Learning Tonotactic Patterns over Autosegmental Representations

Li Han

09/29/2024

Overview

1. Tonotactic Learning over autosegmental representations
   a. ARs can be viewed as graphs with labeled nodes and edges (Jardine, 2019; Jardine & Heinz, 2015).
   b. ARs Well-formedness is defined as not containing forbidden subgraph (Jardine, 2017; Rogers and Lambert, 2019).
   c. The banned AR structures in a model theoretic representation can be successfully learned by the BUFIA (Chandlee et al., 2019).

2. Today:
   a. Shows how tonotactic patterns can be learned over ARs using Hausa as a case study
   b. Ran BUFIA on 664 monomorphemic forms (26 distinct ARs) and found 7 constraints (syl, tone ≤ 3)
   c. Compared BUFIA constraints with linguistically-attested constraints (Rule constraint) and see to what extent they match:
      i. A BUFIA constraint = a Rule constraint.
      ii. A BUFIA constrain is more specific than a Rule constraint
      iii. A BUFIA constraint has no corresponding Rule constraint

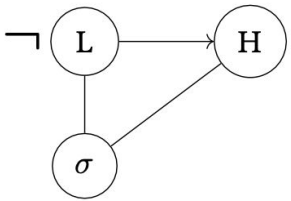
ARs as Graph

1. Jardine (2017) explains how ARs can be expressed as graphs (Coleman and Local, 1991; Jardine and Heinz, 2015)

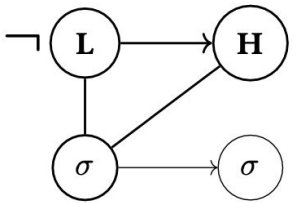
a.	A set of labeled node	elements from the two tiers	(circles)
b.	Directed edges	linear order of tones and TBUs	(lines w arrowhead)
c.	Undirected edges	associative lines between tones and TBUs	(lines w/o arrowhead)

2. A well-formed AR is defined as one not containing any forbidden subgraph.

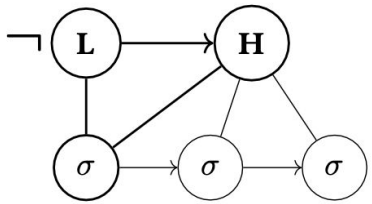
3. Formally, if grammar *G* is the set of constraints such that *G* = *{r1, r2, . . . , rn}*, a well-formed AR = ¬*r1* ⋀ ¬*r2* ⋀ ¬*r3* ⋀ . . . ⋀ ¬ *rn* (Jardine, 2017; Rogers and Lambert, 2019).



(a) *σ́



(b) *σ́σ

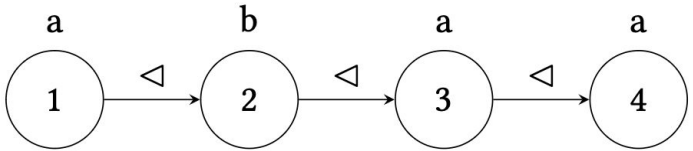


(c) *σ́σ́σ́

A model-theoretic approach

A string can also be represented in a model-theoretic representation, which provides a uniform framework for representing objects and their relations to one another.

Example: Σ = {a, b}, a model-theoretic representation of a string *abaa* would be M = {D | Ra, Rb, ◁}

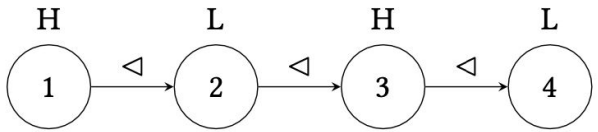


$$\mathcal{M} = \{D \mid R_a, R_b, \lhd\}$$
$$\mathcal{D} = \{1, 2, 3, 4\}$$
$$R_a = \{1, 3, 4\}$$
$$R_b = \{2\}$$
$$\lhd = \{\langle 1, 2\rangle, \langle 2, 3\rangle, \langle 3, 4\rangle\}$$

AR Model

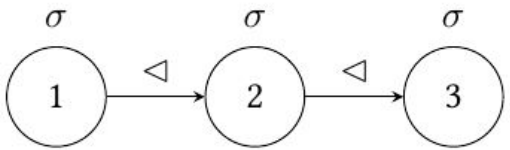
To build an AR model, we need the tonal string ∪ a TBU string ∪ their association

1. Tonal string



$M_t = \{D \mid R_H, R_L, \triangleleft\}$
$D_t = \{1, 2, 3, 4\}$
$R_H = \{1, 3\}$
$R_L = \{2, 4\}$
$\triangleleft = \{\langle 1, 2\rangle, \langle 2, 3\rangle, \langle 3, 4\rangle\}$

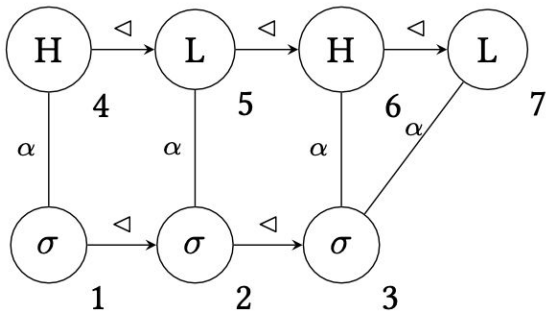
2. TBU string



$M_s = \{D \mid R_\sigma, \triangleleft\}$
$D_s = \{1, 2, 3\}$
$R_\sigma = \{1, 2, 3\}$
$\triangleleft = \{\langle 1, 2\rangle, \langle 2, 3\rangle\}$

AR MODEL

The combined model is the union of two string models along with the binary association relation α



$M = \{D \mid R_H, R_L, R_\sigma, \triangleleft, \alpha\}$
$D = \{1, 2, 3, 4, 5, 6, 7\}$ $= D_t + D_s$
$R_\sigma = \{1, 2, 3\}$
$R_H = \{4, 6\}$
$R_L = \{5, 7\}$
$\triangleleft = \{\langle 1, 2\rangle, \langle 2, 3\rangle, \langle 4, 5\rangle, \langle 5, 6\rangle, \langle 6, 7\rangle\}$ $= \triangleleft_T \cup \triangleleft_S$
$\alpha = \{\langle 1, 4\rangle, \langle 4, 1\rangle, \langle 2, 5\rangle, \langle 5, 2\rangle, \langle 3, 6\rangle, \langle 6, 3\rangle, \langle 3, 7\rangle, \langle 7, 3\rangle\}$

A binary relation indicating tone-TBU associations

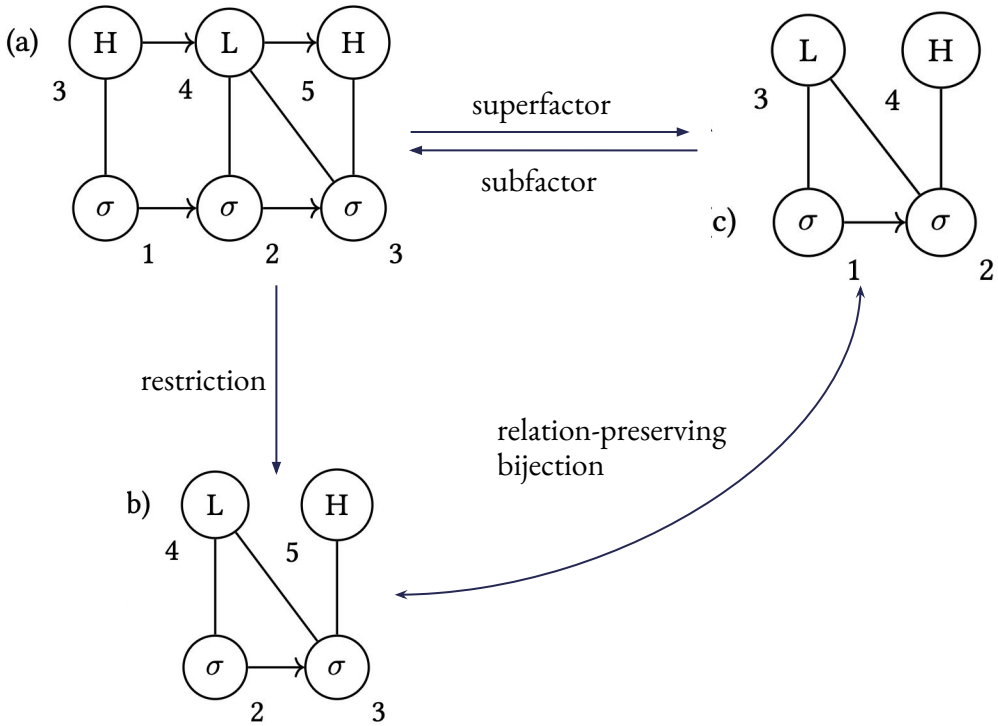
We define the *size* of an AR model as $k = D + |\alpha/2|$. Here, $k = 11$

Containment

Does AR (a) contains (c)?

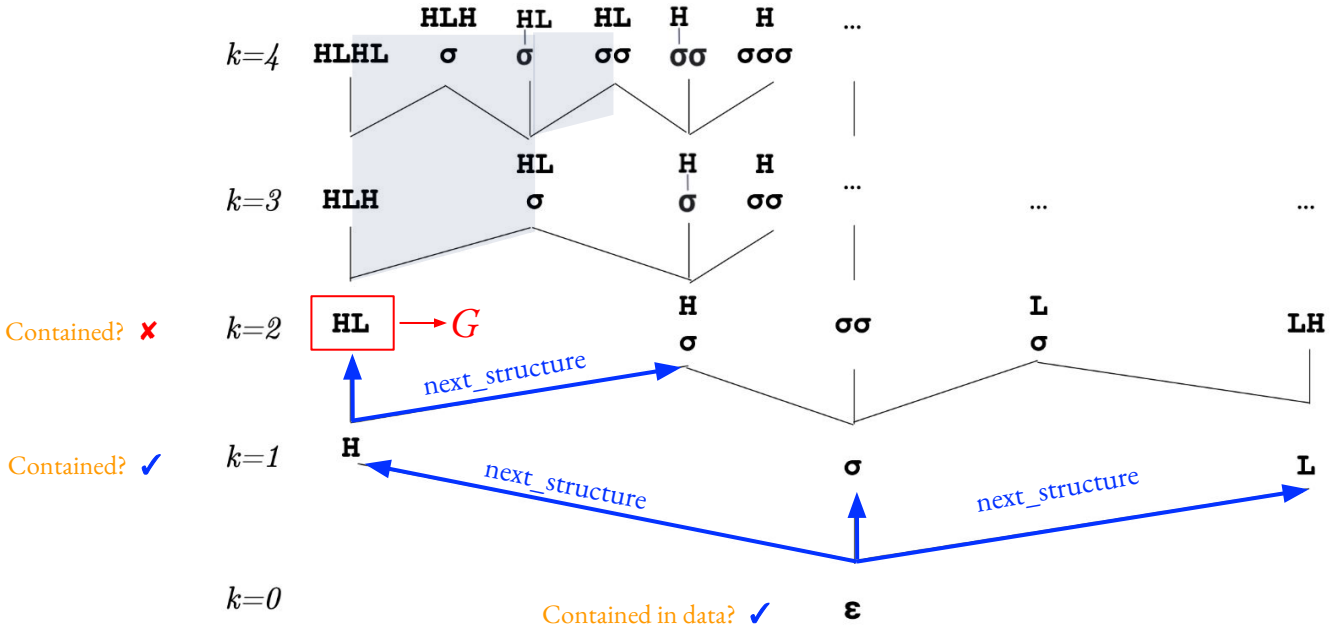
1. Can we find any **restriction** in (a) that holds all same relations in a relative way as in (c)? Chandlee et al. (2019, p.94).

2. If so, (a) is the superfactor of (c), (c) is a subfactor of (a) and a contains (b)



Partial Order of AR Space

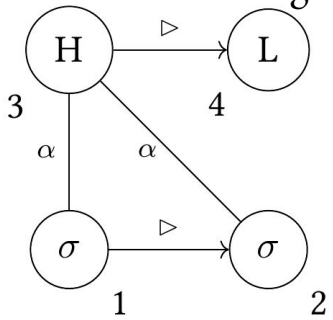
The containment relation yields a partial order over the constraint space



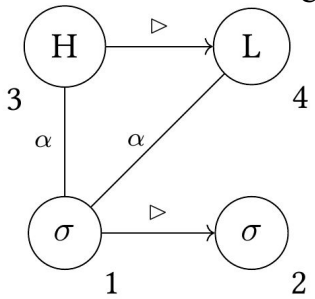
Next Structure

1. When BUFIA identifies the presence of a given AR structure, the NextAR function will generate its immediate superfactors.
2. An AR can be expanded adding a new tone, a new TBU, or a new association depending on whether there are any *floating* units in the AR.

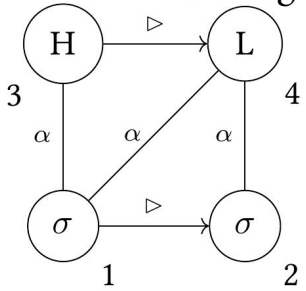
AR with floating tone



AR with floating TBU



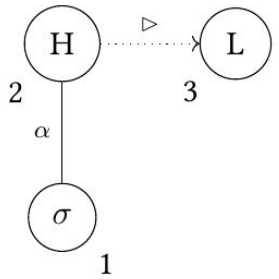
AR w.o floating units



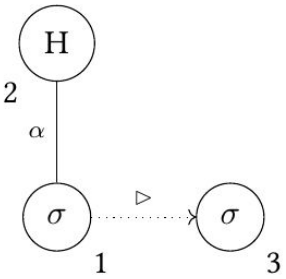
$$\texttt{NextAR(AR)} = \begin{cases} \texttt{addTone(AR)} \cup \texttt{addTBU(AR)} & \text{if } \texttt{containFloat(AR) = False} \\ \texttt{addTone(AR)} \cup \texttt{addTBU(AR)} \cup \texttt{addAssociation(AR)} & \text{if } \texttt{containFloat(AR) = True} \end{cases}$$

Generating Next Superfactor: adding new unit

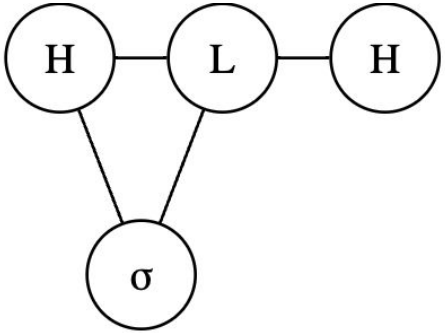
(17) Add a New tone

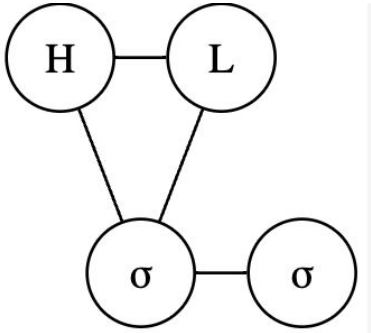


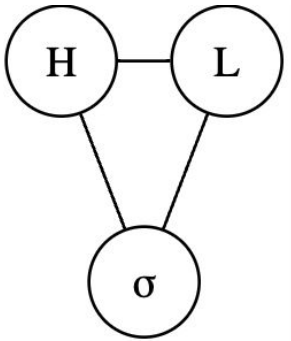
(18) Add new TBU



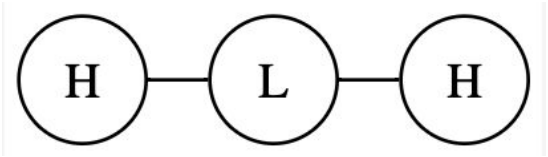
Example

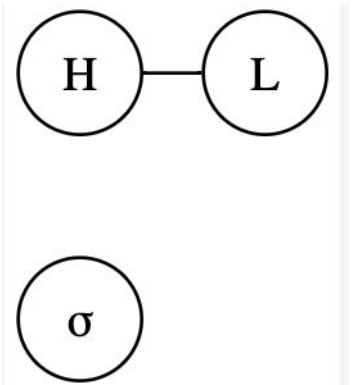


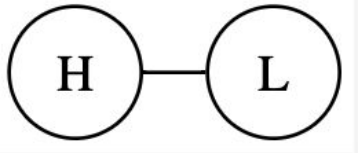




Example



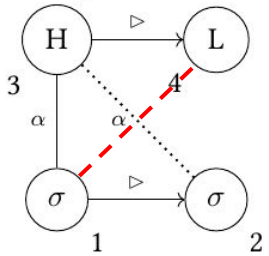




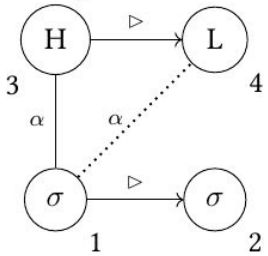
Generating AR Superfactor: adding new association

1. An AR can also be expanded by adding a new association which contain at least one *floating* unit
2. No-Crossing Constraint: (t, s) is well-formed iff there does not exist (t′, s′) ∈ α such that (t′ > t ∧ s′ < s) ∨ (t′ < t ∧ s′ > s). Bird and Klein (1990); Jardine (2013); Jardine and Heinz (2015)
3. An association <1,4> is invalid in (22) since the existing association of <3,2> and 1 < 3 ∧ 4 >2.

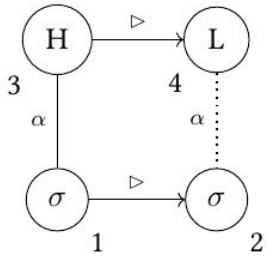
(22) floating TBU to tone

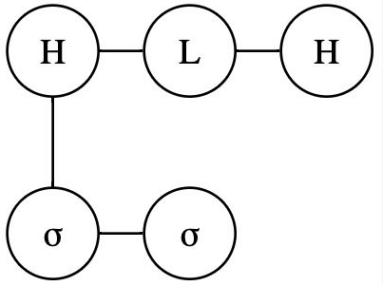


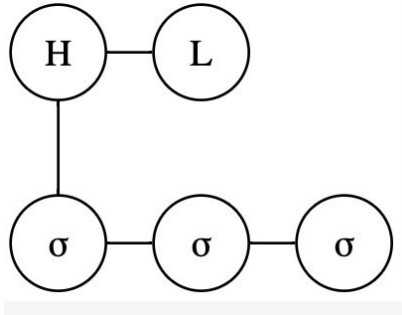
(23) floating tone to TBU

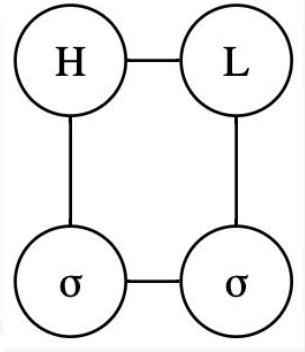


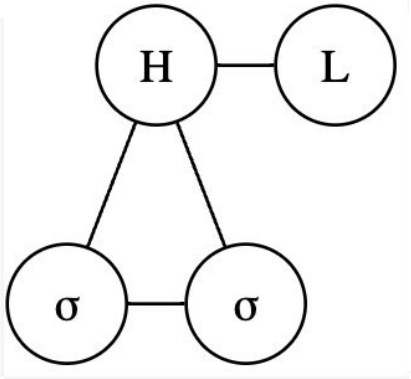
(24) float tone to float TBU

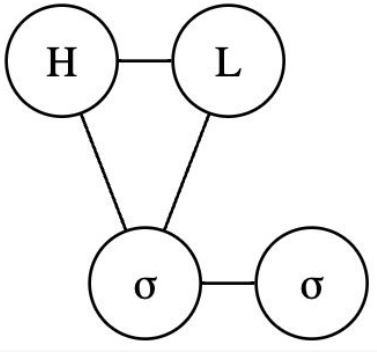


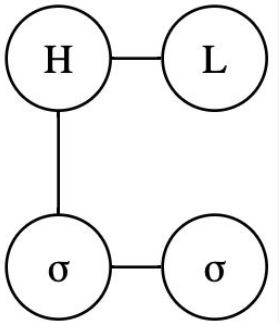

H
L
H
σ
σ


H
L
σ
σ
σ


H
L
σ
σ


H
L
σ
σ


H
L
σ
σ


H
L
σ
σ

Case Study: Hausa

1. Hausa (Chadic) has H and L as two contrast tones, and the TBU is the syllable which consists of at most two moras.

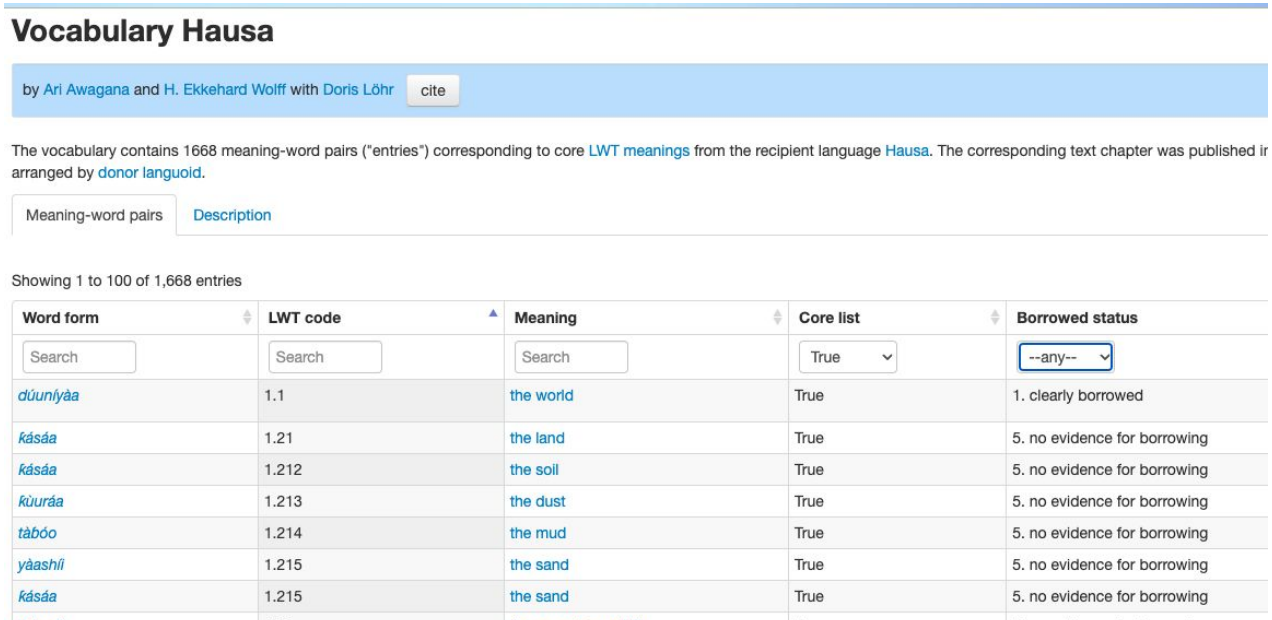
2. Linguistically Attested Constraints (Rule constraints) (Newman 2002 p600, Zoll 2003, Leben 1971)

a.	***Rise**	No monosyllabic rising
b.	***3t-Contour**	No three-tone contour on monosyllabic forms
c.	***Initial-Contour**	No initial HL for di-/polysyllabic forms
d.	***LL**	Avoid LL sequence

Can BUFIA-AR learn these Rule Constraints?

Step 1 Data Preparation

A Hausa mini-dictionary (Awagana et al., 2009) from World Loanwords Database was used.
From a total of 1668 core-meaning words, 664 unborrowed monomorphemic forms were retained.



Vocabulary Hausa

by Ari Awagana and H. Ekkehard Wolff with Doris Löhr cite

The vocabulary contains 1668 meaning-word pairs ("entries") corresponding to core LWT meanings from the recipient language Hausa. The corresponding text chapter was published ir
arranged by donor languoid.

Meaning-word pairs | Description

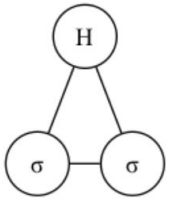
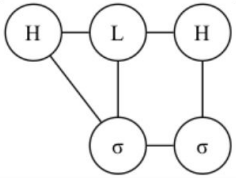
Showing 1 to 100 of 1,668 entries

Word form	LWT code	Meaning	Core list	Borrowed status
Search	Search	Search	True	--any--
dúuníyàa	1.1	the world	True	1. clearly borrowed
ƙásáa	1.21	the land	True	5. no evidence for borrowing
ƙásáa	1.212	the soil	True	5. no evidence for borrowing
ƙùuráa	1.213	the dust	True	5. no evidence for borrowing
tàbóo	1.214	the mud	True	5. no evidence for borrowing
yàashíi	1.215	the sand	True	5. no evidence for borrowing
ƙásáa	1.215	the sand	True	5. no evidence for borrowing

Step 2 Representation Conversion

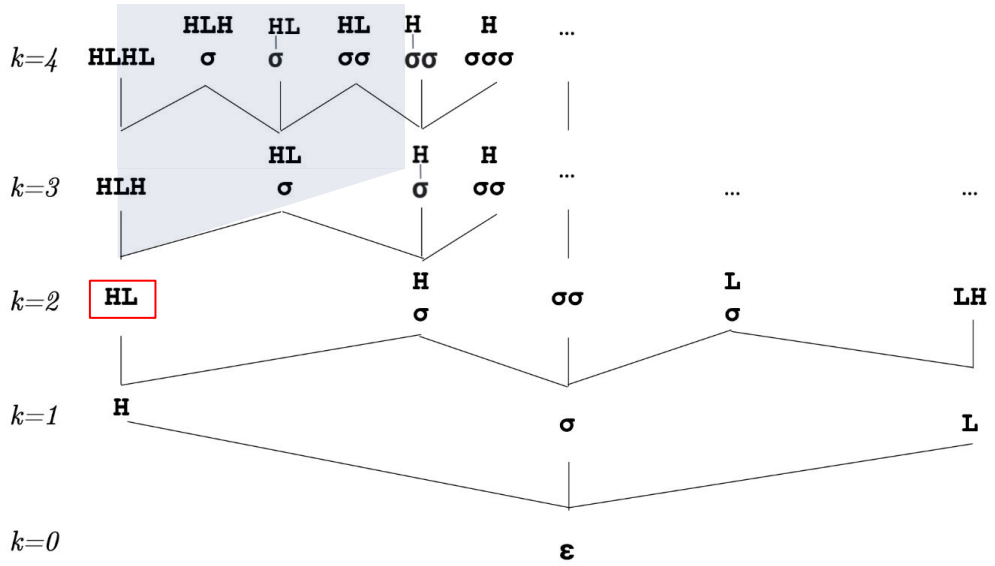
A Python script was created to convert orthographic forms into ARs as in *Coding* which lists information of:

1. Converted tonal sequence
   a. *kásáa* "the soil" will have **H**
   b. *bâutáa* "to worship" will have **HLH**
2. An list of associations **(t,s)** (e.g. **[(1,1),(1,2)]**) indexes the t^{th} tone associates to the s^{th} syllable
3. The last item in the list indicates the number of tones and syllables

Transcription	AR	Coding
kásáa		`('H'[(1,1),(1,2)])`
bâutáa		`('HLH',[(1,1),(2,1),(3,2)])`

Step 3 Implementing BUFIA over ARs

BUFIA initiates from the empty structure
- generates its next superfactors,
- check if each of them is contained in data
  - Not contained:
    - Add to Grammar
    - Ignore its Superfactors
  - Contained:
    - Under a certain size?
    - Contain no constraint?



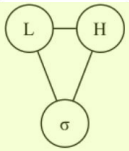
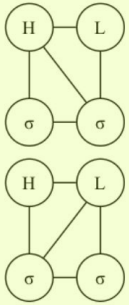
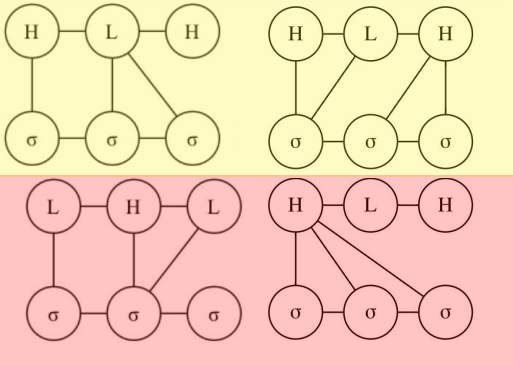
Result

664 native words (in orthography, with tonal markers) from a Hausa mini-dictionary were used as a positive data. BUFIA identified 7 banned structures when syllable and tone numbers are ≤ 3.

Some have been reported before in linguistic analyses

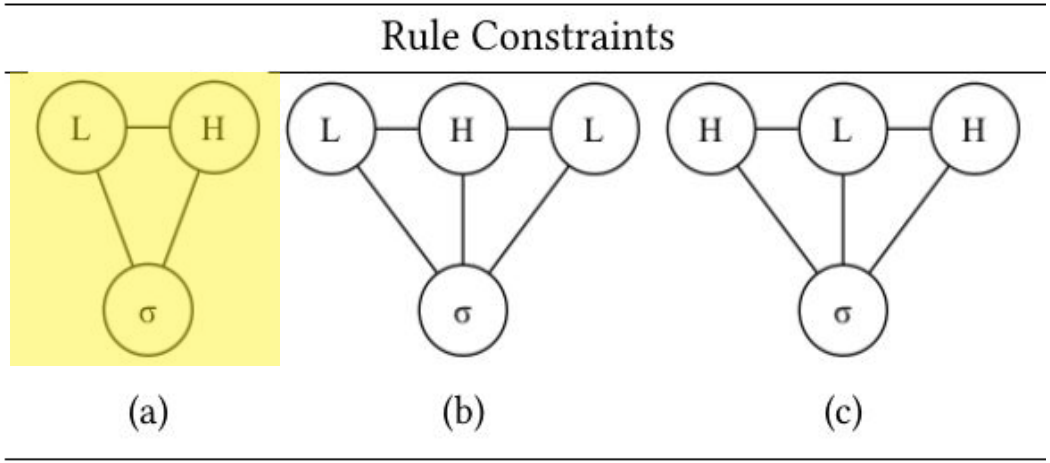
Some are more specific than previously reported generalization

Some have never been discussed

	$t=1$	$t=2$	$t=3$
$s=1$	Found nothing 0		Nothing New Found -
$s=2$	0		-
$s=3$	0	-	

*Rise & *3t-Contour

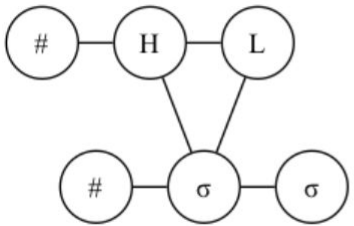
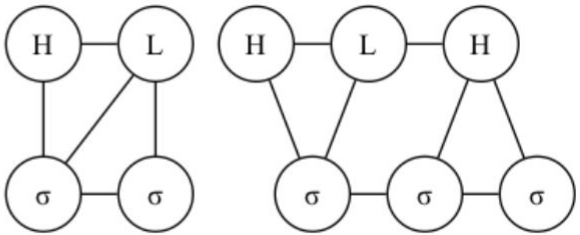
1. ***Rise** BUFIA constraint = Rule Constraint
2. ***3t-Contour** entailed by *Rise which is already banned by BUFIA thus no further check
3. Overall, BUFIA Constraints match with Rule constraint



*Initial-Contour

1. Rule constraint *#σ̂ not match with BUFIA constraint *σ́σ̀ and *σ́σ́σ
2. BUFIA constraints also indicate the existence of Initial Contours
3. 7 words in the data have initial HL: 6 = σ́σ̀ ; 1= σ́σ́σ̀

✓ BUFIA Constraints make more specific generalization than Rule constraint

Rule Constraint	BUFIA Constraint
	
(a) *#σ̂	(b) *σ́σ̀ (c) *σ́σ́σ

Hausa words do contain an initial HL

ƙyânwáa	"the cat"
kûnnée	"the ear"
tsûmmáa	"the handkerchief or rag"
sâiwáa	"the root"
jînƙái	"the pity"
bâutáa	"to worship, to obey"
yânyáawàa	"the fox (fennec of Sahara)"

*LL

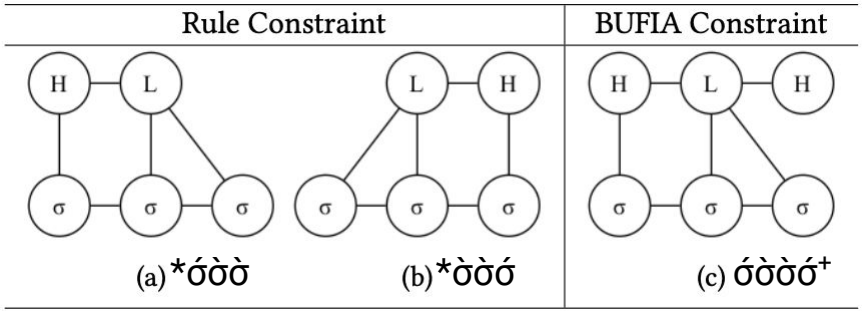
Here I discuss two specific cases: *ὸὸό (*LLH) or *ό̀ὸ̀ (*HLL) which are commonly reported disfavored patterns.

*ὸὸό (*LLH)

- No BUFIA constraint are found
- 16 ὸὸό are found in the data

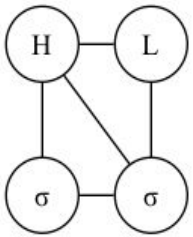
*ό̀ὸ̀ (*HLL)

- BUFIA constraint (c) is the superfactor of (a)
- 1 word *sábòodà* “because”
- Rule revision: *ό̀ὸ̀ό

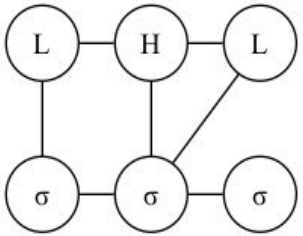


Something New

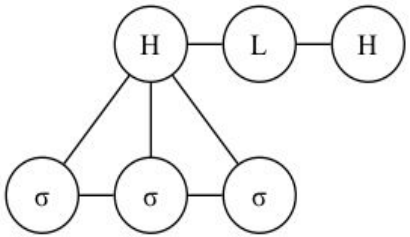
Constraints BUFIA found that are not reported linguistically.



(a) σ́σ̌



(b) σ̀σ̌σ



(c) σ́σ́σ́σ̀⁺σ́⁺

Unexpressible in string representation

Future Questions

1. Enrich the representation to incorporate syllable weight, consonant voicing, register
2. How do we deal with processes over ARs?

Appendix Slides

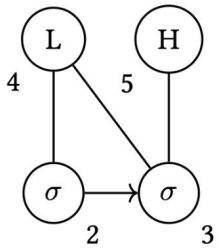
Restriction

Chandlee et al. (2019, p.94).

Definition 1. *$A = \langle D^A; \lhd, R^1_A, \ldots, R^n_A \rangle$ is a restriction of $B = \langle D^B; \lhd, R^1_B, \ldots, R^n_B \rangle$ iff $D^A \subseteq D^B$ and for each m-ary relation R_i, we have $R^i_A = \{(x_1, \ldots, x_m) \in R^i_B \mid x_1, \ldots, x_m \in D^A\}$.*

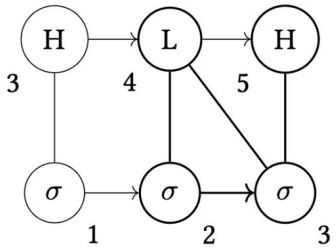
Based on Definition 3, (16) is a restriction of (17) since $D^A = \{2,3,4,5\} \subseteq \{1,2,3,4,5\} = D^B$ and $R^A_H \in R^B_H, R^A_L \in R^B_L, R^A_\sigma \in R^B_\sigma, \lhd^A \in \lhd^B, \alpha^A \in \alpha^B$.

(16)



$$\mathcal{M}^{\mathcal{A}} = \{\mathcal{D} \mid R_H, R_L, R_\sigma, \lhd, \alpha\}$$
$$\mathcal{D}^{\mathcal{A}} = \{2,3,4,5\}$$
$$R^A_\sigma = \{2,3\}$$
$$R^A_H = \{5\}$$
$$R^A_L = \{4\}$$
$$\lhd^A = \{\langle 2,3\rangle\}$$
$$\alpha^A = \{\langle 2,4\rangle, \langle 3,5\rangle, \langle 4,2\rangle, \langle 5,3\rangle\}$$

(17)



$$\mathcal{M}^{\mathcal{B}} = \{\mathcal{D} \mid R_H, R_L, R_\sigma, \lhd, \alpha\}$$
$$\mathcal{D}^{\mathcal{B}} = \{1,2,3,4,5\}$$
$$R^B_\sigma = \{1,2,3\}$$
$$R^B_H = \{3,5\}$$
$$R^B_L = \{4\}$$
$$\lhd^B = \{\langle 1,2\rangle, \langle 2,3\rangle, \langle 3,4\rangle, \langle 4,5\rangle\}$$
$$\alpha^B = \{\langle 1,3\rangle, \langle 2,4\rangle, \langle 3,5\rangle, \langle 3,1\rangle, \langle 4,2\rangle, \langle 5,3\rangle\}$$

No.	Form	Meaning	No.	Form	Meaning
1	**gùngùníi**	“to mumble”	10	jùuyàayíi	“the anxiety”
2	**zàzzàɓíi**	“the fever”	11	tàuràaróo	“the star”
3	**shànshàaníi**	“the centipede”	12	shùugàbáa	“the president”
4	**ƙàiƙàyíi**	“the itch”	13	tàbàráu	“the spectacles/glasses”
5	tùnkùyáu	“the flea”	14	tùrùríi	“the steam”
6	màƙèesúu	“the firefly”	15	sùrùkái	“the parents-in-law”
7	màlàfáa	“the hat or cap”	16	ƙwàràngwál	“the carcass”
8	gìzàagóo	“the adze”	17	sábòodà	“because”
9	màràicée	“the evening”			

Table 12: Hausa Words with either LLH or HLL pattern