacy 2002; Gordon 2023), and the other being the affinity between stress and heavy syllables (Prince 1990; Hayes 1995; Gordon 2006; Ryan 2016). This constraint implies that the combination of two unmarked structures results in a marked configuration: both high-toned foot heads and bimoraic trochees are unmarked individually, but their combination incurs a markedness penalty.

As a result, TrochQuant produces a number of pathological effects. First, consider the constraint grammar in which TrochQuant outranks Max-μ. In this hypothetical language, heavy syllables with a high-toned first mora become light when stressed, as in (76). However, their toneless counterparts surface faithfully, as in (77). In other words, a bimoraic trochee becomes degenerate because its head mora is high-toned.

(76)

/μ́μ/	TrochQuant	Max-H	Max-μ	FtBin
a. ˈ(μ́μ)	∗!			
b. ˈ(μμ)		∗!		
☞ c. ˈ(μ́)			∗	∗

(77)

/μμ/	TrochQuant	Max-μ	FtBin
☞ a. ˈ(μμ)			
b. ˈ(μ)		∗!	∗!

Second, TrochQuant's dominance over Max-H gives rise to a pathological tone deletion pattern whereby high tone deletes in the head position of a bimoraic trochee (as in (78)), but not in the head position of a monomoraic foot (as in (79)), nor in an unstressed syllable (as in (80)):

(78)

/μ́μ/	TrochQuant	Max-μ	Max-H
a. ˈ(μ́μ)	∗!		
b. ˈ(μ́)		∗!	
☞ c. ˈ(μμ)			∗

(79)

/μ́/	TrochQuant	Max-H
☞ a. ˈ(μ́)		
b. ˈ(μ)		∗!

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

(80)

/μ́.μμ/	TrochQuant	FtBin	Max-H
☞ a. μ́.ˈ(μμ)			
b. μ.ˈ(μμ)			∗!
c. ˈ(μ́).μμ		∗!	

Another adverse effect of TrochQuant is illustrated in (81). A H tone is lured away from the foot-heading mora to an unfooted mora to comply with TrochQuant. The pattern is counterintuitive because documented cases of tone shift observe the opposite tendency: highs shift from metrically weak positions to prominent positions (Downing 1990; Bickmore 1995; de Lacy 2002; Breteler 2018).

(81)

/μ.μ́μ/	TrochQuant	Max-H	NoFlop
a. μ.ˈ(μ́μ)	∗!		
b. μ.ˈ(μμ)		∗!	
☞ c. μ́.ˈ(μμ)			∗

Finally, even when TrochQuant ranks below faithfulness and thus fails to induce any unfaithful mappings, it is still capable of producing pathological effects. Suppose there is a language with lexical tone that preferably stresses the leftmost heavy syllable, else defaulting to the leftmost syllable, as in (82). Suppose further that this hypothetical language ranks TrochQuant over WSP. This ranking gives rise to the pathological mapping in (82c): the leftmost heavy, the first mora of which is high-toned, repels stress to the light peninitial syllable for no reason other than having a high-toned first mora.

(82) a. Stress leftmost heavy:
μ.μμ.μ → [μ.ˈμμ.μ]
b. Else leftmost syllable:
μ.μ.μ → [ˈμ.μ.μ]
c. Do not stress leftmost heavy if high-toned:
μ́μ.μ.μ → [μ́μ.ˈμ.μ]

This repulsion effect runs afoul of the propensity of high tone to attract stress (Goldsmith 1987; Gordon 2023).

4.2.3. Undergeneration. Zec (1999) discusses ML without addressing the other processes that conspire against toneless degenerate feet: TF and PL. This section explores whether Zec's account extends to these other functionally related processes in BCMS.

To accomplish this, I conducted a learning simulation using the OT-Help software. The constraint set used in this simulation mirrors the one used in §3.4, the only difference being the inclusion of Zec's TrochQuant in place of the Head-H & FtBin conjunction from §3.4.

The simulation found that Zec's account captures only a subset of the relevant processes in BCMS. Specifically, TrochQuant struggles to model a subgroup of the OS dialects that display TF (as in (37)). The grammar that Zec's constraint set fails to derive is given in (83). These data match the stress pattern of the Kosovo–Resava dialect of BCMS (Ivić 1958; Simić 1972).

(83) *Phonological grammar problematic for TrochQuant*

a. /bɔg/	→	[ˈ(bɔɔg)]	'god.NOM.SG'
b. /brát/	→	[ˈ(brát)]	'brother.NOM.SG'
c. /vɔ.dá/	→	[ˈ(vɔ́).da]	'water.NOM.SG'
d. /ruu.ká/	→	[ˈ(ruu).ká]	'arm.NOM.SG'

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

e. /u.raa.dí.ʃɛ/ → [u.raa.ˈ(dí).ʃɛ] ‘do.AOR.3PL’

f. /sɛ.kí.ra/ → [sɛ.ˈ(kí).ra] ‘axe.NOM.SG’

g. /prɔ.dáaʊ.ni.ʦa/ → [prɔ.ˈ(dáaʊ).ni.ʦa] ‘store.NOM.SG’

The attentive reader will notice that the surface forms in (83c), (83e) and (83f) show monomoraic rather than bimoraic feet (e.g., [ˈ(ʊ́).da] rather than [(ˈʊ́.da)] in (83c)). This adjustment was necessary in order for Zec’s (1999) TROCHQUANT-based account to be made to work, given that TROCHQUANT penalises bimoraic trochees with a high-toned head mora. On the present account, monomoraic feet posited in (83c), (83e) and (83f) would violate FTBIN for no obvious reason since HEAD-H&FTBIN does not favour monomoraic over bimoraic feet. However, for Zec’s account, it is crucial to assume that these OS forms have monomoraic rather than bimoraic feet.

Even when this adjustment in favour of TROCHQUANT is made, the algorithm finds no OT grammar that derives all mappings in (83). The RCD algorithm stalled without reaching the target grammar. Six out of nine constraints were left unranked: HEAD-H, FTBIN-μ, WSP, NOFLOP-H, DEP-μ and TROCHQUANT.

No totally ranked hierarchy was established for the above six constraints because of a ranking inconsistency involving TROCHQUANT. The RCD algorithm demotes DEP-μ below FTBIN on the basis of vowel lengthening in toneless monosyllables (as in (83a)). Since there is no lengthening in high-toned monomoraic forms (see (83b)), the algorithm demotes FTBIN below TROCHQUANT. The subranking TROCHQUANT ≫ FTBIN ≫ DEP-μ is precisely how Zec (1999: 244–245) accounts for vowel lengthening in toneless monosyllables in BCMS, as discussed in §4.1.

The TF mapping in (83c) provides a rationale for demoting NOFLOP below HEAD-H. This is illustrated in the combination tableau in (84). The only way to justify the absence of TF in (83d) is by demoting HEAD-H below TROCHQUANT (as in (85)). Finally, WSP gets demoted below HEAD-H to accommodate the preference for stressed high-toned lights over stressed toneless heavies in the Kosovo–Resava dialect (as in (83e)), illustrated in the tableau in (86).

(84) *Demote NOFLOP below HEAD-H*

/ʊɔ.dá/	HEAD-H	NOFLOP
☞ a. ˈ(ʊ́).da		1
b. ˈ(ʊɔ).dá	W1	L

(85) *Demote HEAD-H below TROCHQUANT*

/ruu.ká/	TROCHQUANT	HEAD-H
☞ a. ˈ(ruu).ká		1
b. ˈ(rúu).ka	W1	L

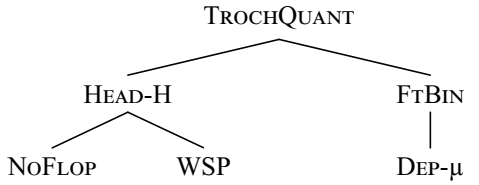
(86) *Demote WSP below HEAD-H*

/u.raa.dí.ʃɛ/	HEAD-H	WSP
☞ a. u.raa.ˈ(dí).ʃɛ		1
b. u.ˈ(raa).dí.ʃɛ	W1	L

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

After these five iterations of constraint demotion, the learner posits the following constraint grammar:

(87)



The grammar in (87) is inconsistent with the mapping /prɔ.dáaʊ.ni.ʦa/ → [prɔ.ˈ(dáaʊ).ni.ʦa] in (83g). Taken in isolation, (83g) requires Head-H to dominate TrochQuant. However, this ranking contradicts the ranking established in (85):

(88) *RCD crashes*

prɔ.dáaʊ.ni.ʦa	TrochQuant	Head-H	FtBin	NoFlop	WSP	Dep-μ
☹ a. prɔ.ˈ(dáaʊ).ni.ʦa	1					
💣 b. prɔ.dáaʊ.(ˈni.ʦa)		1			1	

Thus, the RCD algorithm cannot find a totally ranked hierarchy for the six constraints in (87). Employing TrochQuant to model prosodic minimality in the OS dialects that display TF creates a ranking paradox. The mapping in (88), needed for (83g), cannot be made consistent with the rest of the dialect's phonological grammar in (83).

This undesirable effect of Zec's (1999) TrochQuant is limited to strict-ranking OT. The inconsistency encountered in (88) does not carry over to HG, where Head-H and WSP can gang up to jointly override TrochQuant. This gang effect enables the intended winner [prɔ.ˈ(dáaʊ).ni.ʦa] to prevail against the problematic competitor *[prɔ.dáaʊ.(ˈni.ʦa)] in (88).

4.3. Section summary

I have pointed out that Zec's (1999) account of ML suffers from both overgeneration and undergeneration. The account overgenerates because the TrochQuant constraint predicts a number of pathological effects (§4.2.2). Zec's constraint model likewise fails to generate some of the relevant processes in OS, at least within the bounds of strict-ranking OT (§4.2.3). Further, Zec's account establishes no link between TF, PL and ML, which are evidently functionally related. Moreover, TrochQuant requires unmotivated auxiliary hypotheses, including the preference for monomoraic over bimoraic feet.

None of these adverse effects carry over to the LCC account proposed in this article: it derives all relevant patterns in BCMS (§3.2 and §3.3) and has no additional detrimental effects (§3.4). The ban on toneless degenerate feet falls out from the interaction of two constraints that are independently motivated in BCMS phonology: Head-H and FtBin. Consequently, the present account obviates the need for *ad hoc* constraints *à la* Zec's (1999) TrochQuant to block vowel lengthening in high-toned monomoraic feet or inhibit TF in bimoraic feet. This effectively dispenses with the problematic theoretical assumptions and predictions of Zec's account identified in this section.

5. Conclusion

This article has explored prosodic minimality in BCMS. A key finding is the identification of a relationship between tone and foot size in the language. Individually, BCMS permits both monomoraic feet and feet with a toneless head. However, the language disallows toneless monomoraic feet, the combination of these marked structures.

https://doi.org/10.1017/S0952675726100311 Published online by Cambridge University Press

The present study contributes to our understanding of this complex prosodic pattern in two ways. First, I introduced new data from dialectal BCMS to a wider generative audience. These data show that vowel lengthening in toneless monosyllables, discussed in passing by Zec (1999), is not the sole manifestation of the ban on toneless monomoraic feet. In fact, there is a cross-dialectal conspiracy against this doubly marked structure. Second, this study offers an alternative to the only existing generative account of prosodic minimality in BCMS (Zec 1999), which makes flawed typological predictions and fails to account for the full range of minimality effects in BCMS in a unified fashion.

I attributed the ban on toneless degenerate feet to the joint effect of two markedness constraints independently motivated in the phonology of BCMS. These constraints include HEAD-H, which requires that foot-heading moras be high-toned, and FTBIN, which penalises monomoraic feet. I further demonstrated that the effect of the coincident violation of these two constraints is superadditive. I introduced the local conjunction of these constraints to capture the exacerbated severity of their joint violation in BCMS. The proposed conjoined constraint is necessary in both strict-ranking OT and, importantly, HG, which is often argued to fully dispense with conjoined constraints (Farris-Trimble 2008; Pater 2009b; Potts *et al.* 2010).

The main takeaways of the study are the following:

1. This study identifies a case of cumulative markedness interaction in BCMS whereby the combination of two individually tolerable marked structures is categorically ruled out. This pattern was shown to be superlinear: the effect of simultaneous violation of the relevant markedness constraints goes beyond the combination of their independent effects. This demonstration provides support for the use of locally conjoined constraints as a means of modelling superlinear ganging-up cumulativity in weighted constraint grammar (Albright 2009; Hayes *et al.* 2012; Green & Davis 2014; Shih 2017).
2. Virtually all superadditivity effects have been documented in variable phonological patterns (Shih 2017; Smith & Pater 2020; Breiss & Albright 2022; Kim 2022). This study identifies a superadditive cumulativity effect in a categorical prosodic pattern, thereby expanding the empirical range of attested superadditivity effects.
3. The identification of a tone-sensitive prosodic minimality effect has implications for prosodic typology. In BCMS, lexical highs not only determine the locus of stress, which is a well-known tendency in languages with tone-driven stress (de Lacy 2002; Gordon 2023), but also take part in defining the minimal size of prosodic constituents.
4. The analysis has the added benefit of unifying three superficially distinct processes in BCMS – ML, tonal flop and penultimate lengthening – which had not previously been seen as related.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017S0952675726100311.

Data availability statement. Replication data and code for this study can be found in Harvard Dataverse: https://doi.org/10.7910/DVN/1V64L2.

Acknowledgements. I would like to thank an anonymous *Phonology* reviewer, reviewer Canaan Breiss and the Associate Editor, all of whose extensive comments led to substantial improvements of the initial submission. The article greatly benefited from feedback by audience members at the Harvard PhonLab meeting, CLS 59, FASL 32 and AMP 2023, especially Michael Becker, Canaan Breiss, Aleksei Nazarov and Paul Smolensky. Special thanks go to Kevin Ryan for continuous support and a careful reading of the manuscript, and to Anabelle Caso for an outstanding job editing the revised manuscript. All remaining errors are solely my responsibility.

Competing interests. The author declares that there are no competing interests regarding the publication of this article.

Ethical standards. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

References

Albright, Adam (2009). Cumulative violations and complexity thresholds: evidence from Lakhota. Paper presented at the 17th Manchester Phonology Meeting, University of Manchester, May 2009.

Albright, Adam (2012). Additive markedness interactions in phonology. Ms, Massachusetts Institute of Technology.

Alderete, John (2001). Dominance effects as transderivational anti-faithfulness. *Phonology* **18**, 201–253.
Baković, Eric (2000). *Harmony, dominance and control*. PhD dissertation, Rutgers University.
Batas, Ana (2014). *Fonetska i akcenatska promenljivost reči u kontinualnom govoru*. PhD dissertation, University of Belgrade.
Bethin, Christina (1994). On the phonology of the Neoštokavian accent retraction in Serbian and Croatian. *Die Welt der Slaven* **39**, 277–296.
Bethin, Christina (1998). *Slavic prosody: language change and phonological theory*. Cambridge: Cambridge University Press.
Bhatara, Anjali, Natalie Boll-Avetisyan, Annika Unger, Thierry Nazzi & Barbara Höhle (2013). Native language affects rhythmic grouping of speech. *JASA* **134**, 3828–3843.
Bickmore, Lee (1995). Tone and stress in Lamba. *Phonology* **12**, 307–341.
Bion, Ricardo, Silvia Benavides-Varela & Marina Nespor (2011). Acoustic markers of prominence influence infants' and adults' segmentation of speech sequences. *Language and Speech* **54**, 123–140.
Breiss, Canaan (2020). Constraint cumulativity in phonotactics: evidence from artificial grammar learning studies. *Phonology* **37**, 551–576.
Breiss, Canaan & Adam Albright (2022). Cumulative markedness effects and (non-)linearity in phonotactics. *Glossa* **7**, 32 pp.
Breteler, Jeroen (2018). *A foot-based typology of tonal reassociation: perspectives from synchrony and learnability*. PhD dissertation, University of Amsterdam.
Crowhurst, Megan (2020). The iambic/trochaic law: nature or nurture? *Language and Linguistics Compass* **14**, article no. e12360.
Downing, Laura (1990). Local and metrical tone shift in Nguni. *Studies in African Linguistics* **21**, 261–318.
Downing, Laura (1998). On the prosodic misalignment of onsetless syllables. *NLLT* **16**, 1–52.
Farris-Trimble, Ashley (2008). *Cumulative faithfulness effects in phonology*. PhD dissertation, Indiana University.
Garrett, Edward (1999). Minimal words aren't minimal feet. *UCLA Working Papers in Linguistics* **1**, 68–105.
Goldsmith, John (1987). Tone and accent, and getting the two together. *BLS* **13**, 88–104.
Gordon, Matthew (2006). *Syllable weight: phonetics, phonology, typology*. New York: Routledge.
Gordon, Matthew (2023). The phonetic basis for tone–stress interactions: a cross-linguistic study. In Jeroen van de Weijer (ed.) *Representing phonological detail, part II: syllable, stress, and sign*. Berlin: Mouton de Gruyter, 171–190.
Green, Christopher & Stuart Davis (2014). Superadditivity and limitations on syllable complexity in Bambara words. In Ashley Farris-Trimble & Jessica Barlow (eds.) *Perspectives on phonological theory and development, in honor of Daniel A. Dinnsen*. Amsterdam: Benjamins, 223–247.
Hayes, Bruce (1995). *Metrical stress theory: principles and case studies*. Chicago: University of Chicago Press.
Hayes, Bruce, Colin Wilson & Anne Shisko (2012). Maxent grammars for the metrics of Shakespeare and Milton. *Lg* **88**, 691–731.
Hyde, Brett (2007). Non-finality and weight-sensitivity. *Phonology* **24**, 287–334.
Hyman, Larry (2006). Word-prosodic typology. *Phonology* **23**, 225–257.
Inkelas, Sharon & Draga Zec (1988). Serbo-Croatian pitch accent: the interaction of tone, stress, and intonation. *Lg* **64**, 227–248.
Itô, Junko & Armin Mester (1998). *Markedness and word structure: OCP effects in Japanese*. Ms, University of California, Santa Cruz.
Ivić, Pavle (1957). O govoru Galipoljskih Srba. *Srpski dijalektološki zbornik* **12**, 1–519.
Ivić, Pavle (1958). *Die serbokroatischen Dialekte: ihre Struktur und Entwicklung*. The Hague: Mouton.
Jäger, Gerhard & Anette Rosenbach (2006). The winner takes it all—almost: cumulativity in grammatical variation. *Linguistics* **44**, 937–971.
Jović, Dušan (1968). Trstenički govor. *Srpski dijalektološki zbornik* **17**, 1–239.
Kager, René (1993). The moraic iamb. *CLS* **27**, 291–305.
Kapović, Mate (2015). *Povijest hrvatske akcentuacije: fonetika*. Zagreb: Matica hrvatska.
Kawahara, Shigeto (2006). A faithfulness ranking projected from a perceptibility scale: the case of [+voice] in Japanese. *Lg* **82**(3), 536–574.
Kawahara, Shigeto (2020). A wug-shaped curve in sound symbolism: the case of Japanese Pokémon names. *Phonology* **37**, 383–418.
Kawahara, Shigeto & Canaan Breiss (2021). Exploring the nature of cumulativity in sound symbolism: experimental studies of Pokémonastics with English speakers. *Laboratory Phonology* **12**, article no. 3, 29 pp.
Kawahara, Shigeto & Gakuji Kumagai (2021). What voiced obstruents symbolically represent in Japanese: evidence from the Pokémon universe. *Journal of Japanese Linguistics* **37**, 3–24.
Kenstowicz, Michael (2009). Two notes on Kinande vowel harmony. *Language Sciences* **31**, 248–270.
Kim, Seoyoung (2022). A MaxEnt learner for super-additive counting cumulativity. *Glossa* **7**, 34 pp.
Kirchner, Robert (1997). Synchronic chain shifts in Optimality Theory. *LI* **27**, 341–350.
Kisseberth, Charles (1970). On the functional unity of phonological rules. *LI* **1**, 291–306.
de Lacy, Paul (2002). The interaction of tone and stress in Optimality Theory. *Phonology* **19**, 1–32.
Langston, Keith (1997). Pitch accent in Croatian and Serbian: towards an autosegmental analysis. *Journal of Slavic Linguistics* **5**, 80–116.
Legendre, Géraldine, Yoshiro Miyata & Paul Smolensky (1990). Harmonic Grammar—a formal multi-level connectionist theory of linguistic well-formedness: theoretical foundations. In *Proceedings of the 12th Annual Conference of the Cognitive Science Society*. Hillsdale, NJ: Erlbaum, 388–395.

Lehiste, Ilse & Pavle Ivić (1986). *Word and sentence prosody in Serbocroatian*. Cambridge, MA: MIT Press.
Ligorio, Orsat (2016). *Serbo-Croatian accent retraction: its course and character in the dialect of Dubrovnik*. PhD dissertation, Leiden University.
Łubowicz, Anna (2003). *Contrast preservation in phonological mappings*. PhD dissertation, University of Massachusetts Amherst.
Lyman, Benjamin Smith (1894). The change from surd to sonant in Japanese compounds. In *Oriental studies: a selection of the papers read before the Oriental Club of Philadelphia, 1888–1894*. Boston: Ginn, 160–176.
Martinović, Martina (2012). The interaction of animacy with phonetic and phonological factors in Neoštokavian pitch accents. *WCCFL* **29**, 161–168.
McCarthy, John (2003). Comparative markedness. *Theoretical Linguistics* **29**, 1–51.
McCarthy, John & Alan Prince (1986). *Prosodic morphology*. Ms, University of Massachusetts Amherst and Brandeis University.
McCarthy, John & Alan Prince (1994). The emergence of the unmarked: optimality in prosodic morphology. *NELS* **24**, 333–379.
McPherson, Laura (2016). Cumulativity and ganging in the tonology of Awa suffixes. *Lg* **92**, e38–e66.
Milenković, Aljoša (2023). Towards a more comprehensive theory of tone–stress interaction. *NELS* **53**, 209–218.
de la Mora, Daniela, Marina Nespor & Juan Toro (2013). Do humans and nonhuman animals share the grouping principles of the iambic–trochaic law? *Attention, Perception, & Psychophysics.* **75**, 92–100.
Nishimura, Kohei (2003). *Lyman's Law in loanwords*. Master's thesis, Nagoya University.
Pater, Joe (2009a). Review of: Paul Smolensky and Géraldine Legendre (2006). *The harmonic mind: from neural computation to Optimality-Theoretic grammar.* Cambridge, MA: MIT Press. Vol. 1: *Cognitive architecture.* Pp. xxiv+ 563. Vol. 2: *Linguistic and philosophical implications.* Pp. xxiv+ 611. *Phonology* **26**, 217–226.
Pater, Joe (2009b). Weighted constraints in generative linguistics. *Cognitive Science* **33**, 999–1035.
Pater, Joe (2016). Universal grammar with weighted constraints. In John McCarthy & Joe Pater (eds.) *Harmonic Grammar and Harmonic Serialism*. Bristol, CT: Equinox, 1–46.
Potts, Christopher, Joe Pater, Karen Jesney, Rajesh Bhatt & Michael Becker (2010). Harmonic Grammar with linear programming: from linear systems to linguistic typology. *Phonology* **27**, 77–117.
Prince, Alan (1983). Relating to the grid. *LI* **14**, 19–100.
Prince, Alan (1990). Quantitative consequences of rhythmic organization. *CLS* **26**, 355–398.
Prince, Alan & Paul Smolensky ([1993] 2004). *Optimality Theory: constraint interaction in generative grammar*. Oxford: Blackwell. Originally published in 1993 as technical report no. 2 of the Rutgers Center for Cognitive Science.
R Core Team (2021). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. https://www.R-project.org/.
Remetić, Slobodan (1985). Govori centralne Šumadije. *Srpski dijalektološki zbornik* **31**, 1–555.
Ryan, Kevin (2016). Phonological weight. *Language and Linguistics Compass* **10**, 720–733.
Ryan, Kevin (2017). Attenuated spreading in Sanskrit retroflex harmony. *LI* **48**, 299–340.
Shih, Stephanie (2017). Constraint conjunction in weighted probabilistic grammar. *Phonology* **34**, 243–268.
Simić, Radoje (1972). Levački govor. *Srpski dijalektološki zbornik* **19**, 1–618.
Smith, Jennifer (2022). Some formal implications of deletion saltation. *LI* **53**, 852–864.
Smith, Brian & Joe Pater (2020). French schwa and gradient cumulativity. *Glossa* **5**, article no. 24, 33 pp.
Smolensky, Paul (1993). Harmony, markedness, and phonological activity. Paper presented at Rutgers Optimality Workshop 1, Rutgers University, October 1993.
Smolensky, Paul (2006). Optimality in phonology II: harmonic completeness, local constraint conjunction, and feature domain markedness. In Paul Smolensky & Géraldine Legendre (eds.) *The harmonic mind: from neural computation to Optimality-Theoretic grammar, volume 2: linguistic and philosophical implications*. Cambridge, MA: MIT Press, 27–160.
Staubs, Robert, Michael Becker, Christopher Potts, Patrick Pratt, John McCarthy & Joe Pater (2010). OT-Help 2.0. Software package. Available at https://people.umass.edu/othelp/.
Stevanović, Mihailo (1933). Istočnocrnogorski dijalekat. *Južnoslovenski filolog* **13**, 1–129.
Tesar, Bruce & Paul Smolensky (2000). *Learnability in Optimality Theory*. Cambridge, MA: MIT Press.
Tomanović, Vaso (1935). Akcenat u govoru Lepetana (Boka Kotorska). *Južnoslovenski filolog* **14**, 59–141.
Werle, Adam (2009). *Word, phrase, and clitic prosody in Bosnian, Serbian, and Croatian*. PhD dissertation, University of Massachusetts Amherst.
Yip, Moira (2001). The complex interaction of tone and prominence. *NELS* **31**, 531–545.
Zec, Draga (1999). Footed tones and tonal feet: rhythmic constituency in a pitch-accent language. *Phonology* **16**, 225–264.
Zec, Draga & Elizabeth Zsiga (2010). Interaction of tone and stress in Standard Serbian: phonological and phonetic evidence. In Wayles Browne, Adam Cooper, Alison Fisher, Esra Kesici, Nikola Predolac & Draga Zec (eds.) *Formal Approaches to Slavic Linguistics 18: the second Cornell meeting, 2009*. Ann Arbor, MI: Michigan Slavic Publications, 537–555.
Zec, Draga & Elizabeth Zsiga (2022). Tone and stress as agents of cross-dialectal variation: the case of Serbian. In Haruo Kubozono, Junko Itô & Armin Mester (eds.) *Prosody and prosodic interfaces*. Oxford: Oxford University Press, 63–94.
Zsiga, Elizabeth & Draga Zec (2013). Contextual evidence for the representation of pitch accents in Standard Serbian. *Language and Speech* **56**, 69–104.
Zuraw, Kie & Bruce Hayes (2017). Intersecting constraint families: an argument for harmonic grammar. *Lg* **93**, 497–548.