Affix-specificity makes stress learnable

Michael Becker

michael.becker@stonybrook.edu

LSA 91st Annual Meeting, Austin TX

January 5–8, 2017

Highlights:

- I show how Turkish stress (Inkelas & Orgun 2003), and lexical stress more broadly, is now learnable thanks to Sublexical Phonology (Becker & Gouskova 2016; Allen & Becker 2015).

- Egyptian Arabic, including (lots of) its non-concatenative morphology, is learnable as well in the Sublexical approach.
  My analysis offers better empirical coverage than Hayes (1995), with lesser computational complexity.

Introduction

(1) The SPE-style analysis privileges a single grammar + single UR for each morpheme (Chomsky & Halle 1968; Kenstowicz & Kisseberth 1977, 1979).

(2) The single grammar/single UR analysis was never shown to be learnable from realistic language data, despite many attempts (Tesar & Smolensky 1998, 2000; Tesar et al. 2003; Jarosz 2006; Riggle 2006; Rasin & Katzir 2016).

(3) Some progress from giving up on a single UR (Pater et al. 2012).

(4) Much progress from giving up on URs (Albright & Hayes 2002, 2003, 2006).

Towards learnable analyses:

(5) Learning morphological categories one at a time, Sublexical Phonology fully covers a language's alternation patterns.

(6) Together with the UCLAPL (Hayes & Wilson 2008), language-wide and category-specific phonotactic generalizations are learnable as well.

Sublexical phonology (Becker & Gouskova 2016; Allen & Becker 2015):

(7) Morphological categories/affixes are learned one at a time.
One consequence: cyclic derivation.

(8) Each morphological category/affix is associated with sublexicon(s).

Sublexicon:

- Morphophonologically uniform set of paradigms,
e.g. [ara'ba ∼ araba'lar, ka'nat ∼ kanat'lar, o'da ∼ oda'lar].
- A deterministic, productive operation,
e.g. "destress last vowel, add ['lar] at right edge".
- GateKeeper Grammar: a MaxEnt phonotactic grammar over the bases, e.g. GateKeeper says: "base must have final stress".

(9) Learns from pairs of surface forms, following Albright & Hayes (2002, 2003, 2006).
See also Moore-Cantwell & Staubs (2014).

(10) Implementation: `http://sublexical.phonologist.org/`
Includes all of today's simulations.

Sublexical analysis of Turkish stress

(11) All generalizations and evidence from Inkelas & Orgun (2003).

(12) Sublexical analysis: Turkish only has two kinds of suffixes.

Stress-neutral suffixes (I&O's prestressing):

(13) Affix is stressless on the surface, doesn't change the base stress.

INTERROGATIVE [mɯ]:

base	interrogative	
araba-ˈlar	araba-ˈlar-mɯ	car-PL
jaʃa-ˈdɯ	jaʃa-ˈdɯ-mɯ	live-PAST
ˈtoka	ˈtoka-mɯ	belt
jaˈʃa-ma-dɯ	jaˈʃa-ma-dɯ-mɯ	live-NEG-PAST

(14) Sublexical analysis of stress-neutral suffixes:

One sublexicon for INTERROGATIVE, "add [mɯ] at right edge".
All constraint weights remain at zero = permissive grammar.

(15) Ignoring vowel harmony today.

(16) Similarly, one sublexicon for NEGATIVE [ma], COMITATIVE [(j)la], etc.

Stress-attracting suffixes (I&O's neutral and stressed)

(17) Two surface forms = two sublexicons per suffix

PLURAL ['lar]/[lar]:

		base	plural	
a.	Non-final stress on base	ˈpenaltɯ ˈtoka	ˈpenaltɯ-lar ˈtoka-lar	penalty belt
b.	Final stress on base	araˈba ˈʤan	araba-ˈlar ʤan-ˈlar	car soul

(18) Sublexical analysis of stress-attracting PLURAL:

- "add [lar]"
  GateKeeper: *FINALSTRESS $w = 10.2$, *MONOSYLLABIC $w = 1.8$
- "add ['lar], destress last vowel"
  GateKeeper: *NONFINALSTRESS $w = 7.8$

(19) Similarly two sublexicons for ADV ['arak]/[arak], LOCATIVE, PAST, etc.

Place name stress

(20) I&O: “Sezer stress ... a zero-derivation process”

prosody	base	place name	sublexical operation
HL'X#	tor.ba.'lɯ	'tor.ba.lɯ	“destress, stress antepenult”
final stress	be.'bek	'be.bek	“destress, stress penult”
other	af.'ri.ka	af.'ri.ka	“do nothing”

(21) Sublexical analysis of place names: three sublexicons

- “destress last vowel, stress antepenult” , bases HL'X#
- “destress last vowel, stress penult”, other bases with final stress
- “do nothing”, applies to bases with non-final stress

(22) Currently implementation is imperfect; doesn’t know about penult, antepenult nuclei.

Turkish: broader picture

The disyllabic domain of Turkish stress:

(23) I&O: lexical stress (=fixed stress) always on a disyllabic domain, e.g. *[kaˈvan ∼ kaˈvan-lar].

(24) Sublexical analysis: PLURAL "add [lar]" paired with *FINALSTRESS. Similarly for LOCATIVE, POSSESSIVE etc.
Captured for each suffix separately, not as a cross-affixal generalization.

(25) Compare to minimally different Hebrew PLURAL [ˈim]/[im], which allows monosyllabic domains, e.g. [saˈlat ∼ saˈlat-im] 'salad(s)'.

Generalizations across affixes:

(26) Central to the single-grammar generative analysis.

Accidental/conspiratory in the Sublexical analysis, where each suffix is learned and analyzed separately.

Any **empirical** evidence for cross-affixal speaker knowledge?

(27) If indeed cross-affixal generalizations are real:
Sublexicons **could** influence each other, e.g. via Bayesian priors.

I&O's leftmost wins: the leftmost lexically marked morpheme wins.

(28) Once non-final stress is present in the base or generated by one suffix, following suffixes choose stressless allomorphs.
Captured in the Sublexical analysis as a conspiracy.
Descriptively adequate and learnable.

Generalizations over monomorphemic forms:

(29) Beyond the purview of sublexicons, which are affix-specific.

(30) Capturable by phonotactic grammar(s) à la Hayes & Wilson (2008).

Language-wide:	one stress per word, etc.
Nominal roots only:	final stress is common.
Verb roots only:	final stress is required.

(31) Allen & Becker (2015): the probability of a word is the product of its probability in all of the grammars it belongs to.
Cf. Shih (4pm).

Turkish summary:

(32) Together with wider phonotactic grammars, the Sublexical analysis captures all known speaker knowledge of Turkish.

The first learnable analysis of Turkish stress alternations.

Irregular stress in Egyptian Arabic

Overview:

(33) Hayes (1995): L→R moraic trochees, stress last foot, delete other feet.

(34) Graf (2010): dependency between the left edge of the word and the stressed syllable is unbounded. Makes phonology >star-free regular.

(35) My analysis: Non-directional moraic trochees + lexical exceptions. Better empirical coverage than Hayes (1995), solves Graf's (2010) computational complexity problem.

(36) Sources: Mitchell (1960), Broselow (1976), Hinds & Badawi (1987) , a.o.

Heavy syllable in the three-syllable window:

(37) Predictable final stress if the ultima is heavy

LLˈH#	ga.la.ˈsaːt	‘meetings’
HLˈH#	jas.ta.ˈħimm	‘he bathes’
HHˈH#	tag.rib.ˈteːn	‘two tests’

(38) Predictable penultimate stress if the penult or antepenult are heavy

HˈLL#	ʔit.ˈwi.lid	‘he was born’
	maʕ.ˈma.ʕa	‘turmoil’
HˈHL#	ʔis.ˈtaʕ.mil	‘he used’
LˈHL#	ji.ˈʁaj.jar	‘he changes’

Following everyone: word-final CVC is light, otherwise CVC heavy.

(39) Heavy syllable inside 3σ window → 100% predictable stress.
Across all lexical categories, loanwords, etc.

Analysis:

(40) Non-directional moraic trochees, WSP, main stress on final foot. Language-wide phonotactic grammar.

	Parse-σ	Rightmost
a. ('ʔit)wi.lid	*	
b. ('ʔit)(wi.lid)		*
c. ☞ (ʔit)('wi.lid)		

(41) The same analysis correctly predicts LL'LL#

	Parse-σ	Rightmost
a. ('ka.ta)bi.tu	**	
b. ka('ta.bi)tu	**	
c. ('ka.ta)(bi.tu)		*
d. ☞ (ka.ta)('bi.tu)		

(42) My analysis allows two options for LLL#, HLLL#

	Parse-σ	Rightmost
a. ☞ (ʔit)(ˈka.ta).bu	*	
b. ☞ (ʔit)ka.(ˈta.bu)	*	

(43) Another option:
Footless implementation with UCLAPL (Hayes & Wilson 2008).
Uglier constraints, but descriptively adequate.

(44) Hayes (1995): L→R moraic trochees: (ˈLL)L okay, *L(ˈLL) impossible.

(45) Our analyses agree everywhere else, e.g. (LL)(ˈLL), (H)(ˈLL), etc.

Triple light words:

(46) ˈLLL# is intuitively more much more common, in types and tokens. Supported by OO-Faith, MSA, neighboring dialects.

ˈLLL#	ˈka.ta.bit	‘she wrote’
	ˈʃa.ga.ra	‘one tree’
	ˈbu.χa.la	‘misers’

(47) LˈLL# is less common, but includes Egyptian innovations, found neither in MSA nor in neighboring dialects.

LˈLL#	ra.ˈmi.tu	‘she threw it/him’
	tu.ˈku.sa	‘taxis’
	ʕa.ˈla.mi	‘world-ADJ’

(48) **Triple light words never monomorphemic in Egyptian Arabic. Stress on triple lights is predictable from the affixes involved.**

Both types of triple lights deserve to be analyzed.

Exhaustive list of triple light sources: Verbs

(skipped)

Exhaustive list of triple light sources: Nouns and adjectives

(49) Nominal: mass noun + singulative [a]

mass noun	singular		
ˈʁa.nam	ˈʁa.na.m-a	ˈLLL#	‘sheep’
ˈma.kan	ˈma.ka.n-a	ˈLLL#	‘machine’

Predictable trochaic stress on [ˈʁa.nam],
OO-Faithful stress on [ˈʁa.na.ma],
Or the singulative [a] specifies antepenult stress.

(50) [ˈma.kan] is backformated < Italian macchina [ˈmak(ː)ina],
Showing that LLL is productively monomorphemic.

(51) Nominal/adjectival broken plurals: ˈCaCaCa, ˈCuCaCa

singular	plural		
ˈsaː.fil	ˈsa.fa.la	=Levantine, MSA ˈsa.fa.la	‘vile (person)’
ba.ˈχiːl	ˈbu.χa.la	=Levantine, < MSA bu.χa.ˈlaːʔ	‘miser(ly)’

High type frequency.

(52) Nominal broken plurals: CuˈCuCa, CiˈCiCa

singular	plural		
ˈda.kar	du.ˈku.ra	cf. MSA ðu.ˈkuː.r(a)	‘male’
ˈnimr	nu.ˈmu.ra	cf. MSA ˈnu.mur	‘tiger’
ˈkart	ku.ˈru.ta	< French	‘card’
ˈtaks	tu.ˈku.sa	< Italian/English	‘taxi’

Egyptian innovation. Mitchell (1960); Hinds & Badawi (1987): stigmatized, avoided by educated speakers.

(53) Noun summary: plural templates must specify stress, ˈLLL vs. LˈLL.

Sublexical analysis of Egyptian triple light plurals:

(54) Biggest sublexicon, operation made of three unordered parts:

- insert [a] after 4^{th} segment of the base,
- change last nucleus of the base to [u],
- insert [ˈu] after 3^{rd} segment of the base

[táks] ‘taxi’ → táksa → tuksa → [tukúsa] ‘taxis’.

(55) Other sublexicons for disyllabic bases (e.g. [ˈda.kar ∼ du.ˈku.ra] ‘male’), CiˈCiCa plurals, ˈCaCaCa plurals, ˈCuCaCa plurals.

(56) Find the simulation at `http://sublexical.phonologist.org/`. The only place you’ll find a learner of Arabic broken plurals.

Egyptian summary:

(57) Egyptian stress 100% predictable with a heavy in the three-syllable window and with LLLL# words.
Language-wide phonotactic grammar; learned with the UCLAPL (Hayes & Wilson 2008) without feet, or analyzed by hand with moraic trochees.

(58) In triple light words, general markedness is more permissive.
Stress is predictable from morphological information.

(59) Single grammar analysis: lowly ranked OO-Faith and other effects emerge, including faithfulness to lexically specified plural templates.

(60) Markedness dominating faithfulness to lexical stress is well-known, e.g. verbal stress in Hebrew, Spanish, Japanese (Smith 2002, 2010), English secondary stress (Pater 2000), etc.

(61) Sublexical analysis:

- Each verbal suffix is analyzed separately, specifying stress on outputs.
- Each broken plural pattern is analyzed separately, specifying stress on outputs.

Stress assigned to derivatives is redundant where predictable; informative on triple light words.

(62) The Sublexical analysis is needed to learn productive affixation, even when the affix has uniform stress, e.g. object 1SG always with penult stress:

3SG.M	+ 1SG	
ˈka.tab	ka.ˈtab.ni	‘he wrote’
kat.ˈta.bit	kat.ta.ˈbit.ni	‘she dictated’
ˈra.ma	ra.ˈmːa.ni	‘he threw’

Even when stress is predictable, we still need a mechanism for learning that the suffix is [ni]/[ːni].

Conclusions

Analysis of lexical stress in Turkish:

(63) Inkelas & Orgun's (2003) single-grammar analysis:

Turkish assigns final stress by default.
Unstressed URs receive this stress, e.g. [araˈba ∼ araba-ˈlar] 'car-PL',
cf. /bɯrak-ˈɯver-ˈerek/ → [bɯraˈkɯvererek] 'leave-SUDDEN-ADV'.

(64) Sublexical analysis:

Turkish prefers final stress phonotactically.
Some affixes are stressed and remove stress from the base;
other affixes are unstressed and leave the base stress as is.

This partially recapitulates the language-wide requirement for culminativity.

Better coverage + learnable analysis in Egyptian Arabic:

(65) Hayes (1995): Egyptian stress is predictable, assigned by R→L moraic trochees + erasure of secondary stress.
Doesn't cover productive L'LL patterns that are unique to Egyptian.

(66) Language-wide phonotactic grammar + Sublexical analysis:

Stress is predictable everywhere except triple light words.
Learning morpheme-specific operations ensures correct stress assignment when triple light words are derived.

No need to assign stress to triple light bases — they don't exist.

Towards a universal learner of alternations:

(67) Learning simulations of Turkish, Egyptian, and others:
`http://sublexical.phonologist.org/`

(68) No Arabic-specific machinery. All done with machinery that had been built for Russian vowel deletion, English past tense, Tagalog infixation, etc.

(69) The Sublexical analysis is fragmented; might seem inelegant.
But elegance must not be used to prefer a hand-crafted analysis over a learnable analysis.

(70) Some things to do:

- Integrate Sublexical grammars and (hierarchically structured) phonotactic grammars for wider empirical coverage.
- Collect evidence for cross-affixal generalizations.

Thank you

For their thoughtful comments and suggestions, I thank Honaidah Ahyad, Ellen Broselow, Thomas Graf, Joe Pater and John McCarthy.

Thanks to Maria Gouskova and Blake Allen for being a constant source of insight and inspiration.

Thanks to Stephanie Shih and Claire Moore-Cantwell for organizing this spesh sesh.

Bonus slides

MSA (Modern Standard Arabic) provides longer words:

(71) Hayes (1995), data from Mitchell (1960):

MSA	gloss	Egyptian
ʔad.wi.ja.ˈtu.hu.maː	‘their.DUAL medications’	ʔad.wij.ˈtu.hum
ʃa.ʤa.ˈra.tu.hu	‘his tree’	ʃa.ˈgar.tu

(72) Positive evidence for a MSA phonotactic grammar, separate from Egyptian: presence of [ʤ], unsyncopated high vowels, unstressed long vowels.

(73) Educated speakers have a separate grammar for MSA.

Exhaustive list of triple light sources: Verbs

(74) Verb stem + V-initial subject agreement [-it], [-u]

stem	stem + 3SG.F		
ˈfa.taħ	ˈfa.ta.ħ-it	ˈLLL#	‘she opened’
ʔit.ˈfa.taħ	ʔit.ˈfa.ta.ħ-it	HˈLLL#	‘she was opened’
ˈra.ma	ˈra.m-it	ˈLL#	‘she threw’
ˈʃaːf	ˈʃaː.f-it	ˈHL#	‘she saw’

Single grammar analysis: (H)ˈLLL# determined by OO-Faith.

(75) Verb stem + 3SG.F [-it] + V-initial object agreement [-u], [-ak], [-ik]

stem + 3SG.F	+ 3SG.M		
ˈfa.ta.ħ-it	fa.ta.ˈħ-i.t-u	LLˈLL#	‘she opened it/him’
ˈra.m-it	ra.ˈm-i.t-u	LˈLL#	‘she threw it/him’
ˈʃaː.f-it	ʃa.ˈf-i.t-u	LˈLL#	‘she saw it/him’

Single grammar analysis: stress always on [it]; lexical stress emerging when markedness allows, i.e. in triple light words.

(76) Verb summary: predictable stress given morphological structure.

Sublexical analysis of verbs: subject 3SG.F [-it], object 3SG.M [-u]

(77) Trained on pairs of verbal forms as above.

(78) Adds affixes and adjusts stress correctly, e.g. [ʔab.ˈba.sit] → [ʔab.ba.ˈsi.tu].

(79) Recapitulates language-wide stress pattern when heavy syllable present, e.g. [ˈʔab.bas] → [ʔab.ˈba.sit]. But informative in triple light words.

(80) Find the simulation at `http://sublexical.phonologist.org/`.

References